dasc°pm v1.1

Michael Schulz ▪ Uwe Neuhaus ▪ Jens Kaufmann ▪ Stephan Kühnel ▪ Emal M. Alekozai ▪ Heiko Rohde ▪ Sayed Hoseini ▪ René Theuerkauf

DASC-PM v1.1

A Process Model for Data Science Projects

Daniel Badura ▪ Ulrich Kerzel ▪ Carsten Lanquillon ▪ Stephan Daurer ▪ Maik Günther ▪ Lukas Huber ▪ Lukas-Walter Thiée ▪ Philipp zur Heiden ▪ Jens Passlick ▪ Jonas Dieckmann ▪ Florian Schwade ▪ Tobias Seyffarth ▪ Wolfgang Badewitz ▪ Raphael Rissler ▪ Stefan Sackmann ▪ Philipp Gölzer ▪ Felix Welter ▪ Jochen Röth ▪ Julian Seidelmann ▪ Uwe Haneke

The current version of the work
DASC-PM v1.1: A Process Model for Data Science Projects
is based on the publication

Schulz, Michael; Neuhaus, Uwe; Kaufmann, Jens; Kühnel, Stephan; Alekozai, Emal M.; Rohde, Heiko; Hoseini, Sayed; Theuerkauf, René; Badura, Daniel; Kerzel, Ulrich; Lanquillon, Carsten; Daurer, Stephan; Günther, Maik; Huber, Lukas; Thiée, Lukas-Walter; zur Heiden, Philipp; Passlick, Jens; Dieckmann, Jonas; Schwade, Florian; Seyffarth, Tobias; Badewitz, Wolfgang; Rissler, Raphael; Sackmann, Stefan; Gölzer, Philipp; Welter, Felix; Röth, Jochen; Seidelmann, Julian; Haneke, Uwe (2022)
DASC-PM v1.1 – Ein Vorgehensmodell für Data-Science-Projekte
NORDAKADEMIE gAG Hochschule der Wirtschaft, Elmshorn 2022
DOI: 10.25673/85296

which was published under the Creative Commons (CC) license BY 4.0
(https://creativecommons.org/licenses/by/4.0/)

The list of authors includes everyone who actively participated in revising and editing Version 1.1 and consented to being listed.
They would like to thank everyone who participated in Version 1.0 for their work on the previous draft.

This work is licensed under a Creative Commons Attribution 4.0 International License.
https://creativecommons.org/licenses/by/4.0/

ISBN: 978-3-9824465-1-6

Elmshorn 2022

info@dasc-pm.org

Publisher:

NORDAKADEMIE gAG Hochschule der Wirtschaft
Köllner Chaussee 11
25337 Elmshorn

Supported by the
NORDAKADEMIE foundation

Graphic design: Supported by Fritjof Wild

Table of Contents

List of figures

List of tables

Foreword to Version 1.1

In February 2020, the first version of a comprehensive process model for data science projects appeared: the Data Science Process Model (DASC-PM). The positive feedback we have received indicates we were able to contribute to the discussion of data science activities that we were hoping for. Over the last two years, the DASC-PM has found its way into practice, book contributions (such as Alekozai et al., 2021), and scientific conferences (such as Schulz et al., 2020).

We would like to sincerely thank all the readers who have shared their experiences with us and drawn our attention to the model's strengths and potential improvements. Of course, special thanks go to those who actively participated in developing the model further. Without them, the path to this Version 1.1 would have been impossible.

This version addresses feedback from theory and practice, as well as a few topics we feel strongly about. For example, we have made the document more legible by giving it a more compelling structure and shorter introductory texts. The model itself now more clearly defines the key areas and phases and their characteristics and shows how their interaction can look in various project configurations, including agile ones. We have examined all the terms used in the document with a critical eye and adjusted and standardized them where necessary. To that end, we have also addressed suggestions for a less formal visualization that is more plausible in practice, and—hopefully, at least—made both the document and the actual model more graphically appealing. Since the DASC-PM was created "by many for many," we felt it was worthwhile to make the overall presentation of the model more accessible, even if it might be a little less scientifically precise.

In terms of content, while developing Version 1.1, we focused on the "project order" phase. Important decisions are made and framework conditions are established at the beginning of data science activities. To that end, we are offering a more comprehensive description of the phase and a practical and applicable questionnaire as a concrete basis for both new and experienced users of data science. Just as in Version 1.0, the results should be seen as the aggregate experiences of all the participants of this working group. This English translation of the original German model makes it possible to use it in international projects, more easily supporting the interdisciplinarity that is intrinsic to data science.

All the results presented in the DASC-PM are still mostly based on the feedback of a diverse working group and constitute a state of debate that is meant to serve as a stimulus and support but never claims to have the last word in the very active field of data science. We are pleased that this living vitality will continue to motivate us to discuss and modify the DASC-PM and make it available to a wide audience. If you are interested in participating or want to be kept up-do-date about current developments of the model, contact us at the address given below.

Elmshorn, Halle (Saale), Hamburg, Krefeld, Mönchengladbach and Stuttgart in June 2022

The DASC-PM Core Team

Contact: info@dasc-pm.org

Foreword to Version 1.0

In recent years, the topic of data science has attracted more and more attention in many organizations. However, what distinguishes this discipline from others, what special features are entailed in a data science project's operational sequence, and what competencies are needed to carry out such a project remain unclear.

Hoping to make a small contribution toward rectifying this uncertainty, we prepared this document from April 2019 to February 2020 in an open, virtual working group with representatives from theory and practice. The document describes a process model for data science projects—the DASC-PM. Our goal was not to develop new approaches, but to bring together available knowledge and give it a suitable structure. What we have drafted should be seen as the aggregate experiences of all participants of this working group.

This document is intended for anyone who participates in data science projects either directly or indirectly. It presumes a basic knowledge of the complex of analytical information systems. The process model should help all interest groups in data science projects to understand the necessary tasks and contexts. In addition, students can use it to become more familiar with the topic area.

Data science is still at the beginning of its development, so this document should not be seen as a conclusive work. We hope it will be considered during future data science projects. Findings gained during those projects should be used both to question, complete, and add greater detail to the current version of this document.

Please contact us if you have suggestions for improving the process model or would like to actively participate in developing it further. The next meeting of the virtual working group is planned for September 2020.

We would like to thank everyone who participated in the working group. In an atmosphere that was both productive and constructive, we believe we have achieved a beneficial result that will promote greater understanding—and learned a lot about data science ourselves.

Hamburg, February 2020

Uwe Neuhaus and Michael Schulz

Contact: michael.schulz@nordakademie.de

Part A
General thoughts and the overall model

1 Data Science

Despite the increased attention paid to it, there is currently no generally accepted, standard definition of data science. Although the term is interpreted differently in the scientific publications of many disciplines, definitions from practice are notable mostly for their heterogeneity. This leads to different expectations—and sometimes misunderstandings—among the groups of participants.

While creating DASC-PM v1.0, the working group participants repeatedly brought up two definitions or summaries of central aspects of data science: van der Aalst (2016) and Provost & Fawcett (2013). However, both list only a few of the aspects that the members of the working group named as relevant for a comprehensive data science definition. They also consider aspects of varying degrees of detail.

Based on the contributions made by the participants of our working group, we recommend the following concise definition that concentrates on the general aspects of data science:

> ***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.***

Figure 1 shows the characteristics of this data science definition that will be considered more precisely in the following.

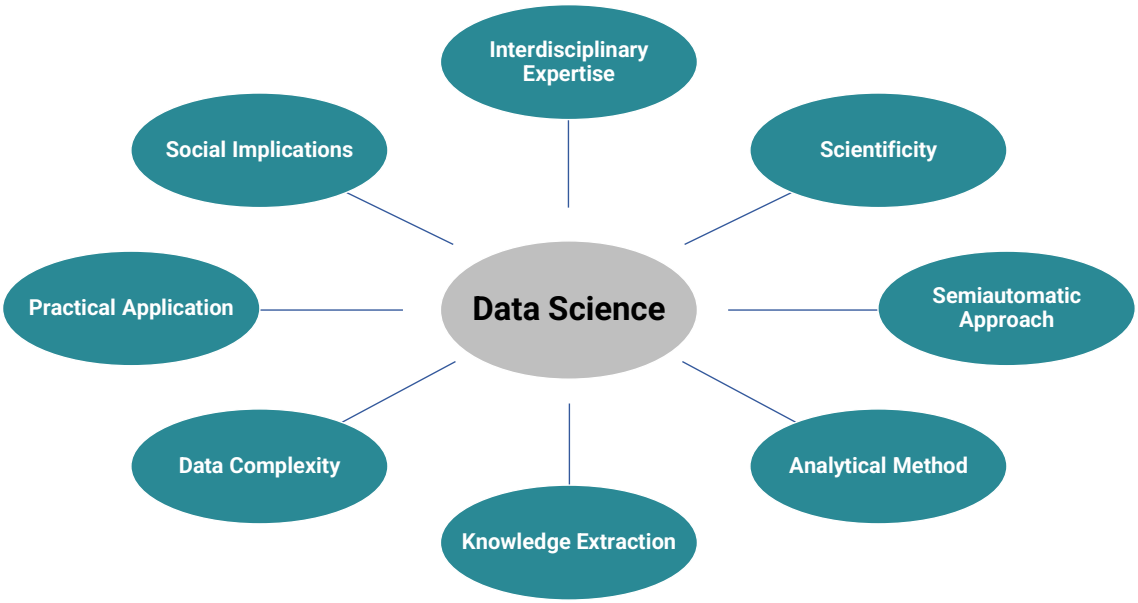


Figure 1: Characteristics of data science

Interdisciplinary Field

We view the interdisciplinary nature of data science as highly relevant. This is founded in the frequently necessary, strong cooperation between various research disciplines, such as mathematics (especially statistics and numerical mathematics), informatics, artificial intelligence, and linguistics. In all these disciplines, there have long been approaches to discussing data scientifically (Provost & Fawcett, 2013).

The emerging field named *data science* appears to have been founded wherever the resources of traditional disciplines no longer suffice to meet current challenges (such as larger, frequently unstructured, or dynamically changing amounts of data). The rapid growth of publicly available information through the internet, drop in prices for computer storage, and growth of processing capacity make it possible to use increasingly complex analytic processes (McAfee & Brynjolfsson, 2012), which are leading more and more toward the emergence of a separate field of expertise.

Furthermore, as an example of interdisciplinarity, we should mention the various domains that use data science in an auxiliary capacity. The fact that many texts on this topic traditionally limit the domain environment to the economy can no longer be justified. Other supportive sciences, such as mathematics and informatics, are used in more than one domain, so interpreting a discipline broadly appears neither new nor problematic. In domains, such as biology, medicine, physics, astronomy, and many more, the use of data science is not only helpful but is already occurring.

Scientificity

The very term *data science* makes it clear that a scientific process is involved. This scientificity is frequently reflected *inter alia* in the general goal of generating knowledge. The focus is not always on using the analysis results directly; sometimes it is on exploring general aspects (such as a procedure's suitability for certain issues), the informative value of individual procedures in relation to the underlying data basis, or evaluating the complexity of various procedures.

Furthermore, scientificity means the examined issue is not trivial. and the chosen procedure should be completely understood, objectively comprehensible, reproducible, documented, and systematically applied.

From a company viewpoint, scientificity also means that analytical methods are borrowed from the scientific world (which is frequently a novelty in a business environment). The actual depth of the scientific discussion varies (including here, primarily regarding the application in a business context), depends on the domain, and can be restricted to an "engineering" approach.

When data analyses are observed using a scientific process, the phrase *data mining* is mentioned frequently—sometimes as a synonym for data science. This is due, among other reasons, to the well-known data mining process models *Knowledge Discovery in Databases* (KDD) (Fayyad et al., 1996) and *Cross Industry Standard Process for Data Mining* (CRISP-DM) (Wirth & Hipp, 2000), which are widely used (at least in business contexts) and impose a "scientific" process on data mining that is structured within limits.

This has influenced the understanding of the term. In the KDD process, only one step—the actual data analysis—is called *data mining*. However, other data mining definitions contain both the data-oriented process steps and the task of examining the analysis results. With its Business Understanding phase, CRISP-DM also explicitly adds a non-technical, application-specific process step.

The necessity of supplementing this initially narrow data analysis by a suitable process model is understandable. However, broadening the original understanding of “data mining” makes it more difficult to use the term in a standardized way. The boundaries between data mining and data science begin to blur.

Analytical Method

We will not explicitly name any particular algorithms or algorithm groups within our data science definition, since we do not want to imply any restrictions on these. As the field is rapidly developing, listing specific algorithms always runs the risk that they will soon become outdated. Even the term “algorithm” itself is less than ideal in the given context, since not all analyses are (solely) algorithm-based and not all algorithms should be considered part of data science.

For that reason, the definition uses the term *analytical method*, which, when applied to data, addresses the core of data science. Analyses that test hypotheses, analyses that are free of hypotheses but have a descriptive, predictive, or prescriptive objective, and other analyses are carried out. To that end, data science can be used to discover patterns, trends, and contexts, as well as optimize purposes.

Existing analytical methods can be applied depending on the use cases given (application cases / problems). But it can also be necessary to enhance analytical methods or develop new ones because no suitable approaches exist.

Semiautomatic Approach

Analytical methods are “semiautomatic” (involving both human and mechanical work steps). Besides the fact that procedures cannot usually be completely automated, we must also mention hybrid learning procedures that are specially developed to encounter problems in the interaction between expert knowledge and analytical methods (Olivotti et al., 2018). Frequently (but not always), this requires high-performance hardware and software that form a complex infrastructure when combined.

Furthermore, complete automation can be sought, depending on the scenario. This will require preparatory, manual work steps. And, in the end, knowledge can be gained only through human participation.

Knowledge Extraction and Data Complexity

One goal of data science is to extract knowledge from mostly complex data that differ in structure, quality, completeness, size, and dimensionality. This can involve static data or data streams, and data can have complex interrelationships.

The development of data science has been strongly driven by the increase in available data quantities (Dhar, 2013). Analyzing enormous, heterogenous data sets has made it necessary to create new procedures, which are frequently listed under the term *big data*. But data science is not limited to big data applications.

Before analytical methods can be applied to data, the data needs to be extracted from the source system, prepared, and made available. Complex infrastructures are often used for this as well.

Practical Application

Data science entails not only extracting knowledge, but also applying it in practice. This can consist of providing findings to domain experts or other recipients, integrating them into existing systems, and/or automatically applying them to new data. When it comes to data science projects, various authors emphasize creating an economic value. Our definition, however, speaks of “practical application” in general to address both scientific and economic objectives.

Extracting findings semiautomatically, the complexity of providing and preparing data, and subsequently applying the data in practice in the form of a software system frequently requires providing or developing a specific IT infrastructure in data science projects. This involves hardware and software components that must be adapted to the project’s framework conditions. The key terms here are scalable architectures, working with distributed data, and cloud connection. The specific IT competencies needed to that end are often brought in from project team members called data engineers. This division of labor allows analysts and IT experts to concentrate on their areas of specialization.

Social Implications

Examining the social consequences of data science by actively participating in the discourse on the resulting ethical and legal issues regarding both the analysis results and the data used as raw material for the analyses should also be considered.

2 Data Scientists

Articles such as the one by Davenport and Patil (2012), in which the job of the data scientist was dubbed *The Sexiest Job of the 21st Century*, can give the impression that all the expertise needed for this area can (or must) be unified in a single person. This viewpoint has been adopted in many publications but is problematic. An overview of existing data scientist definitions is found in Chatfield et al. (2014).

In various sources that build on the article by Conway (2010), data scientists need competencies in three areas:

- *Mathematical-statistical knowledge*
- *Knowledge of information technology*
- *Application-specific knowledge*

Anyone without expertise in all of these areas is not qualified as a data scientist. According to Dhar (2013), abilities in the aforementioned areas must be satisfactory, which is apparent through the overlapping of the individual specialty areas in Figure 2. This frequently cited diagram of Conway (2010) is seen below.

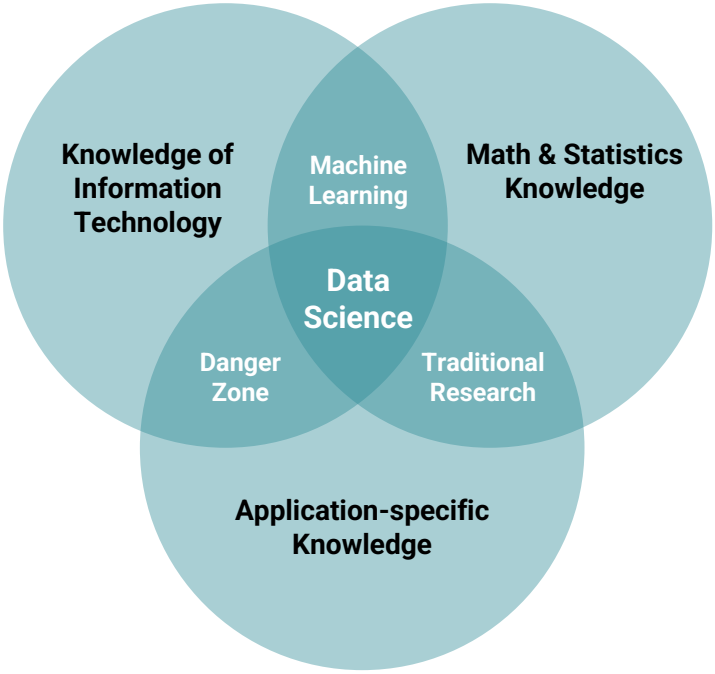


Figure 2: Competencies required of data scientists based on Conway (2010)

Superficial knowledge is not usually enough: Some applications require more in-depth competencies in one or more of the three areas mentioned. Data scientists are also required to be able to

- *communicate with all stakeholders in an appropriate language (Davenport & Patil, 2012),*
- *take on the management of a data science project, and*
- *classify activities strategically.*

Figure 3 summarizes all the competencies needed to carry out data science projects.

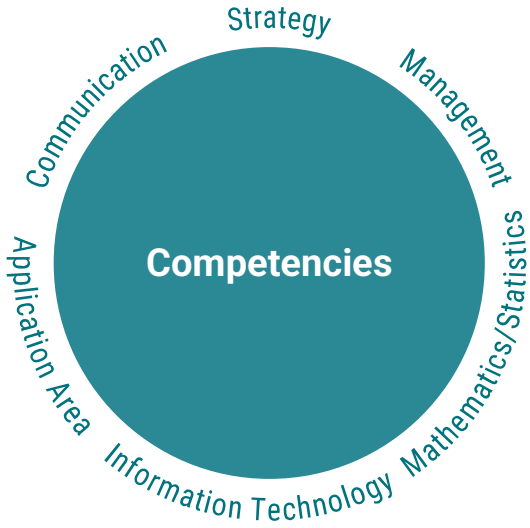


Figure 3: Competencies needed for a data science project

Typically, it is impossible for any one person to gain profound abilities in all the areas named (Zschech et al., 2018). Data scientists can, therefore, either specialize in one discipline, or a few disciplines, or take on more overarching (but less data-oriented) roles.

Such specialization increasingly involves distinguishing between different roles. What follows are the roles identified by the working group participants as relevant to all the activities needed for a data science project. In large projects these roles are frequently divided into sub-roles, although the illustration does not reflect this, since it is intended as an overview. References to the sub-roles can be found in the following descriptions. Additionally, one role does not always need to be mapped to one person. Multiple people can fill one role, or one person can take on multiple roles. The considerations on roles are consolidated in Figure 4. To that end, a distinction is made between four key roles and two supplementary roles.

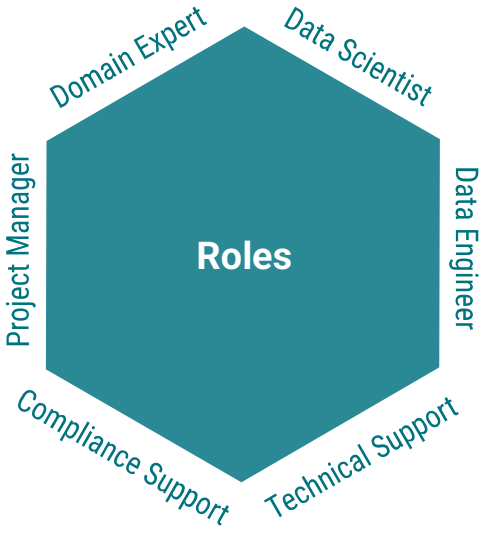


Figure 4: Roles in a data science project

Key role "Data Scientist"

The term *data scientist* is used in practice to represent two different types. First, it is used as an umbrella term for anyone involved in a data science project, and second, it specifically refers to people who specialize in data analysis.

In the sense of the umbrella term, data scientists are responsible for implementing all aspects of a data science project. They work together with domain experts, but are responsible for all methodological, technical, and organizational issues. Although this understanding of data scientist is still frequently encountered in operational practice, one person—as explained above—can take on such a comprehensively defined role in small data science projects.

In the following, this document will always use the term *data scientist* to refer to a specific role:

In a more specific sense, "data scientist" means a specialist in the analysis component of a data science project who is particularly responsible for selecting the analysis methods and tools, performing analyses, and interpreting the results.

In the remaining areas of a data science project, they are active only in an advisory capacity. This means that other tasks must be taken on by other roles.

In large, complex data science projects, the role of the data scientist can be divided into sub-roles:

- ***Data Analyst***
  "Data analyst" is a description that frequently occurs in job advertisements to refer to people who work with various aspects of preparing, analyzing, and evaluating data. They use data-driven analytical methods, statistical models, and methods of data visualization. So in terms of content, there are very large overlaps in the tasks of a data scientist in the specific sense, so that data analyst is frequently considered an older synonym for data scientist. Whenever "data analyst" is seen as a sub-role of "data scientist," the distinction is usually made through prioritization: Data analysts are increasingly active in explorative data analysis and forecasting. In so doing, a data analyst uses more traditional analysis methods and a data scientist focuses more on the analysis model, using new, complex methods.

- ***Method Specialist***
  Method specialists are concerned with researching and enhancing data science methods (data analysis, data transformation, etc.). For example, they draft new analysis algorithms and examine the effect of the parameters that are relevant to that end. Moreover, they are informed on the current state of research in the data science field. Although method specialists can also contribute in the context of application-based data science projects, their focus is more strongly on theory and research.

- ***Data Scientist Consultant***
  Data scientist consultants possess satisfactory methodological, technical, and domain-specific knowledge to give advice on defining suitable analysis issues or application cases. Ideally, they are experienced data scientists who provide their know-how to companies, organizations, or organizational units that must carry out data science projects but lack the necessary expertise.

If necessary, the role of a data scientist can be subdivided further based on background experience (junior data scientist, senior data scientist, advanced data scientist, etc.) or training (graduate of special courses of study or acknowledged professional development or certification programs, lateral entrants, and practitioners).

Key role "Data Engineer"

Data engineers are concerned with procuring, storing, preparing, structuring, and forwarding data. They are particularly active in the preliminary stages of the actual analysis. As data scientists, they have a more technical focus and deal with the IT infrastructure needed for the data science project. The term *data architect* is occasionally used for this role.

One sub-role of a data engineer that is used separately (particularly in large data science projects) is that of *data steward* (also called *data manager* or *data quality engineer*). This role continually deals with the access to the data and its protection, as well as permanently guaranteeing high data quality. A data steward, therefore, has strong points of contact with the technical application area.

Key role "Domain Expert"

Domain experts are key users or their representatives. They have specific knowledge of the application domain and understand the content of the problem or particular application. Domain experts can set priorities for aspects to be modeled or analyzed and are linked to the methodical and technical experts.

Domain experts can be divided into sub-roles. *Business developers* are frequently encountered in business contexts. They develop a domain-specific use case underlying a project and, therefore, form a link between corporate objectives and data analyses, and *business analysts*, who use the developed analysis model during their specialized tasks.

Key role "Project Manager"

Project managers plan, monitor, and coordinate the overall process of a data science project. To do so, they not only need traditional project management competencies but also a good understanding of the methodical and technical aspects of data science, knowledge of suitable process models, and insight into the application domains.

Small projects in particular are frequently managed by people who also function as a data scientist or data engineer. However, project management can also be taken on by people without specific data science expertise if the right experts are available. Such experts, also called *methodical leads* or *technical leads*, possess more profound background knowledge to support the project methodically and technically. Together with domain experts, they determine the scope of the analysis and implementation.

Supplementary role “Technical Support”

Besides the four key roles of a data science project, in which the content is strongly tied to the project goal, two supportive roles are relevant. Although people in supportive roles are necessary to carrying out the data science project successfully, the project results are only indirectly significant for their work. Those people, therefore, contribute to the project’s success as part of their normal activity without being directly affected by the project’s aspects that are specific to data science.

Technical support encompasses all tasks that must be performed to create the technical conditions needed to carry out the data science project. Typical sub-roles of technical support are *IT infrastructure architect*, responsible for drafting a suitable IT foundation for the project, and *IT technician / IT administrator*, who provides the necessary hardware and software and configures the underlying systems. To our understanding, technical support also includes application developers, who deal with implementing application software and tools for using the analysis results productively.

Supplementary role “Compliance Support”

Compliance support is responsible for compliance with statutory requirements, the data science project being compatible with internal regulations, and the project team members performing correctly. It is also responsible for general security management and guarantees data protection (especially for personal data).

3 Key areas for structuring the tasks of a data science project

This chapter will start by considering the fundamental requirements for data science process models and combining these with experiences from additional models. It will then develop key areas of data science projects that can be used to structure a data science process model.

3.1 Fundamental requirements for data science process models

In general, using a process model should increase the quality of data science projects. Traversing through all steps, from conceiving the project to using the findings gained, should be documented as part of this process. In particular, it must be possible to recognize places in which findings were gained by using analytical methods and interpretations were supplemented through domain knowledge. This can ensure that the results can be reproduced, reused, and generalized. In addition, the process model must be scalable to support projects of various sizes. Accordingly, a distinction is made between project activities to be carried out and qualitative requirements for coordinating and organizing the project.

When a process model is developed, it is significant to choose the abstraction level for the contained tasks. Selecting an abstraction level that is too high results in minimal benefits that are limited to the conceptual level. Selecting an abstraction level that is too low makes it more difficult to generalize the model (which is important due to the various applications of data science) and to comprehend the model (which jeopardizes its acceptance).

Dividing the model into levels involving different degrees of abstraction allows it to remain easy to understand and still be helpful for questions about details. At lower levels of abstraction, a modularization can also be useful. This makes it possible to skip over irrelevant model components in the application. As an alternative to a modularization, specialized variants of the process model can arise depending on the domains considered and/or the analytical methods used. This results in the following framework conditions for developing the process model in this document:

- ***Abstraction level***
  At first, a model on a high level of abstraction is developed, which will be suitable for broad use in data science projects. The working group does not initially focus on the final determination and description of all details. Instead, by using the model in real projects, findings should be gained and problems identified that are considered during a continual enhancement process.

- ***Role models and communication***
  Particularly in large projects, the project participants should be able to use the model to identify their own tasks and comprehend those of others. Thus, it is expedient to define groups of people with each group receiving a suitable description and taking on a defined

task spectrum. The process model offers a framework for a standardized understanding of a term to simplify communication between the different groups of people.

- ***Considering the project work and project environment***
  Since data science projects often involve numerous analytical methods, the familiarization period for new topic areas and the testing and discarding of various analytical methods must also be considered. Although under certain circumstances those tasks do not contribute directly to the project success, they are a necessary component of the project sequence. Since data science has consequences on economic, social, and ecological dimensions, those aspects must also be considered in the process model. This can, however, occur in the context of the specific application domain.

3.2 Process models arising from the data science sector

Besides the process models KDD and CRISP-DM mentioned in Chapter 4, whose underlying logic is conceptually connected to data science, other related models must be considered that were developed especially for the data science sector. Primarily in connection with CRISP-DM, however, promising efforts can be identified that should be adapted to the requirements of data science projects, such as the CRISP-ML(Q) (Studer et al., 2021).

The KDD process is clearly comprehensible and features an unambiguous work path from the data source to the knowledge obtained. Moving back and forth along that path is possible, but not necessary in the sense of an iterative procedure. The process is strongly centered on the use of analytical methods. Upstream and downstream tasks, such as building a domain understanding, the deployment of analysis results, and transferring a model into productive operations, are not the focus of consideration.

With CRISP-DM, the possibility of iterating the individual process steps within an analysis project is clearer. However, the model remains simple and easy to understand. The less pronounced differentiation of the individual process steps requires closer cooperation among the groups of participants with less delineation of the individual tasks. The model also contains a process step (*Business Understanding*) strongly focused on the domain, in addition to those related to data, analysis, and evaluation. CRISP-DM was developed from within the industry. The process step of deployment is available, but only minimally pronounced, which can be problematic in light of the data-driven products and services that are common today.

Another model is the Team Data Science Process (TDSP), developed and published by Microsoft. It is described as “an agile, iterative data science methodology to deliver predictive analytics solutions and intelligent applications efficiently“ (Microsoft, 2017). TDSP focuses on data science projects in a business context and includes many aspects that can also be found in other, older process models. In comparison to those models, it explains roles and assigned tasks and adds the areas of IT infrastructure and customer acceptance. This last-named concept emphasizes the domain relevance more strongly than CRISP-DM and also focuses on the corporate context. With respect to the areas of data science, TDSP is more complete than the previously mentioned models, but does not always define the sequence of the process steps in detail. In the documentation, there is often a reference to Microsoft technologies due to the origin of the model and the associated focus on a market segment.

3.3 Key areas of Data Science

Considering the previous discussions and process models named, key areas can be identified on which a process model can be derived and structured. Based on the critical discussion of KDD, CRISP-DM, and TDSP, seven key areas of data science can be deduced. The general context is shown in Figure 5.

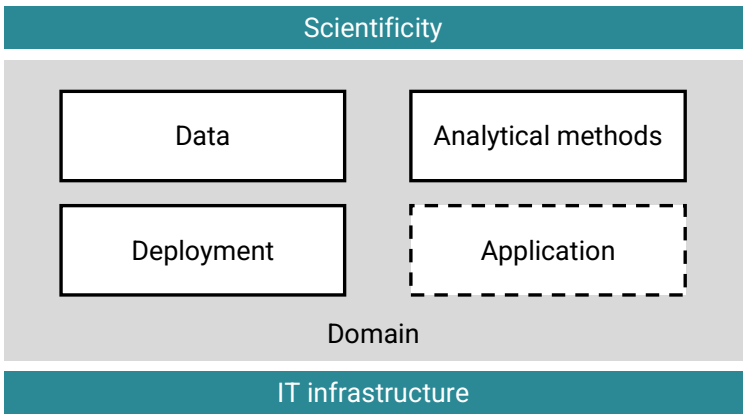


Figure 5: Seven key areas of data science

In the center are the areas of *data* and *analytical methods*, which are also addressed within the context of various process steps by the three process models considered.

Deployment, which is also in focus, is outlined in the KDD, explicitly addressed in CRISP-DM, and especially emphasized in the TDSP through the area of customer acceptance.

The *application* of the analysis artifacts arising from the deployment, to the contrary, is not considered in any of the aforementioned process models. Large parts of this key area are also not seen as a key element of a data science project. However, artifacts that must either be adjusted during the use by data scientists or those that are included in future (further) development projects frequently arise. To take the ambivalent role of use in data science projects into account, it is graphically delineated in the figure showing the previous three key areas and more strongly connected with the domain.

The four named key areas are embedded in the key area *domain*. Considering the definition of the data science term, the domain is not restricted to a business context as it is in the case of other process models.

The described areas are flanked by the overarching key area *scientificity*, which has not previously been considered in any process model but forms a key component of the data science discipline.

The accompanying key area *IT infrastructure* is becoming more important in many data science projects (see Chapter 1). The following sections contain a brief characterization of the seven key areas. A detailed description is contained in subsequent chapters.

Key area “Data”

Data are seen as the “raw materials” of data science (Palmer, 2006). Data are directly connected with numerous work steps that, taken together, frequently form the main area of expenditure in a data science project. Those work steps include procuring, integrating, cleansing, transforming, and storing data. Someone must clarify whether the data needed to fulfill the project goal are available in adequate quantity and quality, whether and how they may be used (data protection), and what structure they possess. However, the data are connected with still another important area: developing a shared understanding of data in the context of the application problem. Accordingly, an important work step is explorative data analysis, possibly including an initial data visualization. Preparing the data for the analytical method used later is another important task, since that method will determine the special requirements made of the data’s form.

Key area “Analytical Methods”

The use of a suitable data analysis procedure is the central step in the data science process, since this (normally) delivers the cornerstone for the gain of knowledge sought. This step is frequently also called modeling, since, by applying analytical methods to data, a model of the examined effectiveness area arises, which can then be used subsequently (to classify new data or to forecast future values, for example).

It is important to select an analytical method and its suitable parametrization that are appropriate for the application and the given data. A broad spectrum of procedures is available, ranging from statistical methods to traditional data mining to neuronal networks, deep learning, and general methods from the area of *artificial intelligence*. If no suitable analytical methods are available, existing procedures must be adapted or new ones might need to be developed. Preparatory steps are the procedure-specific data preparation and feature engineering. An additional task in this key area is the evaluation.

Key area “Deployment”

Deployment justifies the necessary financial expense and time commitment. If the results are later used inadequately, even a theoretically successful project might fail in practice. The simplest (but vaguest) form of deployment is to prepare the results as a final report or publication. If expedient, the goal should be to transfer the developed analysis model into a permanently usable form. Depending on the application and user group, for example, this can be attained through a software system. Hence, numerous professional and technical aspects must be considered, such as the form of preparation for the analysis results or the provision of an adequate user interface.

Key area “Application”

It appears uncontested that the purely operative business—applying the developed analysis artifacts—is not deemed a part of data science projects. But this does not apply to the monitoring of the analysis quality during application with the goal of either using the model or identifying its need for adjustment (“model lifecycle management”). Therefore, application is considered a key area of a comprehensive data science process model.

Key area “Domain”

The four central key areas can be concretized only in the context of the domain, thus the application field. Broad background knowledge of this application field is relevant at many points during the data science process, such as during the identification of a rewarding analysis goal, the correct understanding of data, its origin, quality, and contexts, the evaluation and classification of the analysis results obtained in the context of application, and using the results later in practice. In addition, the assessment of the strengths and weaknesses of existing solutions, the technical requirements analysis, the support in model parametrization, and the final evaluation of the project success are allocated to this area. In conclusion, the legal, social, and ethical aspects of the data science project must be included at this point.

Key area “Scientificity”

A data science project should follow a proven process model and be carried out based on the current state of scientific knowledge. Important aspects include the standardization and structuring of the procedure, suitable project management, and communication to the participating shareholders.

In research and business contexts alike, an adequate working method must be ensured, such as establishing a hypothesis, evaluating applied methods in terms of appropriateness and efficiency, examining the validity of the results, ensuring they can be reproduced, and making sound decisions. It also entails recording new, generalizable knowledge of data sets and methods (as opposed to project-specific knowledge) and publishing the generalizable findings.

Key area ”IT Infrastructure”

Practically all the tasks of a data science project—data management, the actual data analysis, and the evaluation and deployment of analysis results—are implemented with the help of specialized software products. The range and complexity of those products is especially high in larger projects. Providing and operating the necessary IT infrastructure is a demanding task that requires appropriate special IT knowledge (in regard to working with distributed data, cloud connection, sandboxing, scalable architectures, distributed calculation of models, and automation).

4 Data Science Process Model DASC-PM

Building on the previously described drafts, the Data Science Process Model (DASC-PM) is introduced in this chapter (see Figure 6). The visualization of the DASC-PM is derived from the previously identified key areas of a data science project and from the dependencies between them already indicated.

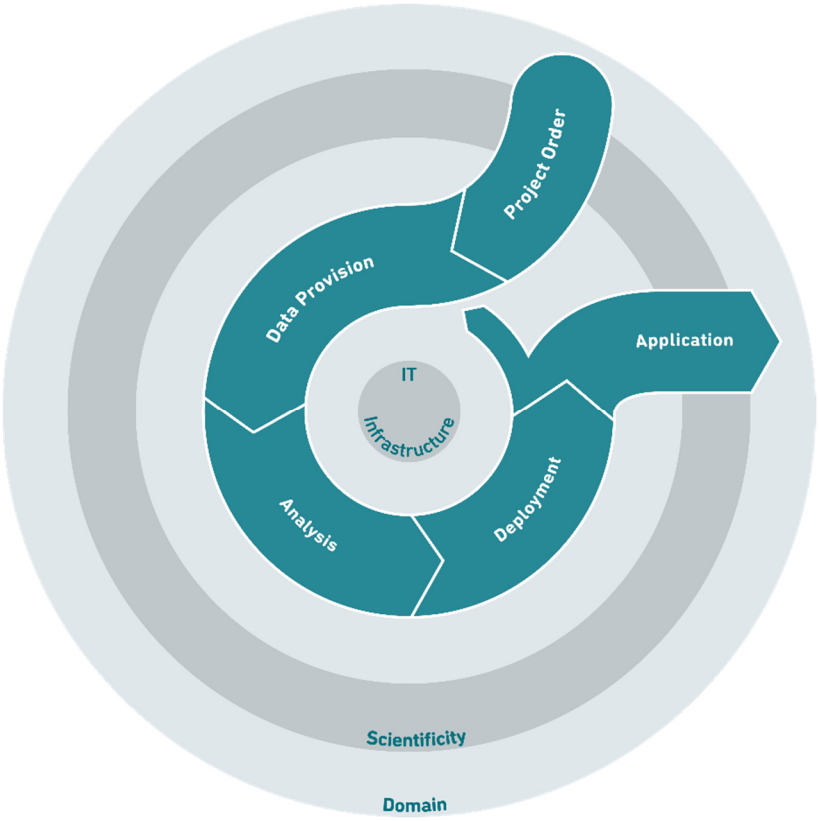


Figure 6: Data Science Process Model DASC-PM

The arrows shown in the model indicate the primary path for using the DASC-PM. They represent individual phases that are derived from the four central key areas and the project order as a phase-similar activity from the overarching key area domain.
By using IT infrastructure and considering scientificity, project phases can be traversed multiple times before an extensive use occurs outside the actual data science project and/or leads to a new/altered project order.

The key areas introduced previously merge into the introduced model. The four central key areas and the tasks derived from them are transferred into iteratively traversable phases. The three general key areas form the framework for this process that arises in this way. The key area domain is given a special position, which forms the broadly encompassing framework of the project and contains tasks that are to be considered like phases and, therefore, will be included separately in the model as a project order. The following list concisely summarizes the components of the DASC-PM.

Phases

- ***Project Order***
  Problems existing within a domain trigger a use-case development. The promising use cases are subsequently configured to a data science project outline. All associated tasks are reflected in the project order phase. Through the early, relatively comprehensive consideration of the project, comprehensive abilities in almost all areas of competence are also frequently required here.
- ***Data Provision***
  Within the data provision phase, all activities that are allocable to the data key area are summarized, which is why the term used is broadly formulated. The phase contains the data preparation (from recording to storage), data management, and an explorative analysis. This phase results in a data source that is suited to further analysis.
- ***Analysis***
  In a data science project, either existing procedures can be used or a new procedure developed—the decision in question is a separate challenge. The phase, therefore, includes not only performing the analysis, but also related activities. The artifact of the phase is an analysis result that has traversed a methodical and technical evaluation.
- ***Deployment***
  In this phase, an applicable form of the analysis results is created. Depending on the project, this can entail comprehensively considering technical, methodological, and professional tasks, or it can be handled pragmatically. The analysis artifact can include results as well as models or procedures and is provided to its target recipients in various forms.
- ***Application***
  Using artifacts after the project performance is not considered a primary part of a data science project. Monitoring is necessary, however (depending on the form of deployment), to check the model's continuing suitability in the application and obtain findings from the application for ongoing and new developments (including developments for the purposes of iterative approaches).

Overarching key areas

- ***Domain***
  Besides the project order as a main component, the explicit requirements or circumstances of the other phases frequently constitute domain-specific framework conditions that influence the tasks. The domain must, therefore, be considered the whole time.
- ***Scientificity***
  Just because data science projects are scientific in nature does not mean they claim to be complete, formalized, academic, or consistently research-oriented in general. Although this might certainly be the case for research projects, the aspect of their scientificity within a

business context primarily refers to a solid methodology: a typically expected characteristic or minimum requirement of scientific work.
The defined project order must be processed methodically in every project phase. Special mention must be made here of the project management and a structured processing that is placed in the foreground by using a process model. Details on the degree of scientificity required must be established while considering the project situation and domain specifics.

- ***IT Infrastructure***
  All the steps that a data science project traverses depend on the underlying IT infrastructure; the actual extent of IT support, however, should be individually assessed for each project. Even if the use of specific hardware and software is frequently determined within the organization, the limiting and empowering characteristics of the IT infrastructure (as well as the possibility of expanding the infrastructure, if applicable) must be considered in all project phases.

Iterations and cancellations during model use

The option to break off the data science project must be considered in every project phase (although we omitted this in Figure 6 to make the figure easy to understand). Although doing so normally precludes attainment of the goal defined in the project order, this does not mean the project has failed. Findings that are collected by the time of the cancellation can be helpful for developing an understanding of the problem or problems considered.

To the extent that iterations are intended, the individual phases must be newly traversed only insofar as the iteration in question is necessary and beneficial. For example, deployment and (project-internal) use of a created model can cause the analysis phase to be newly traversed, possibly to a lesser extent, to improve the model by omitting a new provision of data. Obviously, going back and forth between the phases is also possible during the project if the project team deems it necessary.

Part B
Phases in the model

Notes on Part B

In the following chapters, the individual phases of DASC-PM will be considered in detail. The presentations of the phases and the tasks allocated to them are based on the expertise of the participants in the working group. There is no claim that the content of each phase described is complete; instead, the description should serve to give readers a feeling for relevant aspects within the various project steps.

As an introduction, *Figure 7* shows the nomenclature used both as a structured overview and in an exemplary notation for the detailed description used in the individual chapters to show the connections. The terms shown in italics in the overview are described in greater detail during the phase in question; the terms shown in bold for each phase are explained in subchapters.

Key area

Phase
- Area
  - ***Feature-bearing area***
  - Task
    - ***Core task***
    - ***Accompanying task***
- *Interface (tasks)*

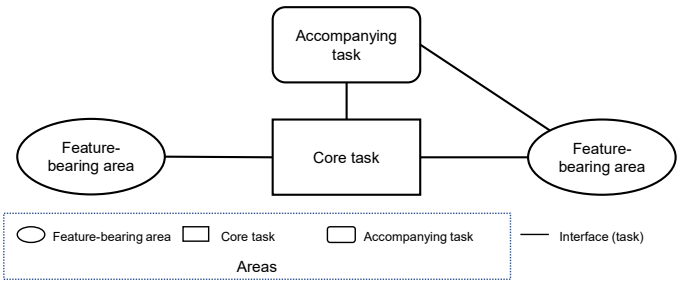


Figure 7: Nomenclature used and notation in the phases

All chapters begin with an introductory and summary page on which the individual areas of the phase in question are listed in a simplified form. A distinction is made between areas based on characteristics and those based on tasks. The latter is further divided into core tasks of the area and accompanying tasks. The connecting borders represent interfaces and contain tasks that arise through the interlocking of the individual areas.

The introductory overview per phase is supplemented by the combined depiction of the profiles of necessary roles and competencies identified, which come into play in the individual areas of the phase. These presentations are based on Figure 3 and Figure 4 as developed in Chapter 2. In the first image, the competence profile of people who specialize in the respective task area are depicted using a radar graph; the second image suggests the role relevance for the respective task area. The images were based on aggregated feedback from participants and should enable a rough estimate of the competence profile for a typical data science project, although the particular manifestations can vary strongly in individual projects and contexts.

Each chapter is followed by a more detailed presentation of the connections between the areas of the phase and its interfaces. Characteristics, core tasks, and accompanying tasks are subsequently detailed in separate subchapters.

5 Project Order

Problems existing within a domain trigger a use-case development. The promising use cases are subsequently configured to a data science project outline. All associated tasks are reflected in the project order phase. Through the early, relatively comprehensive consideration of the project, comprehensive abilities in almost all competence areas are also frequently required.

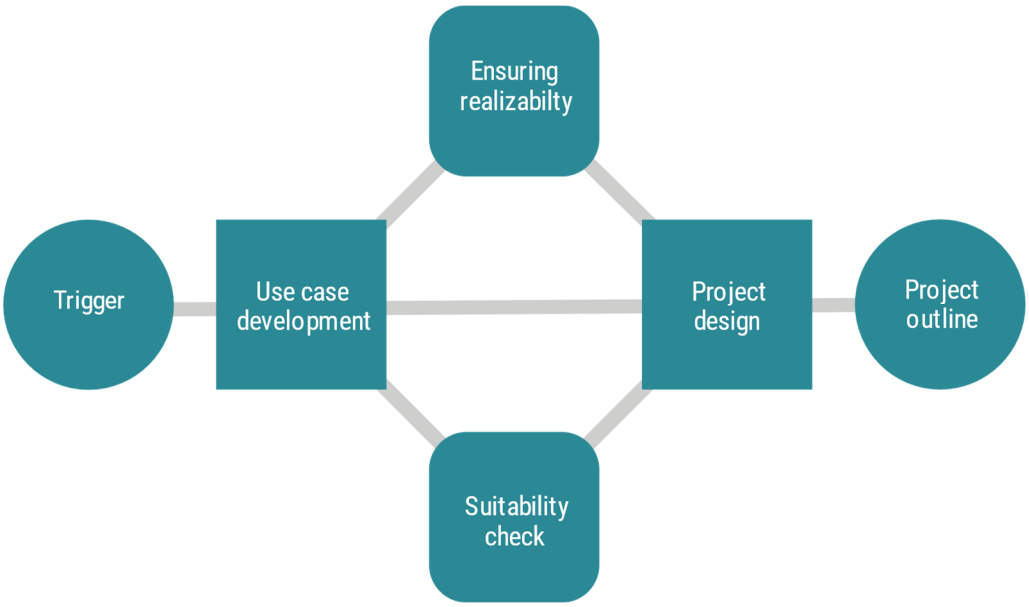


Figure 8: Brief overview of the "Project Order" phase

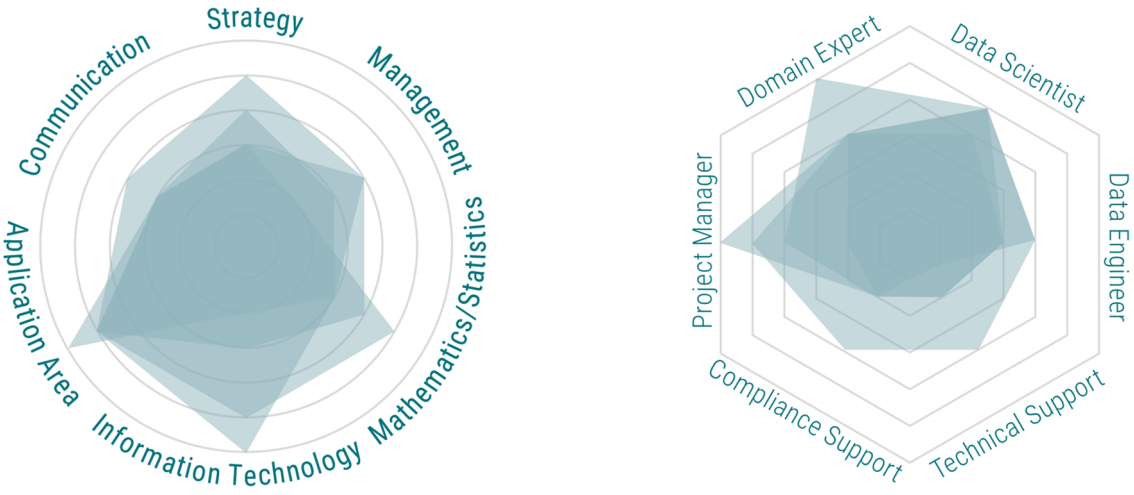
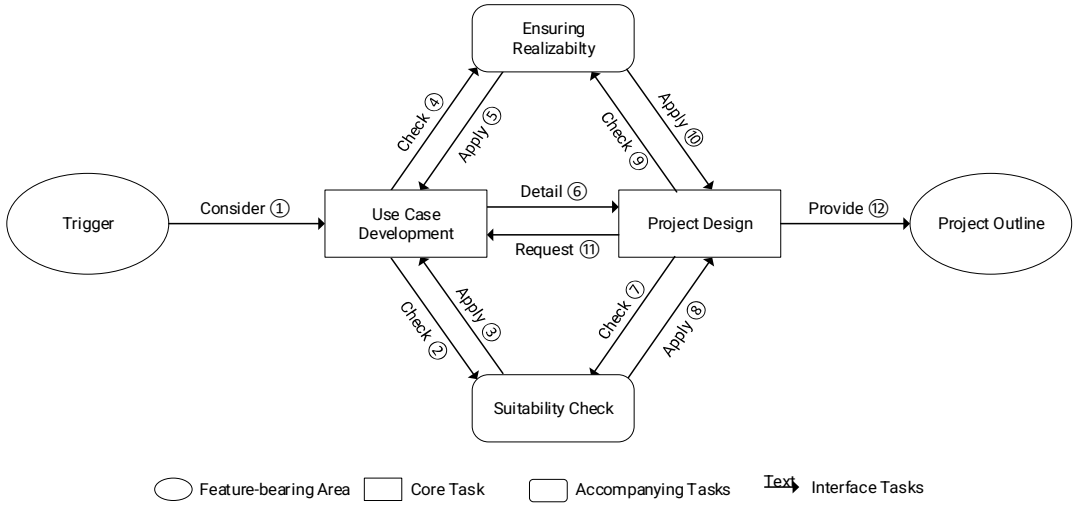


Figure 9: Competence and role profile for the "Project Order" phase

Detailed presentation of the “Project Order” phase



①	Events occurring in the domain trigger the development of use cases.
②	The use case should be checked to determine whether it is suitable for being treated in a data science project.
③	The results of the suitability check should be taken into account while the use case is being developed.
④	While the use case is being developed, its realizabilty during a project should be ensured.
⑤	If special requirements are identified that influence the use case development, they must be considered.
⑥	Selected use cases must be configured for a project.
⑦	The selected use case must be checked to determine whether it is suitable for being treated in a data science project.
⑧	The results of the suitability check must be considered while the project is being designed.
⑨	The realizabilty of the selected use case in a project must be ensured.
⑩	If special requirements are identified that influence the project design, they must be considered.
⑪	If it is recognized during the configuration that a possible project is not useful, a return to the use case development must be made.
⑫	A project outline must be provided as a result of this subprocess.

Figure 10: Detailed presentation of the “Project Order” phase

5.1 Feature-bearing area “Trigger”

The initialization of a data science project is triggered by an event —usually a problem— in the domain. In a researching or explorative context, however, it can also be a matter of open-ended questions that do not constitute a “problem” per se in common parlance.

Table 1 lists and describes characteristics of triggers frequently mentioned by participants.

Table 1: Description of the characteristics of the “Trigger” area

Characteristic	Description
Objective	A decision must be made on how the results of a data science project triggered by a problem will be used later (for example, whether the project objective aims at gaining knowledge or operating models in production).
Functional purpose	Defining the functional purpose of the solution to be worked out can help establish the scope of the project. It is also possible to determine the relevance of a solution to a problem.
Requirements	A description must be made of the requirements for solutions to be worked out.
Participating areas	The areas participating in carrying out the project on the domain side must be specified.
Functional domain	The functional domain within which the problems must be worked on should be described.
Application area	The level of abstraction must be established. Does the solution to be developed involve only recommended actions for the department, or does it involve strategic decisions by an entire group?
Complexity	Only by estimating the complexity of the problem considered can a suitable allocation be enabled.
Alternate courses of action	This concerns both alternatives in the execution of the data science project and alternatives to execution.

A definition and concretization of the problems to be solved in one or more use cases is normally possible only in the subsequent steps, which is why no particular formal requirements are made for formulations or types of documentation at this point. In addition, a specific abstraction level of the descriptions must not be prescribed, since triggers can vary greatly from one another.

5.2 Core task “Use Case Development”

A frequently presented problem when organizations first concern themselves with the discipline of data science is identifying and selecting feasible use cases. In this document, a use case is defined as a self-contained unit that can be deconstructed into work packages. A project can consist of multiple use cases with each one delivering a clear benefit. Ideally, a use case is handled by an iteration of the DASC-PM, but it can also be necessary for multiple iterations to be traversed for one use case.

Table 2 lists and describes the subtasks of a use case development frequently named by participants.

Table 2: Frequently mentioned subtasks of the “Use Case Development” task

Subtask	Description
Develop an understanding of the discipline	The expectations of the data science discipline frequently do not coincide with their possibilities. In addition, data science initiatives frequently lack a clear focus. An understanding of the discipline must be developed.
Use case identification	For one thing, breaking down organizational goals in the use cases that can be considered within a project can be challenging. For another, use cases must result from requirements and problems in an organization. Which results should be expected when the use case in question is carried out often remains unclear. The task consists of developing ideas that are both creative and feasible.
Prioritizing use cases	Selecting suitable use cases for implementation is not always possible, since the potential for one’s own organization might not be directly visible before the project is executed. Input data and (interim) metrics for evaluating alternative use cases (such as those based on profits, possibly from capital investments) might be lacking or imprecise. This means that efforts to prioritize use cases might not be profitable, since they might later turn out to be flawed.
Coordinating groups of participants	The groups of people in an organization that can be helped by the implementation of use cases in the form of projects are often unaware of the benefits that data science can offer them. Conversely, data specialists might not know the most relevant use cases in an organization. Communication and coordination between those two groups of people is, therefore, highly relevant.

The participants had no clear answer to the question of which groups of people or departments in an organization should forward the development of problems into use cases. Domain experts and data scientists were named most frequently, but other groups of people should also be considered.

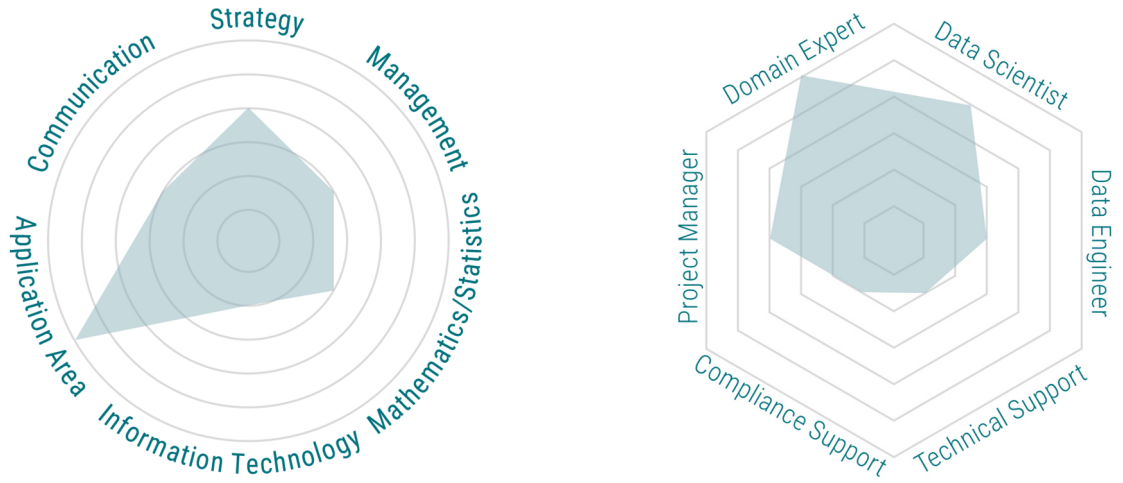


Figure 11: Competence and role profile for the "Use Case Development" task

When selecting suitable use cases, it is important to create a basis of trust by and between the groups of participants. Domain experts should be able to report honestly about difficult, expensive, or challenging tasks without being influenced in doing so by anticipatory solution proposals. It is worthwhile to hold preparatory bilateral conversations to sound out the framework conditions, build up good relationships with the contact person, and prepare interviews or workshops. In so doing, the methods selected must be adapted to the given environment of the organization.

The two formats for identifying suitable use cases mentioned most frequently by participants are interviews and workshops. Interviews are more suited to smaller areas with fewer participants. The interviews should be held by experienced interviewers as a conversation among equals. Workshops are also suited to larger groups and areas. Workshops with many participants are especially used for brainstorming. Use cases can be concretized and prioritized afterward in smaller workshops. To illuminate as many facets as possible, the participant field should be as heterogenous as possible, meaning it should cover various areas (such as domain experts, data scientists, and deciders) and seniority levels. Possible agenda items of such a workshop include presenting exemplary use cases, mutual brainstorming, and evaluating suggestions according to relevance and realizabilty. All this can be backed up with selected data science if appropriate. The workshops should result in selecting and designing a use case that is as specific as possible. Exactly what is to be attained by considering the use case in a project and what business benefits (business value or impact) this might bring about should be recognizable.

Various methods can be used to design the workshop. In general, the methods used should provide a basic structure and a central theme while leaving room for flexible, creative elements. Such a structure allows the workshop participants (especially those having a similar state of knowledge) to examine questions brought about or treated on a mutual scientific basis. The creative element can increase motivation, stimulate various approaches and ways of thinking, and promote group interaction between very different participants.

General methods, such as focus groups, fishbowls, design thinking, and hackathons have been tried and tested for identifying and selecting suitable data science use cases. There are also more specialized methods that are tailored to the context of *data science*, *artificial intelligence*, *machine learning* and *big data*, such as the Enterprise AI Canvas (Kerzel, 2021), the Machine Learning Canvas (Dorard, 2015) and the procedures described by Bill Schmarzo (2015). It is also important to make

sure the methods selected fit the problem and group of people at hand (key word "acceptance," especially for "exotic" methods).

Table 3 lists the advantages and disadvantages of general methods mentioned most frequently by participants.

Table 3: Frequently mentioned advantages and disadvantages of general methods

Advantages	**Disadvantages**
Method: **Focus groups**	
• *Helpful in identifying use cases* • *Helpful in prioritizing use cases* • *Often delivers quick results* • *Useful for unifying perspectives and states of knowledge among the participating groups* • *Incorporates different viewpoints and opinions* • *Promotes the development of new ideas and gives room for spontaneous inspiration* • *Focus groups can ideally develop into task forces*	• *Requires an experienced, neutral moderator (who is, therefore, not a central provider of ideas)* • *Good structuring is needed to attain results (through orientation to leading questions, for example); otherwise, it runs the risk of aimlessness* • *Number of participants is limited (more than twelve often make focused discussions impossible)* • *High requirements for group composition (a diverse group of individuals that still share a comparable state of knowledge; otherwise, individual participants can dominate discussions). Results often strongly depend on who makes up the group.* • *Ideas are frequently only "snapshots" that require more in-depth follow-up considerations* • *Structured follow-up work and analysis (summary transcript, final report, etc.) takes time*
Method: **Fishbowl (inner-outer circle method)**	
• *A larger group of people can participate (active in the inner circle, passive in the outer circle)* • *Discussion participants can be replaced to match different topics, perspectives, and competencies* • *Less pressure on participants, since they can leave the inner circle at any time* • *Well suited to purposefully help the inner circle identify use cases*	• *A core of motivated people is needed for the inner circle* • *Smaller groups in the inner circle can shape the discussion, and the diversity will suffer* • *People in the outer circle might not feel comfortable conversing with those in the inner circle (especially during in-person meetings)*
Method: **Design thinking**	
• *Focusing on the end users' perspective creates the greatest possible benefit and acceptance* • *Interdisciplinary teams make it possible to consider diverse aspects*	• *Design thinking is more of a creative approach for designing products for a certain target group. Application in processing use cases within a data science context is possible without adjustments only in special cases (such as the development of a management dashboard).*

Advantages	Disadvantages
Method: **Hackathons**	
• *Realizes a prototype or a minimum viable product (MVP) in a short time* • *Promotes deep discussions of solution approaches* • *Heterogenous working groups possible* • *Results are frequently very specific and usable* • *Competition incites people to develop innovative solutions* • *Possible (supportive) realization forms for quick wins or proofs of concept*	• *Higher organizational effort* • *Risk of focusing on specific use case; broader viewpoint sometimes neglected* • *Focus on technical implementation, thereby neglecting further-reaching aspects (such as data procurement, compliance, or social consequences)*

In addition, the participants of the working group recommended the following best practices for selecting suitable data science use cases:

- *By **analyzing the organizational or area strategy**, strategically fitting use cases can be identified, allowing those cases to experience a higher degree of attentiveness and support.*
- ***Quick wins** are use cases that deliver specific company benefits and have realistic chances of success while incurring little effort or expense. Quick wins frequently help people who are initially skeptical to become convinced, so they secure resources for subsequent, more expensive projects.*
- ***Lighthouse projects** are "showcase projects" created with great energy and expense. They are meant to illustrate what can be attained with a well-planned and well-executed data science project. Lighthouse projects should be more highly visible, convince people who are initially skeptical (through testimonials, for example), and, therefore, encourage other departments or organizational units to imitate them. Parts of the lighthouse project can be ideally adapted for subsequent projects and reused.*
- *Data science use cases in which technical realizabilty is questionable can be started with a **proof of concept**. To that end, predefined and usually smaller resources (such as time and personnel) are used to prove that the objective sought is generally attainable under the given framework conditions (existing data, available methods, given IT infrastructure, etc.).*

Table 4 lists the most frequently mentioned advantages and disadvantages of best practices.

Table 4: Frequently mentioned advantages and disadvantages of best practices

Advantages	**Disadvantages**
Best practice: **Alignment with the organizational or divisional strategy**	
• *Such alignment is always recommended in principle, but conflicts with the organizational strategy should be avoided at the very least* • *Helpful in maximizing benefits* • *More attention for the project from the company or divisional management (depending on strategy level)*	• *Problematic as the sole criterion under certain circumstances, since this might draw focus to complex, expensive projects.* • *It might be difficult to obtain support from individual target groups*
Best practice: **Quick wins**	
• *Results deliver benefit with little effort* • *Well suited for gaining attention and support for subsequent projects* • *Demonstrates pragmatic, resource-sparing actions* • *Quick success also motivates the project team itself*	• *Limited number of topics that can be realized as quick wins* • *Focusing on reaching an objective quickly might lead to the neglect of other aspects (such as data quality, standardization of the database, and user-friendliness)* • *Risks bringing about data silos or isolated solutions, since a holistic solution is too expensive*
Best practice: **Lighthouse project**	
• *Illustrates the results and advantages attainable through data science* • *Attracts a lot of attention, including across organizational boundaries* • *Can deliver "blueprints" of how data science projects can be optimally carried out*	• *Greater expense (this is visible from the outside only to a limited extent, however)* • *Higher need for resources (such as time and money) can have negative effects if the benefit gained is not great enough*
Best practice: **Proof of concept**	
• *Quick, resource-saving development of a prototype* • *Insightful for further decisions and developments*	• *Might lead to "quick and dirty" solutions, which might nevertheless be used in production later under certain circumstances, since they (or at least their essential features) are functional*

5.3 Accompanying task “Suitability Check”

The suitability check aims to determine whether the identified (and later selected) use cases can be successfully implemented in a project. To do so, a check must be performed to determine whether the established requirements can be met with the available resources. Table 5 lists and describes subtasks for suitability checks frequently mentioned by participants. If the use case considered is selected for implementation in a project, a more detailed examination is made during the project design phase (cf. Section 5.5).

Table 5: Description of the subtasks of the area “Suitability Check”

Subtask	Description
Suitability of the use case	A check must be made to determine whether the use case at hand actually calls for the use of data science.
Suitability of methods	A check must be made to determine whether, as a general principle, analytical methods exist or can be developed that will achieve a suitable result with adequate probability. Initial tests must also be performed to that end, if applicable.
Assessing the data basis	It is often difficult to discern which data are available or can be procured for data science projects, the extent to which they are suitable for use in analyses, and their quality. The effort required to prepare data, and the actual benefit those data will bring, can often be hard to estimate in advance. An initial assessment of the data foundation in this early stage is absolutely necessary.
Suitability of the objective	The expected project results must be compared with the use case considered.
Considering past projects	Past projects must be compared with the use case currently considered.
Prioritizing the use case	Keeping in mind that resources are scarce, a check must be performed to determine whether considering the use case is judicious or whether other problems should take precedence.

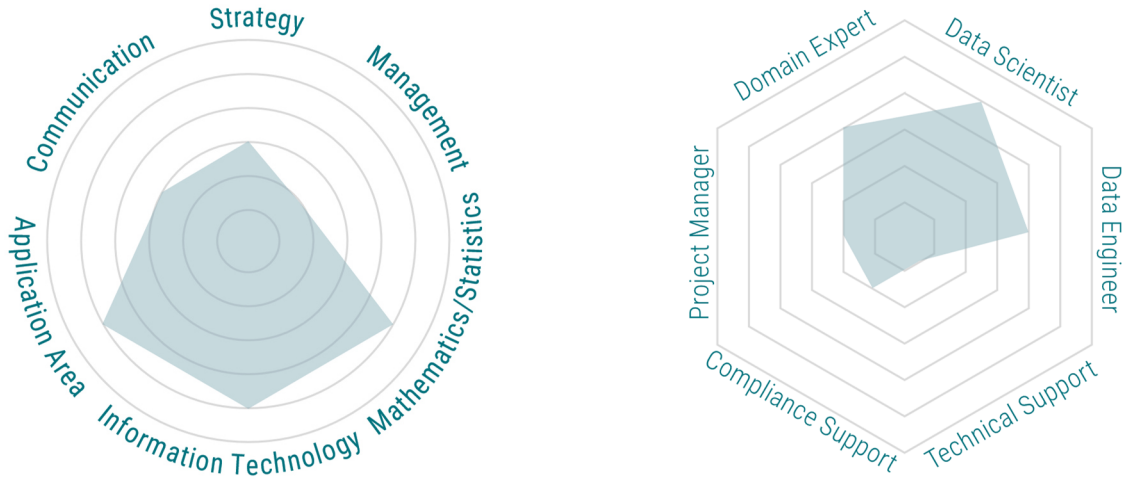


Figure 12: Competence and role profile for the “Suitability Check” task

5.4 Supporting task “Ensuring Realizabilty”

In this step, a check must be performed to determined which project ideas can be implemented. This often entails an iterative process with all interest groups. In some instances when use cases are implemented, it can be uncertain whether the examinations will lead to findings and what those findings will look like. There is often a high percentage of implicit knowledge in organizations, and when a project begins, it may be unclear how that knowledge can be formed into analysis artifacts. In addition, some legal aspects can limit potential use cases. Some of the contact people responsible are unfamiliar with this area and might be overcautious. Furthermore, the expense for implementing use cases and the necessary resources for executing data science projects successfully is frequently underestimated. Table 6 lists and describes the subtasks that participants mentioned most frequently for ensuring realizabilty. If the use case is selected for implementation in a project, a more detailed examination is made during the project design phase (cf. Section 5.5).

Table 6: Description of the subtasks of the area “Ensuring Realizabilty”

Subtask	Description
Checking the IT infrastructure	A check must be made to determine whether the available IT infrastructure is suitable for implementing the considered use case. Alternatively, a check must be made to determine whether other technical possibilities exist, and, if applicable, additional infrastructure can be procured.
Evaluating the expertise	The expertise of the participating individuals must be checked regarding their suitability for implementing the use case considered.
Risk assessment	The risk in implementing the use case during a project must be assessed (probabilities of occurrence for the risk, severity of consequences).
Cost-benefit analysis	Although analyzing the benefit is often difficult, the costs should still be assessed.

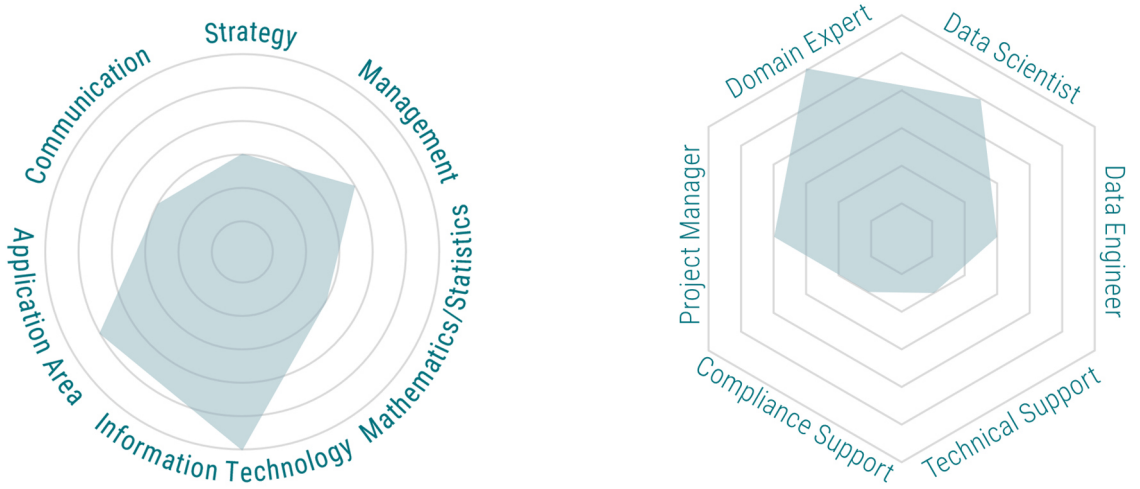


Figure 13: Competence and role profile for the “Ensuring Realizabilty” task

5.5 Core task “Project Design”

Project design aims to determine the necessary work steps that lead to meeting the requirements specified by the use case at hand. This must be done based on the information about the data foundation while incorporating domain specifics. Since the fundamental characteristics of project management are only slightly different from those of other projects, we must refer to the accompanying standard literature. Data science projects, however, possess very specific characteristics. Since their project success is frequently more difficult to assess than project success in other areas, projects must be intensely considered. Under certain circumstances, a distinction must be made between explorative research and development projects and projects specifically targeting an implementation or regular operation.

The questionnaire in the appendix offers assistance in carefully thinking through specific data science project characteristics. If it becomes evident when the selected use case is being processed that it has only limited suitability for implementation during a project, an adjustment must be made (cf. Section 5.2). Parts of the questionnaire can be reconsidered during the project execution whenever new information becomes available.

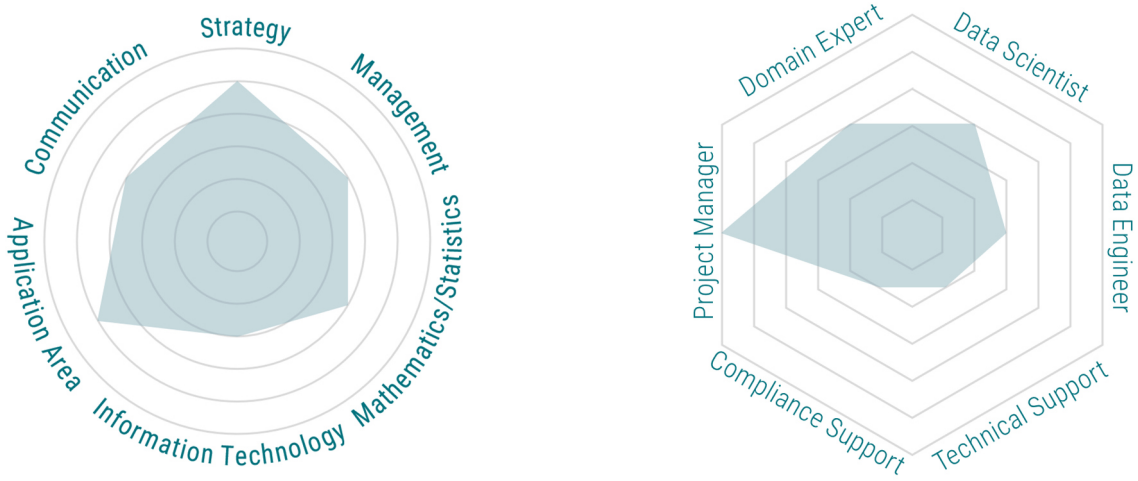


Figure 14: Competence and role profile for the “Project Design” task

5.6 Feature-bearing area “Project Outline”

Unlike many other disciplines, the course of data science projects cannot normally be completely planned and described. Therefore, this subprocess can bring about only a project outline that must be continually reconfigured during the project flow. This can initially result in the creation of a backlog (or comparable collection) of sought-after functionalities and structures for desirable solutions, especially in agile process models like Scrum or Kanban. The term “project outline” used here can be adjusted or transferred to the approach accordingly.

In any case, when the project is being described, it is important to choose an abstraction level that can be used to concisely present all relevant requirements and information from a data, domain, and analysis viewpoint.

Moreover, the project description should outline the work steps and consequences that can already be identified and that help fulfill the established requirements. If changes occur during project execution, the project outline must be adjusted accordingly.

6 Data Provision

Within the data provision phase, the activities that are allocable to the data key area are summarized, which is why the term used is broadly formulated. The phase contains the data preparation (from recording to storage), data management, and an explorative analysis. This phase results in a data source that is suited for further analysis.

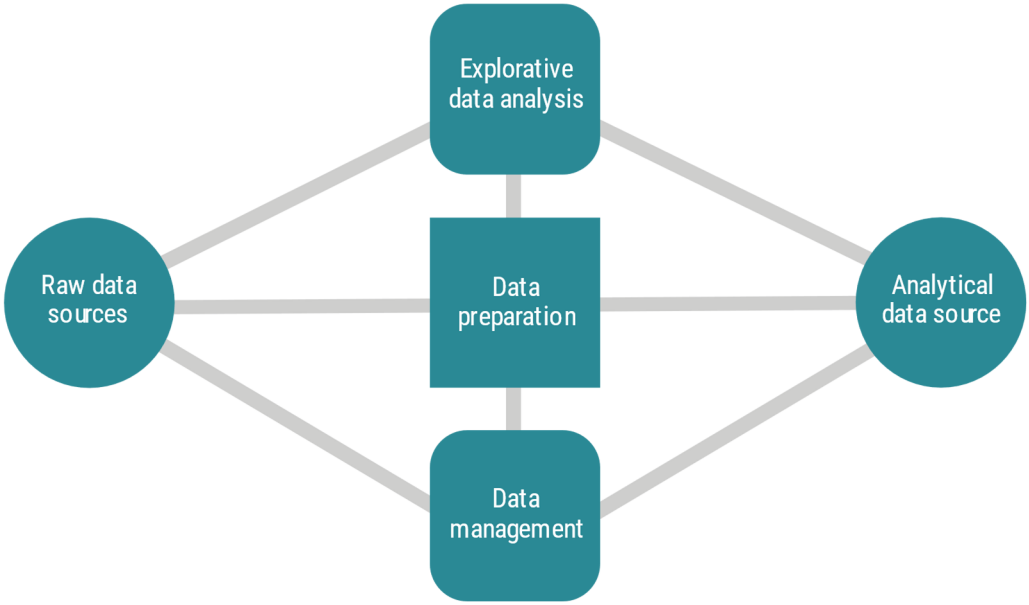


Figure 15: Brief overview of the "Data Provision" phase

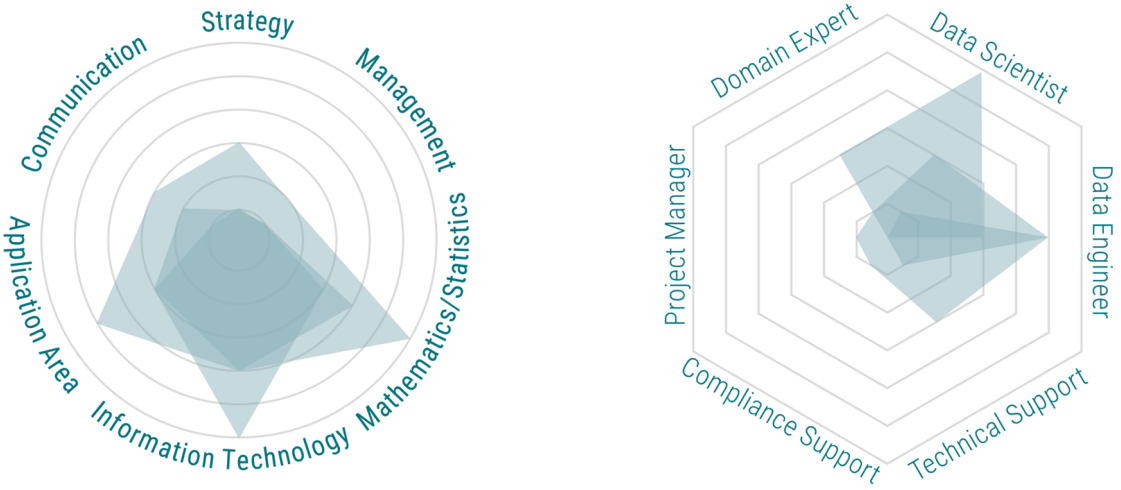
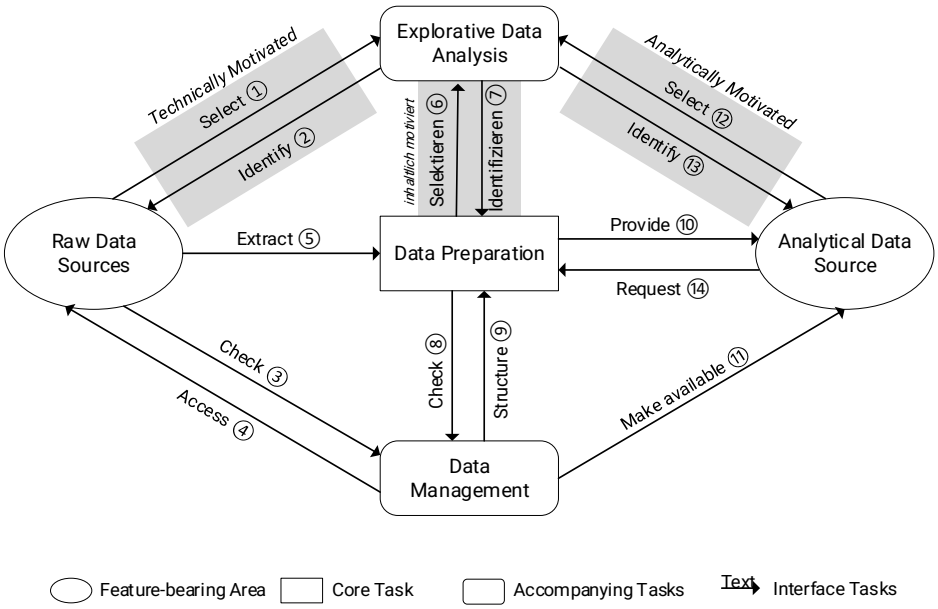


Figure 16: Competence and role profile for the "Data Provision" phase

Detailed presentation of the “Data Provision” phase



①	The data are selected directly from the sources so their properties can be examined. At this point, the data needed to carry out the data science project have already been generally identified. It is also conceivable, however, that checking existing data sources will help generate new ideas.
②	The relevant data are identified so their fundamental properties can be subsequently analyzed.
③	Data management strategies (regarding degree of detail, previous calculations, etc.) are examined and positioned based on the characteristics and functions of the raw data sources.
④	Proper access to the data sources that suits the problems in the data science project is ensured.
⑤	Potentially relevant data are extracted from the sources based on the findings gained in ② and the data access established in ④.
⑥	Data are selected for analysis to investigate characteristics, establish preparation strategies, or validate preparation results.
⑦	The explorative data analysis identifies useful preparation strategies.
⑧	The data management requirements in the data science project are checked.
⑨	Data structuring requirements are considered for the data storage architecture of the current project and the existing IT infrastructure.
⑩	The prepared data are provided for the application of analytical methods.
⑪	The data are stored in a suitable data model, provided via a virtualization layer or as a stream, and possibly protected and archived.
⑫	The data are selected to be evaluated for an explicit analysis project.
⑬	The relevant properties for an explicit analysis project are identified.
⑭	The data preparation based on the findings identified in ⑫ are adjusted.

Figure 17: Detailed presentation of the “Data Provision” phase

6.1 Feature-bearing area “Raw Data Sources”

Data are not usually collected for analytical tasks. If recourse is taken to existing data sources, an understanding must first be established for the process, recording, and the framework conditions under which those sources came about. That metadata must be documented in a suitable form and allocated to the data set. Metadata from multiple data sources should ideally be managed in a metadata repository so that the data sources can be used for the long term, including in other projects.

Characteristics of *data (sources)* can be divided into four categories that can be relevant for a data science project, as shown in Table 7. This presentation does not aim to offer a comprehensive checklist of all conceivable characteristics, but to make it easier to structure an approach to taking a rudimentary inventory. An extensive consideration of the possible characteristics of data sources can be found, for example, in Jayawardene et al. (2013). Despite its scope, the enumeration of the data quality criteria shown in Table 8 and taken from the previous source does not claim to be complete. The relevance of the individual characteristics must be evaluated individually for each project.

Table 7: Description of the categories of characteristics in the “Raw Data Sources” area

Characteristic category	Description
Procurement effort	The availability of data can strongly impact which analyses are performed. For example, if data are already available within the organization and can be loaded automatically, this will incur far less effort than using external data that must first be collected, purchased, or located.
Administrative effort	Various forms of data storage can be necessary depending on the quantity, speed of change, and confidentiality of the data. Another relevant characteristic is whether the data must be accessed only once or more often.
Processing effort	The way in which the data must be transformed to be usable for analyses is influenced by granularity, redundancy, and structuring, and by the preprocessing already performed in the source systems, among other things.
Data quality	The qualities of the data possess depend on how up—to-date they are, the percentage of missing or incorrect values, their relevance to the data science project, and other factors. Producing an accurate picture of the data’s quality requires knowledge of their origin and collection process as well as an explorative data analysis before applying complex analytical methods.

Table 8: Frequently mentioned data quality criteria, taken from Jayawardene et al. (2013)

Data quality criteria (cf. Jayawardene et al. 2013)			
Ability to Represent Null Values	Access Security	Accessibility	Accessibility and Clarity
Accessibility Timeliness	Accessible	Accuracy to Reality	Accuracy to Surrogate Source
Accuracy	Accuracy / Validity	Allowing Access to Relevant Metadata	Applicability
Appropriate Amount of Data	Appropriateness	Authority	Availability
Believability	Business Rule Validity	Clarity	Coherence
Cohesiveness	Comparability	Complete	Completeness
Complexity	Comprehensiveness	Concise Representation	Conciseness
Concurrency of Redundant or Distributed Data	Conformance	Conforming to Metadata	Conformity
Consistency	Consistency and Synchronization	Convenience	Correct Interpretation
Correctness	Credibility	Currency	Currency / Timeliness
Data Coverage	Data Decay	Data Integrity Fundamentals	Data Specifications
Definition Conformance	Derivation Validity	Document Standardization	Duplication / Non-duplication
Ease of Understanding	Ease of Use and Maintainability	Enterprise Agreement of Usage	Equivalence of Redundant or Distributed Data
Fact Completeness	Flexibility	Flexibly Presented	Format Precision
Informativeness / Redundancy	Integrity	Interactivity	Interpretability

Maintainability	Mapped Completely	Mapped Consistently	Mapped Meaningfully
Mapped Unambiguously	Naturalness	Null Values	Objectivity
Perception Relevance and Trust	Perceptions	Phenomena Mapped Correctly	Portability
Precision	Precision / Completeness	Presentation Clarity	Presentation Media Appropriateness
Presentation Objectivity	Presentation Quality	Presentation Standardization	Presentation Utility
Properties Mapped Correctly	Provenance	Record Existence	Referential Integrity
Relevance / Aboutness	Relevance / Relevancy	Reliability	Representation Consistency
Representation of Null Values	Reputation	Secure	Security
Semantic Consistency	Semantic Definition	Signage Accuracy and Clarity	Source Quality and Security Warranties or Certifications
Speed	Structural Consistency	Structured Valued Standardization	Suitably Presented
Timeliness	Timeliness and Availability	Timeliness and Punctuality	Timely
Traceability	Transactability	Type-sufficient	Ubiquity
Unambiguity	Understandable	Understood	Uniqueness / Unique
Usability	Valid	Validity	Value Added
Value Completeness	Value Existence	Value Validity	Verifiability
Volatility			

6.2 Core task “Data Preparation”

Data preparation generally entails transferring the data extracted from one or more source systems in a suitable format for the analytical method to be applied. Another main goal is increasing data quality.

Processing large amounts of data requires the use of powerful hardware and software and some innovative procedures. This is addressed in the key area *IT infrastructure*. Scripts that can be automated, but that document the process and make it repeatable in any case, emerge as possible artifacts of the data preparation. Performing these scripts results in a prepared database that is suited to the data science project and considers the aforementioned tasks. Documenting the preparation steps is just as necessary as documenting the characteristics in a data catalog.

Table 9 (see next page) lists and describes the subtasks frequently mentioned by participants that must be performed in data science projects.

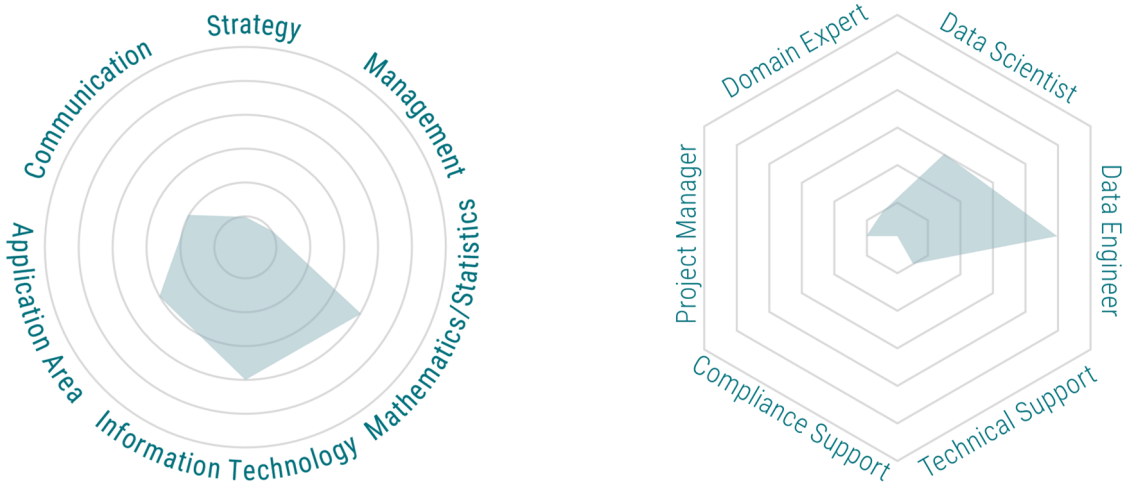


Figure 18: Competence and role profile for the “Data Preparation” task

Table 9: Frequently mentioned subtasks of the "Data Preparation" task

Subtask	Description
Generating characteristics	Additional or alternative characteristics can be derived from existing data.
Data anonymization	If confidential data (such as personal data) are needed in data science projects, they must first be anonymized or pseudonymized as applicable.
Data aggregation	If data possess a high degree of detail, they must be aggregated.
Data annotation	Annotating characteristics is necessary to use supervised learning procedures, among other reasons.
Data cleansing	Identified errors or missing values can be cleansed manually or automatically as applicable. If this is impossible, data filtering or dimensional reduction must be checked.
Data filtering	Unnecessary or inaccurate data must be removed from the database.
Data integration	Data from various sources must be merged and harmonized.
Data structuring	Depending on the analytical methods to be applied, unstructured data must be structured in advance. To that end, for example, methods of natural language processing or image recognition can be used.
Data transformation	Transformations must be performed to prepare data for analysis. This contains both the need for transformation identified during the explorative data analysis and transformations from a more technical viewpoint driven by data management.
Dimensional reduction	Irrelevant or redundant material should be removed from the database.
Creating data preparation plans	Before data are prepared, preparation plans must be created based on data requirements.
Format adjustment	As a rule, source formats have not been defined primarily for the use of analytical methods. Therefore, a transfer into a suitable format is frequently necessary here.
Logging the data preparation	All data preparation steps must be logged. This is important for making sure the project results are representative and can be reproduced, among other reasons.
Process automation	If data must be prepared repeatedly, regarding or based on the application of various analytical methods, the preparation process can be fully or partially automated.
Schema integration	Schemata from various sources must be merged and harmonized.

6.3 Accompanying task “Data Management”

Data management focuses on making the necessary data available without formulating requirements for an IT infrastructure in doing so. Table 10 lists and describes subtasks of *data management* frequently mentioned by participants. As an artifact, an expansion emerges in the data catalog for the traceability of the data management.

Table 10: Frequently mentioned subtasks of the “Data Management” task

Subtask	Description
Data archiving	If analytical methods must be reproducible based on identical data and this option is not ensured by the source system, the data used must be archived. Hence, both technical challenges and topics (such as copyrights) that can preclude permanent storage should be taken into account.
Data protection	Depending on the data used, it might be necessary to protect them from unauthorized access or, if applicable, store them exclusively in anonymized or pseudonymized form while allowing for various access roles and access rights.
Backing up prepared data	A check must be made to determine whether the prepared data must be backed up while the data science project is being executed or whether the scripts developed during the *data preparation* task can be automatically restored.
Storing raw data	A check must be made to determine whether the raw data for the project are backed up separately. If data accumulate or are continually added during the course of the project, suitable processes and infrastructures must be provided for.
Data access	Data can be loaded and processed either once, in defined intervals through batch processing, or in (near) real time as a stream. In the context of open science, third-party access to the data can also be granted if applicable.
Meta data management	Metadata that are extracted from the sources, or supplemented or determined through the tasks performed, must be managed sensibly.

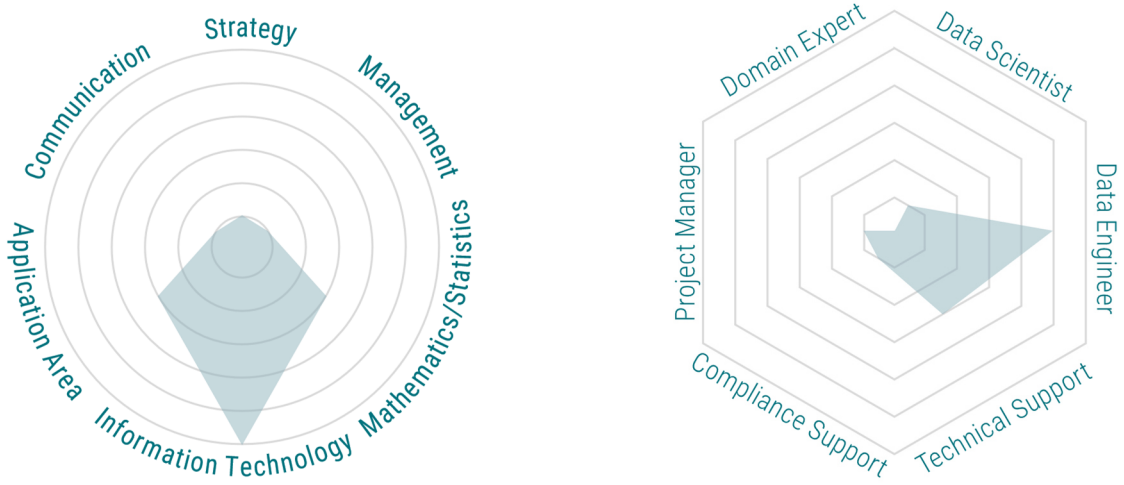


Figure 19: Competence and role profile for the “Data Management” task

6.4 Accompanying task “Explorative Data Analysis”

Explorative data analysis entails obtaining a better understanding of the content of the available data as well as possible points of connection for determining subsequent, deeper analysis. It must be determined whether the quantity and quality of the available data is adequate for the problem selected and whether the planned analysis will require additional data preparation steps. Table 11 lists and describes subtasks of explorative data analysis frequently mentioned by participants.

Table 11: Frequently mentioned subtasks of the “Explorative Data Analysis” task

Subtask	Description
Identifying outliers	Outliers can strongly influence the subsequent analysis. A decision must be made on whether the identified outlier corresponds to real data points or whether other effects have arisen. Accordingly, those values must be filtered out or replaced.
Data validation	Domain knowledge can be used to identify values in data sets that might be unobjectionable from a formal standpoint but whose content is incorrect or not useful.
Data visualization	Simple diagrams (such as histograms, line diagrams, and plot diagrams) can be used to clearly show the distribution of available data and discover simple connections between attributes.
Identifying central attributes	The subsequent data analysis can be performed more efficiently if the data sets possess fewer attributes. Therefore, the goal is to identify as many central, meaningful attributes as possible while excluding insignificant ones. Recourse is frequently taken to knowledge of domains and statistics.
Understanding content	Data must be evaluated regarding their suitability in the specific domains, while allowing for the objectives of the current data science project.
Statistical analysis	Simple statistical measurements such as the median, mean value, standard deviation, and correlation help to better understand the available data and detect unexpected deviations.
Examining the necessity of data transformations	To guarantee comparability of attributes, a standardization of the data is frequently necessary. Another reason for transformations is the subsequent analytical methods to be used, which frequently presume certain data characteristics. Identifying the transformation tasks is part of *explorative data analysis* and the implementation is the responsibility of the *data preparation* area.
Examining missing values	If data sets are missing attribute values, a decision must be made on whether those data sets or the affected attributes can be deleted. Since this can influence the quantity, representative power, and informative value of the underlying data, a use of the missing values is also conceivable. Identifying suitable procedures for handling missing values is part of *explorative data analysis*, and the *data preparation* area is responsible for taking appropriate measures.

Since aspects that are still unknown must be detected during explorative data analysis, there is no fixed sequence for the procedures to be applied. Besides data visualization, various statistical methods are usually used, such as correlation, factor, and cluster analyses as well as statistical and possibly causal modelings. The visualizations and models that arise constitute the artifacts accordingly. Any problems in data quality and changes needed in the data materials must be documented. That being said, documentation of explorative data analysis is not usually highly detailed, since many hypotheses must be explored and possibly discarded in quick succession.

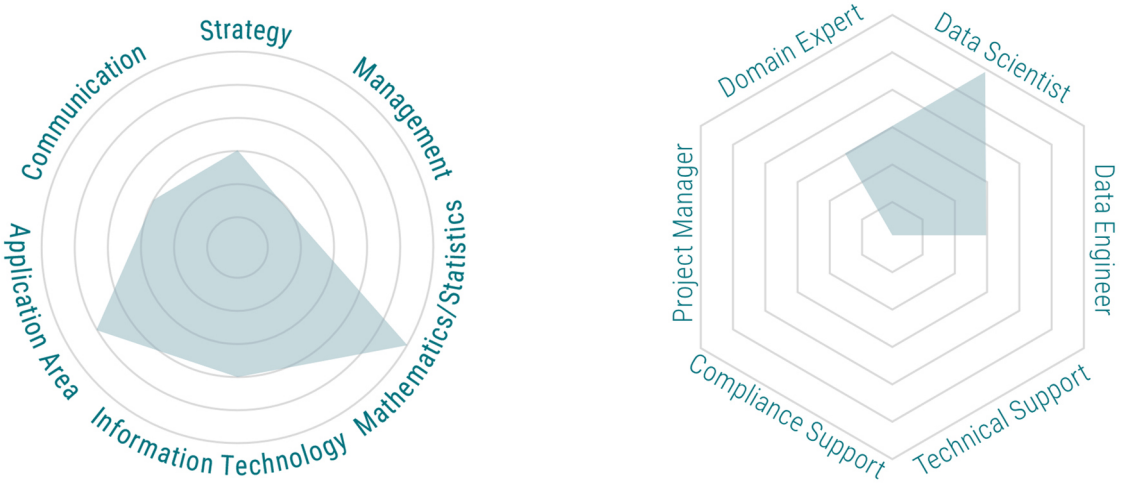


Figure 20: Competence and role profile for the "Explorative Data Analysis" task

6.5 Feature-bearing area “Analytical Data Source”

Although the main characteristics of raw data sources and analytical data sources are similar, details in their respective content, scope, structure, and format differ due to the preparation of the data for data science applications.

For example, regarding the analysis objective, attributes are sought that are cleansed and as free of redundancy as possible. Moreover, the attributes should be especially relevant to the analysis objective. However, evaluating such relevance is not always unequivocally possible in the early stages of a data science project, so an assessment by domain experts is recommended. Data formats and scale levels must be adjusted depending on the analytical methods to be applied. Many learning procedures exclusively process numerical attributes, for example.

Analytical data sources can usually be processed independently by groups of project participants. Hence, data access can occur in real time, continually, or once. Metadata from the data sources can be provided as a supplement.

7 Analysis

In a data science project, either existing procedures can be used or new procedures must first be developed—deciding which course to take is its own challenge. This phase, therefore, includes not only performing the analysis but also related activities. The artifact of the phase is an analysis result that has traversed a methodical and technical evaluation.

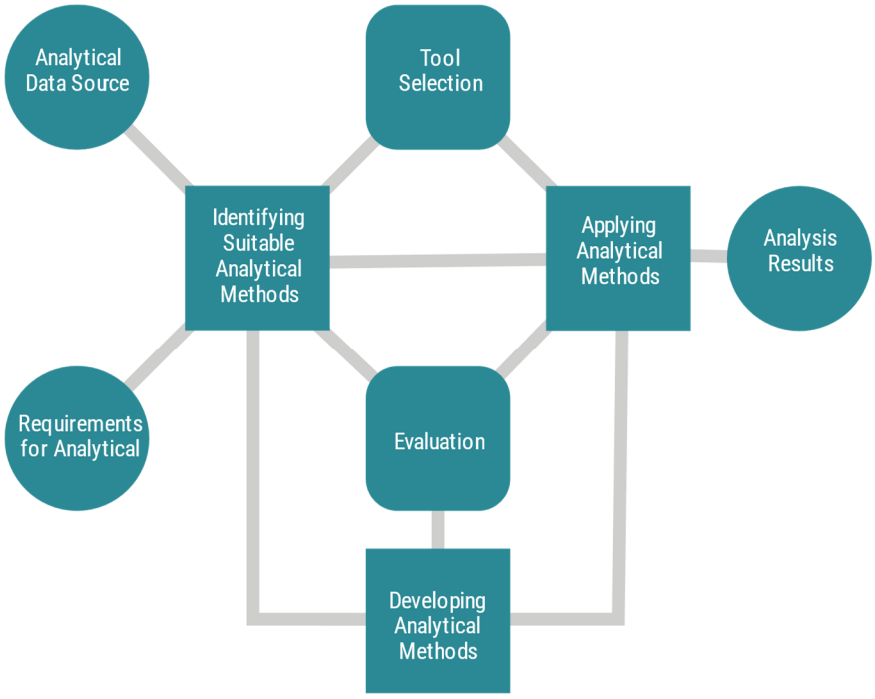


Figure 21: Brief overview of the "Analysis" phase

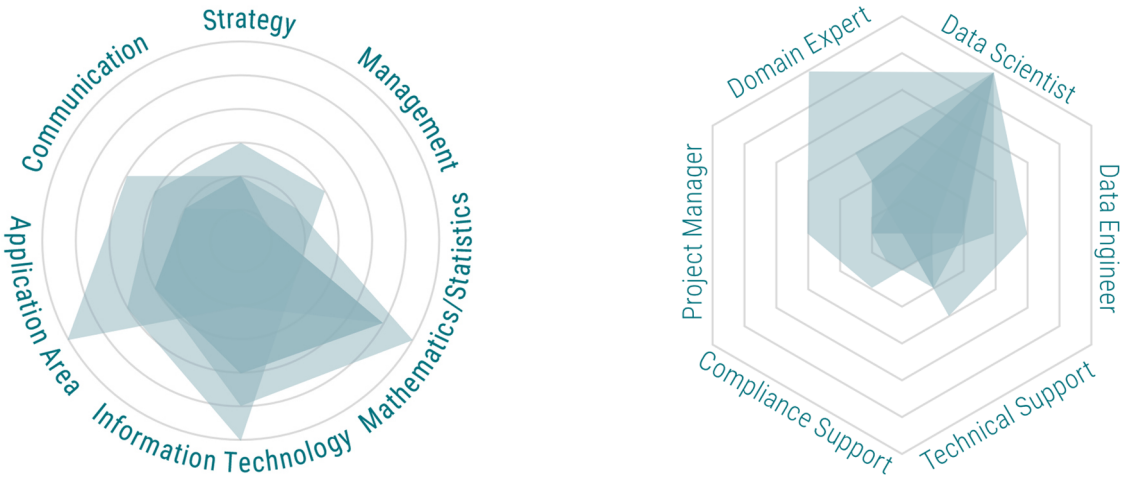
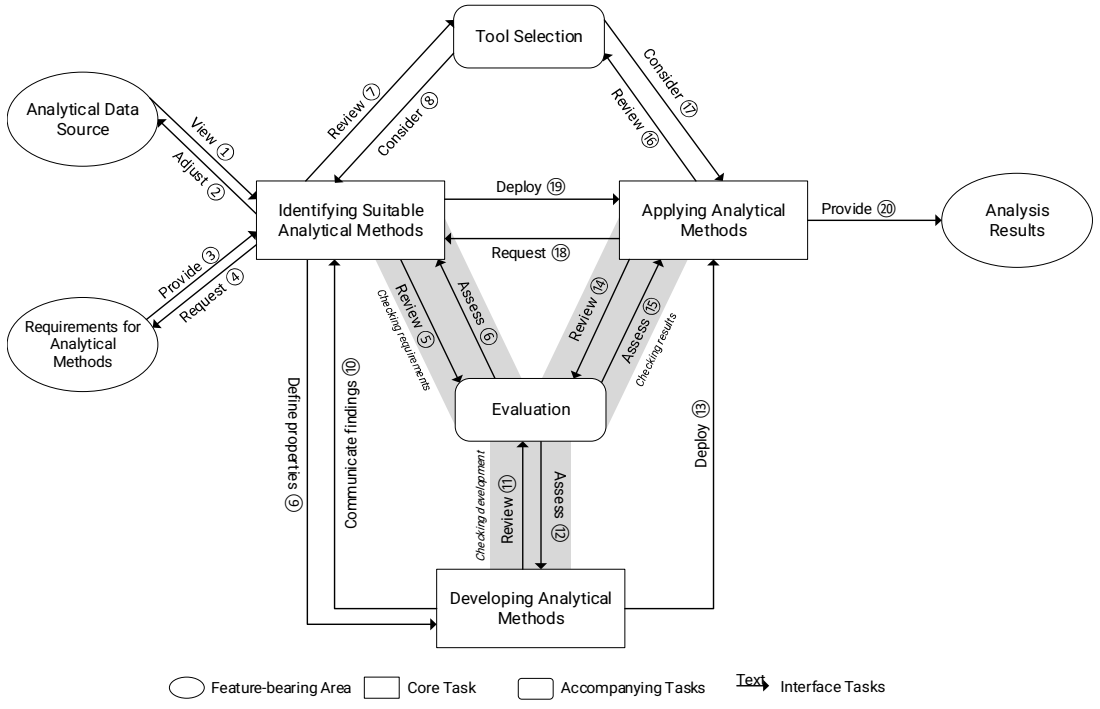


Figure 22: Competence and role profile for the "Analysis" phase

Detailed presentation of the “Analysis” phase



①	The analytical data source is created by handling the tasks described in the key area *data*. Suitable analytical methods can be identified only if characteristics of the available data are considered.
②	After potentially suitable analytical methods are identified, they might need to be adjusted to ensure applicability.
③	When suitable analytical methods are being identified, the defined nonfunctional requirements must be taken into account.
④	If no suitable analytical methods can be identified, it might be useful or necessary to adjust the established requirements.
⑤	Selected procedures must be evaluated to determine whether prescribed analysis requirements can be met.
⑥	The results of the evaluation must be reflected on when suitable analytical methods are being identified.
⑦	The selection of suitable tools must be checked while considering the analytical methods identified.
⑧	The results of the tool selection must be considered when suitable analytical methods are being identified.
⑨	In special cases, analytical methods must be developed. While this is being done, the requirements considered when analytical methods were identified must be taken into account.
⑩	If the step of developing analytical methods fails and this fact does not cause the project to be aborted, the findings gained must be reflected on when suitable analytical methods are identified again.
⑪	The suitability of the analytical methods must be continually evaluated during the development.
⑫	The findings from the evaluation need to be pondered while the analytical methods are further developed.

⑬	The developed analytical method(s) must be provided for application.
⑭	When analytical methods are being used, various parametrizations must be evaluated.
⑮	The results of the evaluation must be considered when analytical methods are being applied.
⑯	The selection of suitable tools for the application of analytical methods must be checked.
⑰	The results of the tool selection need to be considered when analytical methods are being applied.
⑱	If applying analytical methods fails to deliver acceptable results, the process must be aborted or returned to the step of identifying suitable analytical methods.
⑲	Suitable analytical methods can be applied.
⑳	If applying analytical methods leads to acceptable results, they can be provided for the deployment.

Figure 23: Detailed presentation of the "Analysis" phase

7.1 Feature-bearing area “Analytical Data Source”

Identifying suitable analytical methods is based on the characteristics of the available analytical data source (cf. section 6.5).

Special requirements are often made of the data source due to the analytical problem considered or the analysis results required, and vice versa, unchangeable characteristics of the analytical data source can also restrict the number of answerable questions.

During the *analysis* phase, no special features must be recorded beyond that extent.

7.2 Feature-bearing area “Requirements for Analytical Methods”

The characteristics in this section represent the nonfunctional requirements for analytical methods. In the individual project, they can also be seen with explicit limit values already and used as specification requirements.

Table 12 lists and describes characteristics of requirements for analytical methods frequently mentioned by participants.

Table 12: Frequently mentioned characteristics of the “Requirements for Analytical Methods” area

Characteristic	Description
Requirement coverage	The analytical methods selected cannot always fulfill all application requirements completely. However, a high degree of coverage is desirable.
Efficiency	The procedure must be able to be applied to the IT infrastructure at an appropriate time. The less data and research time required, the simpler it is to integrate the procedure into ongoing operations and the more economical it is to use.
Innovative solution	The procedure must solve a problem that has not yet been solved to the same extent or with the same quality by existing procedures.
Replicability	For the result to be replicable (by others) and the procedure usable (ideally in different scenarios), technologies and algorithms must be used that are extensively documented and generally available.
Robustness	The procedures used should be as error-proof as possible. For example, it is helpful if inaccurate data or outliers are recognized automatically or only minimally influence the result.
Scalability	In practice, the quantity and/or dimension of the data to be newly analyzed frequently increases considerably over time. It is, therefore, advantageous if the procedure selected can also process a growing amount of data with reasonable added efforts.
Implementability	The procedure must be implementable with available resources (such as technical infrastructure and specialized personnel). Furthermore, implementing it should require as little effort as possible.
Validity	The forecasts or derived structures must reliably reflect the reality of the situation as accurately as possible. The acceptable error tolerance depends on the problem.
Comprehensibility	The results of the procedure should be traceable, if possible, and easy to communicate and/or visualize.

7.3 Core task “Identifying Suitable Analytical Methods”

Before this task begins, it should already be clear that the question at hand can actually be answered with the help of data science; this means that it constitutes a potentially solvable problem but is not so trivial that it could be solved with, say, the help of a standard report. Identifying suitable analytical methods is often challenging. Although many analytical methods exist, there is a chance that none will suit the problem at hand. In that case, a check must be made to determine whether certain project framework conditions can be changed, whether the development of a new analytical method is conceivable, or whether the project might need to be aborted.

This phase focuses on gaining an overview of existing procedures and identifying the best procedure for the application. Since a final selection cannot be made without further evaluation, multiple procedures for additional assessment should be considered first. The decision to develop new procedures should be made under consideration of the expense and existing uncertainty.

Table 13 lists and describes subtasks for identifying suitable analytical methods frequently mentioned by participants.

Table 13: Frequently mentioned subtasks for the task of “Identifying Suitable Analytical Methods”

Subtask	Description
Identifying requirements	Before various procedures are tested, it must be clear which problems they are to solve.
Determining the problem class	The requirements identified can usually be used to assign the problem to a specific problem class, which can then guide the search for a specific analytical method.
Researching comparable problems	When searching for suitable analytical methods, it helps to research whether there are publications about similar applications.
Determining potentially suitable procedures	In light of the problem class, and based on the research into comparable problems, promising analytical methods or their variants can now be named in principle.
Selection	After the procedures that come into question have been listed, the ones that best fit the project-specific criteria and resources should be selected.

As an artifact to this phase, a list of analytical methods emerges that also contains justifications for why those procedures suit the situation at hand. If no suitable analytical methods are identified, procedures can be selected even if they must be enhanced; a prototype might even be created to ensure the selection will be suitable.

The findings from this phase must be documented in a way that not only justifies the selection for the current project but also sets forth the decisions in a form that can be used for future issues.

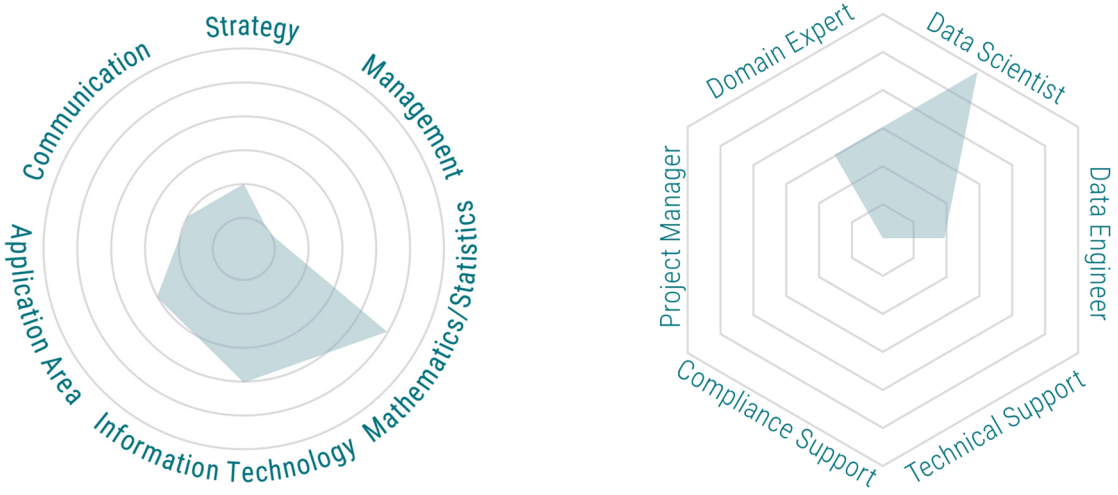


Figure 24: Competence and role profile for the task of "Identifying Suitable Analytical Methods"

7.4 Core task “Applying Analytical Methods”

Applying analytical methods correctly requires detailed knowledge of existing procedures. Using procedures incorrectly can lead to arbitrary results, which will result in inaccurate or false statements.

Applied procedures must fulfill the respective tasks in a suitable manner. This must also play a major role in identifying the procedure (see the previous section). However, this cannot be ensured until the procedure is actually applied to the data to be analyzed. The goal is to find the best analysis result. In detail, this depends on the procedure used and the individual domain requirements. With a few procedures, deciding between aiming for a result that is as exact as possible or a model that is applicable to as many scenarios as possible is important.

Table 14 (see next page) lists and describes subtasks for applying existing analytical methods frequently mentioned by participants.

A large portion of the emerging artifacts and required documentation depends on the individual project, meaning it is inseparably connected to the problem to be solved, the data used, and the analytical methods applied. Basically, what emerges as artifacts is a documentation of the analysis execution and evaluation results (including interim results and graphics), a justification of the selection for the final model, a securing of the development environment, an interface documentation, and the parameter configurations. Technical information should be prepared so the domain experts can easily understand it and should include notes on what errors and anomalies there were and which additional problems were investigated with the help of the analytical methods. Depending on the tools used, basic documentation is created even during the analysis procedure itself.

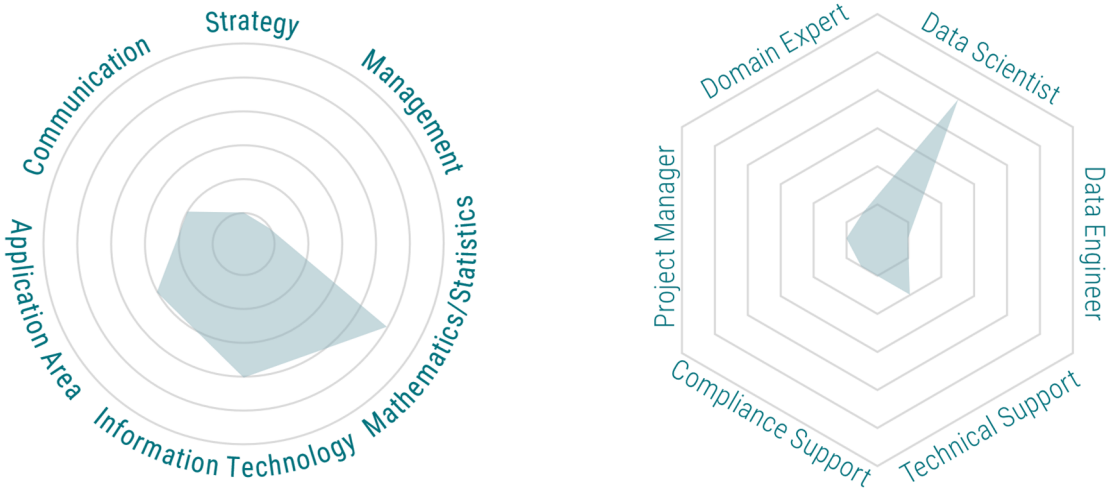


Figure 25: Competence and role profile for the task of “Applying Analytical Methods”

Table 14: Frequently mentioned subtasks for the task of "Applying Analytical Methods"

Subtask	Description
Setting up a development environment	Especially if multiple users are involved, a powerful, easily accessible development environment with version management should exist to guarantee the data science project progresses smoothly over the long term.
Constructing the progress	The individual components of the processes must be created and put in the right order.
Reducing dimensions	Since many algorithms on high-dimensional data do not deliver good results, a check should be made to determine whether data dimensions can be removed or amalgamated.
Ensuring validity	Even while the models are being constructed, the probability of over-adaptation can be reduced (by division into training and test partitions and by cross-validation, for example).
Considering multiple analytical methods	If applicable, multiple analytical methods can be tried or combined by forming ensembles.
Selecting the best parameter configuration	Various combinations must be systematically tested to select suitable or desired settings.
Weighing time against benefit	The quality of the result must suit the problem to be solved, but it is possible that the total research costs for the analysis will not exceed the benefit of the model.
Ensuring replicability and transparency	Replicability and transparency must be ensured by storing the transformed data and all configurations of the training process (such as the seeds used).

7.5 Accompanying task “Tool Selection”

Tool selection aims to identify a suitable implementation infrastructure for the selected procedure. This refers to both hardware and software. This area, therefore, also partially overlaps with the key area *IT infrastructure* (cf. Chapter 12). However, this typically does not belong to the core tasks of data scientists and must be prepared much more extensively.

The term *tool selection* is, therefore, to be understood as selecting individual components of the IT landscape that contribute to a direct solution within the context of the problem. Depending on the organization, the hardware and software might have been previously prescribed, so selecting them might no longer fall within the framework of the project, but using them is considered a requirement.

Table 15 (see next page) lists and describes subtasks for *tool selection* frequently mentioned by participants.

Unlike the requirements for the implementation infrastructure, an extensive documentation of the selection process is normally necessary only for large-scale projects.

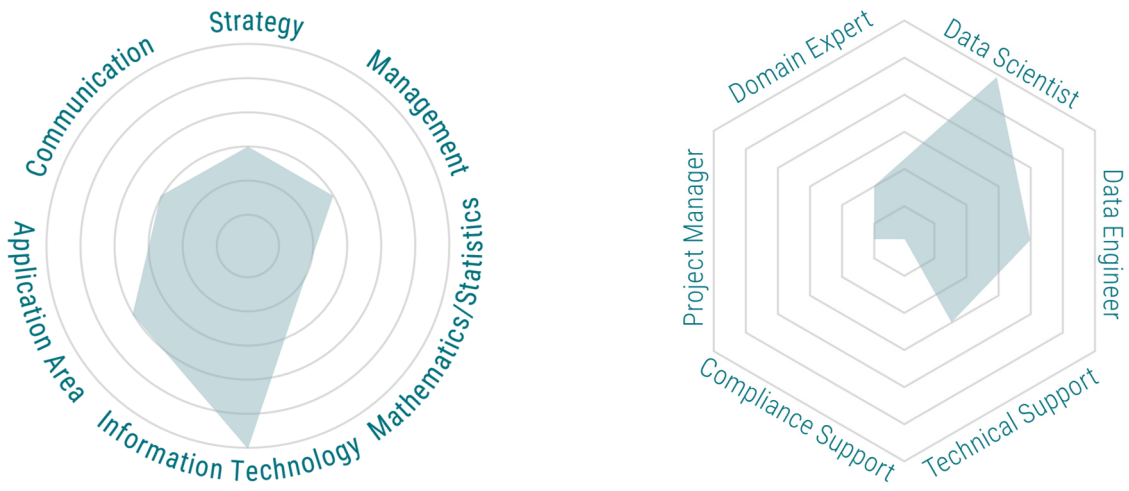


Figure 26: Competence and role profile for the “Tool Selection” task

Table 15: Frequently mentioned subtasks of the "Tool Selection" task

Subtask	Description
Researching suitable software	As soon as the considered analytical methods can be estimated, the software with which the procedure is to be implemented and how the software can be procured or created if they are not available already must be clarified.
Researching suitable hardware	Regardless of how much research is needed and whether the application will be carried out locally or in a cloud, various hardware will be required.
Comparison with the abilities available in the project team	If a tool cannot be operated adequately or at all, either another tool must be selected, a training measure must be introduced, or external resources must be brought in.
Evaluating the tool's suitability	If a tool is not fully compatible with the project's remaining workflow, a compromise between the complete implementation of the procedure sought and integration into the remaining infrastructure must be found.
Quality assurance during implementation	The quality of implementation must be ensured through software validation, peer review, or the like.

7.6 Core task “Developing Analytical Methods”

If there are no suitable analytical methods, existing procedures must be adjusted or merged (if possible) or completely new solutions can be developed. In so doing, it must be established whether the procedure should have an application that is as varied as possible or should be optimized for the special application or available data. The efficiency of the independent development must be observed, thus avoiding unnecessary work (caused by not using existing (auxiliary) procedures, for example). The newly developed procedure must be inserted into the implementation infrastructure, and temporal and budgetary restrictions must be considered.

Table 16 lists and describes the subtasks for developing new analytical methods frequently mentioned by participants.

Table 16: Frequently mentioned subtasks for the task of “Developing Analytical Methods”

Subtask	Description
Establishing criteria	What the procedure should and should not be able to do must be defined clearly and precisely.
Determining differences with relevant existing procedures	The inadequacies of relevant existing procedures for solving the problem (gap analysis) must be determined.
Establishing the procedure	It must be decided whether a completely new procedure should be developed or whether an existing idea can be built on.
Designing the procedure	A technical design of the new analytical method must be created.
Testing the procedure	An empirical model validation and reliability test must be performed as well as a comparison with existing procedures.
Implementation	The analytical method must be technically implemented.

The development of a new analytical method must be carefully and extensively documented. For example, this can include:

- *A reason for the new development*
- *The complete derivation of the procedure*
- *A description of the developed model (including all assumptions made and simplifications undertaken)*
- *The theoretical basis / underlying mathematics*
- *The comprehensive presentation of the developed algorithm*
- *The conditions for the application*
- *A description of the input and output*
- *Presentation of dependencies of existing software*
- *Documentation of the procedures on code level*
- *Various quality criteria (robustness, validity, objectivity, reliability)*
- *A user handbook*
- *Sample applications*
- *A "lessons learned" document*
- *Strengths and weaknesses of the procedure*
- *Potential for further development*

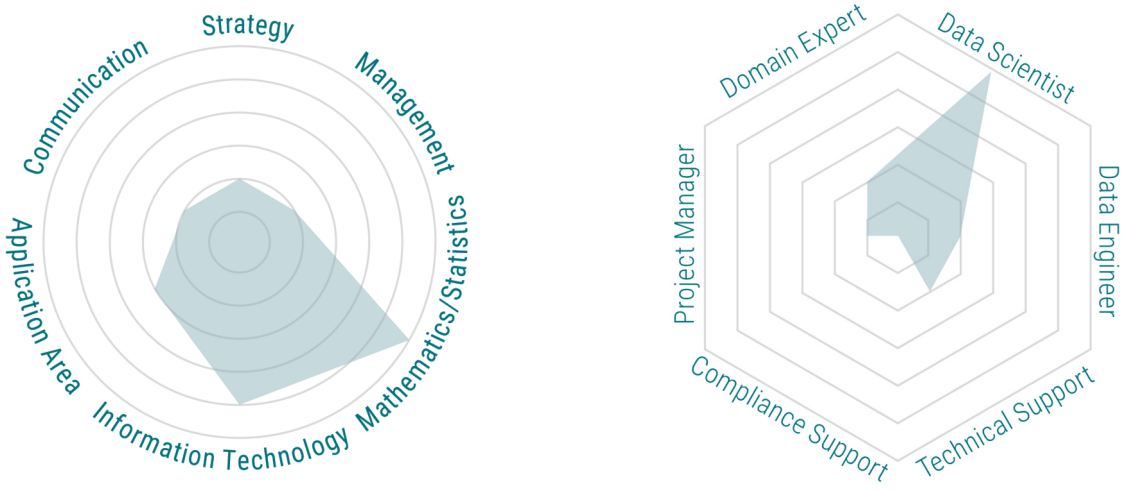


Figure 27: Competence and role profile for the task of "Developing Analytical Methods"

7.7 Accompanying task “Evaluation”

Evaluation is a diverse task in the key area “analytical methods” since it is performed at three points: (1) when potentially suitable analytical methods are selected for the task, (2) when new analytical methods are developed, and (3) when the selected or newly developed analytical method is applied to the specific problem. In all three cases, the goal is a reproducibly evaluation and classification of the results. Each evaluation is based on selecting a suitable metric. Therefore, both technical metrics and (in particular) the central criteria for the application domain must be considered, since only those perspectives allow the actual value of the performed analysis to be determined.

Table 17 lists and describes the subtasks of the evaluation frequently named by participants.

Table 17: Frequently mentioned subtasks of the “Evaluation” task

Subtask	Description
Determining the evaluation criteria	The criteria according to which the evaluation is performed must be selected depending on the domain and in regard to the project objective.
Estimating added value	The benefit to be gained from the performed analysis must be estimated in advance. This can only occur in the context of the domain-specific issue. Estimating added value establishes a framework for a justifiable effort for the analyses.
Reviewing realizabilty	The realizabilty of the analysis must be assessed regarding the attainability of the objective that has been set, the suitability of the available data, and the reasonableness of available resources.
Benchmarking	Suitable comparative criteria (a “benchmark”) for evaluating subsequent results must be selected. This can be an existing procedure that must be replaced, or a very simple comparative procedure that is usable with little effort.
Cost estimate	The cost required to implement the analytical method must be estimated. The estimated cost must be significantly less that the added value expected from the analysis.
Comparing procedures	The fundamental characteristics of the procedures must be worked out and compared with one another. Then the procedure’s suitability for the problem to be worked on must be assessed.
Evaluating results	The results of the performed analysis must be assessed. This typically entails a plausibility check, various statistical evaluations, validating the results, and an examination of the procedure’s robustness. Applicability must be checked from a domain perspective.
Performance tests	If the developed analytical method is later moved into regular operations, the procedure’s performance must be tested (hardware required, scope of the data quantities processed).

The results of the evaluation must be carefully documented. Selecting the procedure primarily entails comparing advantages and disadvantages of the analytical methods and describing suitable applications. In particular, evaluating the results includes presenting the evaluation criteria and their characteristics, the approach selected, the test setup, configuration tables, a list of the parameter combinations examined, and the specific test results (including details on the duration of the performance). The experiences and potential weaknesses collected during the evaluation of the procedure should be recorded. And finally, the decisions made based on the evaluation must be justified transparently and in the context of the problem examined.

Figure 28 displays the competence profile of people who specialize in the area of *evaluation* and tasks directly associate with it, as well as participating roles. Since three different aspects are examined during the evaluation, as described above, these might be conceivably performed by different people. In that case, the people involved do not need to possess the maximum characteristics shown below in every competence dimension.

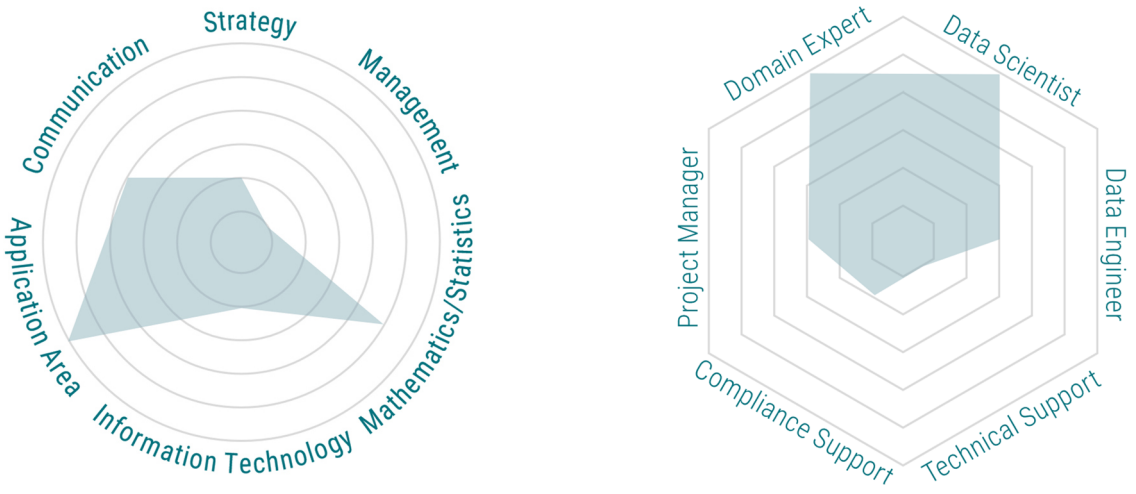


Figure 28: Competence and role profile for the "Evaluation" task

7.8 Feature-bearing area “Analysis Results”

The results of the analysis process can take very different forms depending on the issue at hand, the objective, the methods used, and the available data base. The spectrum ranges from descriptive and diagnostic analyses to predictive and prescriptive models to self-guiding systems. Table 18 lists and describes characteristics of analysis results frequently mentioned by participants.

Table 18: Frequently mentioned characteristics of the area “Analysis Results”

Characteristic	Description
Meaningfulness	What statements can be derived from the analysis results? Does this involve rough estimates or precise statements? Are the results expected to be valid in the future and not only in their current state?
Form of presentation	How are the analysis results conveyed? Are they described in a way that is easy to understand? Are they displayed to make them more vivid? Are the results shown in a detailed or aggregated manner?
Type of result	What type of analysis result is this (describing or clarifying a context, predicting future behavior, deriving a guideline, streamlining a system)?
Ease of generalization	How well can the results be transferred to additional data?
Limits	What restrictions are there on the developed model’s information value? What are the reasons for those restrictions (such as minimal data quantity, missing attributes, or restrictions of the analytical method?) How can this be overcome, if necessary?
Realizabilty	Can (and should) the analysis model be further developed into a software that permanently provides the analysis function for new data?
Complexity	How easy are the results to understand, and how well can measures be derived from them?
Novelty	Are findings gained that otherwise would not have come to light or been available?
Quantitative evaluation	What are the quantitative evaluation criteria (significance level, error rate, etc.)?
Relevance	Do the results contribute to solving the original problem, answer another question, or have additional uses? Are the results trivial, or do they deliver new findings? Can specific action requirements be derived from them?
Transparency	Is the development process for the analysis results transparent and traceable?
Comparability	Can the analysis results be compared with the results of other known procedures?
Comprehensibility	Can the results be understood in and of themselves? Are interpretation aids needed?
Completeness	How complete are the existing results? Were only partial aspects examined, or was there a comprehensive analysis? Is there a recognizable need for further analyses?

8 Deployment

In the “Deployment” phase, an applicable form of the analysis results is created. Depending on the project, this can entail comprehensively considering technical, methodological, and professional tasks, or it can be handled pragmatically. The analysis artifact can include results as well as models or procedures and is provided to its target recipients in various forms.

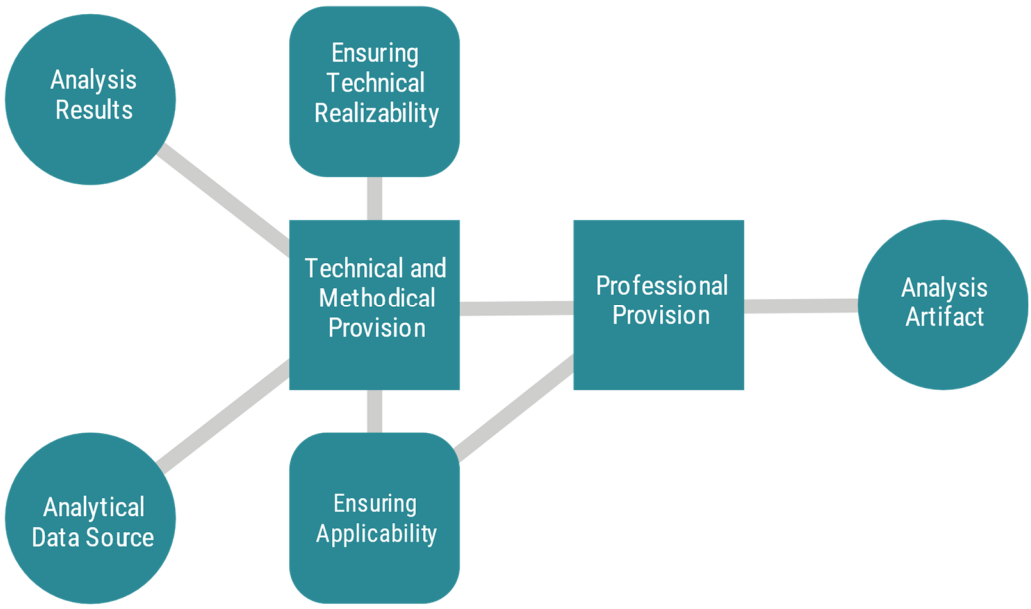


Figure 29: Brief overview of the “Deployment” phase

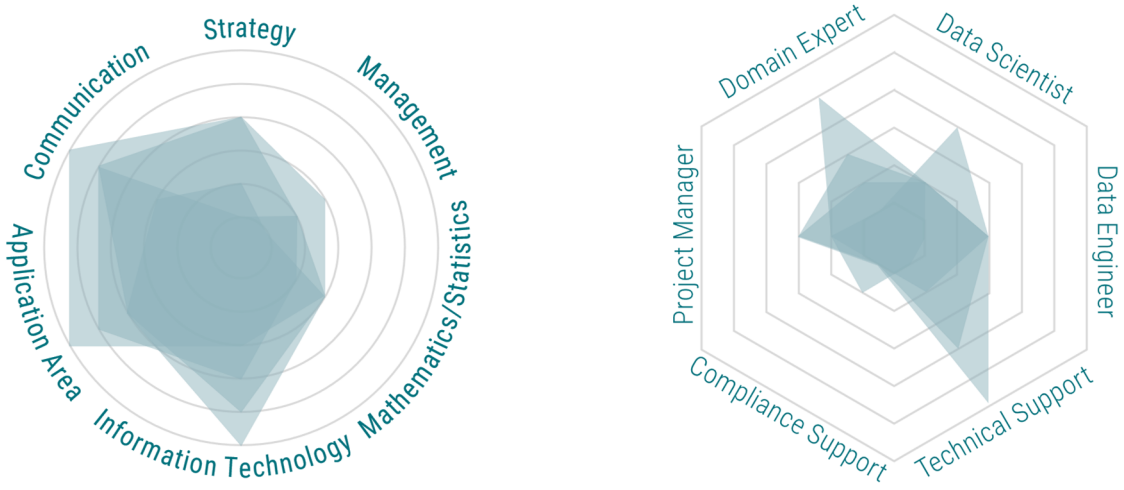
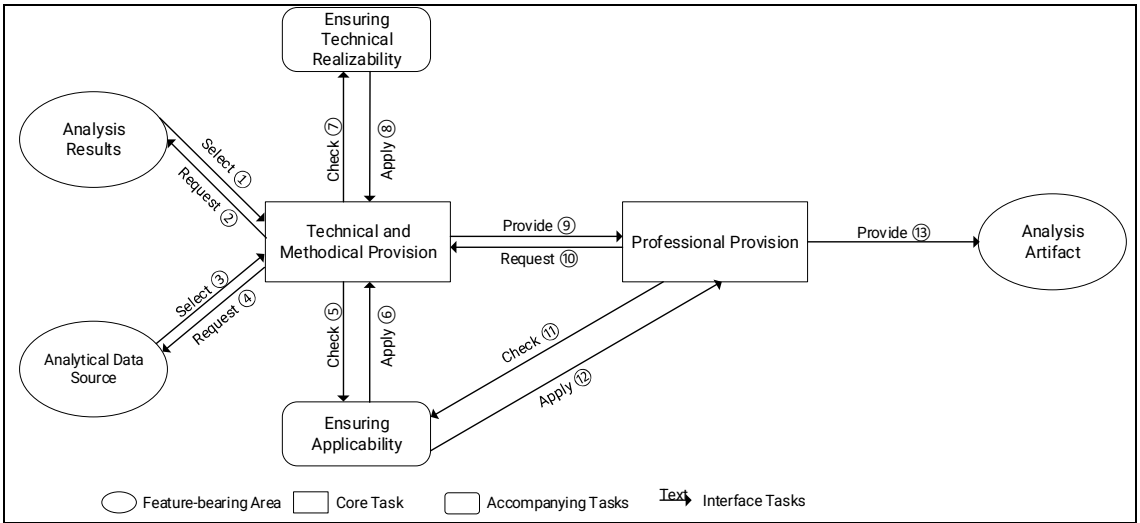


Figure 30: Competence and role profile for the “Deployment” phase

Detailed presentation of the “Deployment” phase



① Selecting the analysis results for the technical and methodical provision

② If the analysis results do not suit the intended technical and methodical provision, changes can be requested.

③ If access to data is required for technical and methodical provision, they must be selected. If new data are used, they must possess the same characteristics as the data with which the model was developed.

④ If the selected data do not suit the intended technical and methodical provision, changes can be requested.

⑤ During the technical and methodical provision, the applicability must be checked by the target group for the analysis.

⑥ Identified opportunities for ensuring applicability must be considered during the technical and methodical provision.

⑦ During the technical and methodical provision, technical requirements must be examined.

⑧ Any technical impossibilities for implementation identified must be considered during the technical and methodical provision.

⑨ The results of the technical and methodical provision form the basis for the professional provision of the prepared analysis results for the target user group. The effects can range from a support of existing processes to a project adjustment to a completely new development (now automated, if applicable) of processes.

⑩ If the professional provision does not lead to the desired results, changes to the technical and methodical provision can be requested.

⑪ During the professional provision, the applicability of the analysis must be checked by the target group.

⑫ Identified opportunities for ensuring applicability should be taken into account during the professional provision.

⑬ Analysis artifacts that arise from the professional provision should be carried over into deployment in a suitable form.

Figure 31: Detailed presentation of the “Deployment” phase

8.1 Feature-bearing area “Analysis Results”

Deployment is based on the characteristics of the analysis results described in Section 0. No additional special features should be recorded at this point.

8.2 Feature-bearing area “Analytical Data Source”

To apply the analysis results in practice, it may be necessary to access the analytical data source again (cf. Section 6.5). No additional special features should be recorded during this phase.

8.3 Core task “Technical and Methodical Provision”

The results of the analysis must be prepared for implementation in a suitable form. Thus, a distinction must be made for the following:

- *A manual use of the results in which they are prepared for the target group and conveyed in seminars or workshops, for example*
- *An implementation of the results, possibly in the form of a report in which the results are prepared once*
- *The application of the trained model so it can be applied to unknown data*
- *Continual learning in which the model can be independently adjusted to unknown data through repeated use*
- *Publishing the developed analytical method (possibly only within the organization) so third parties can use it. This allows model results to be checked independently and weaknesses identified early on*

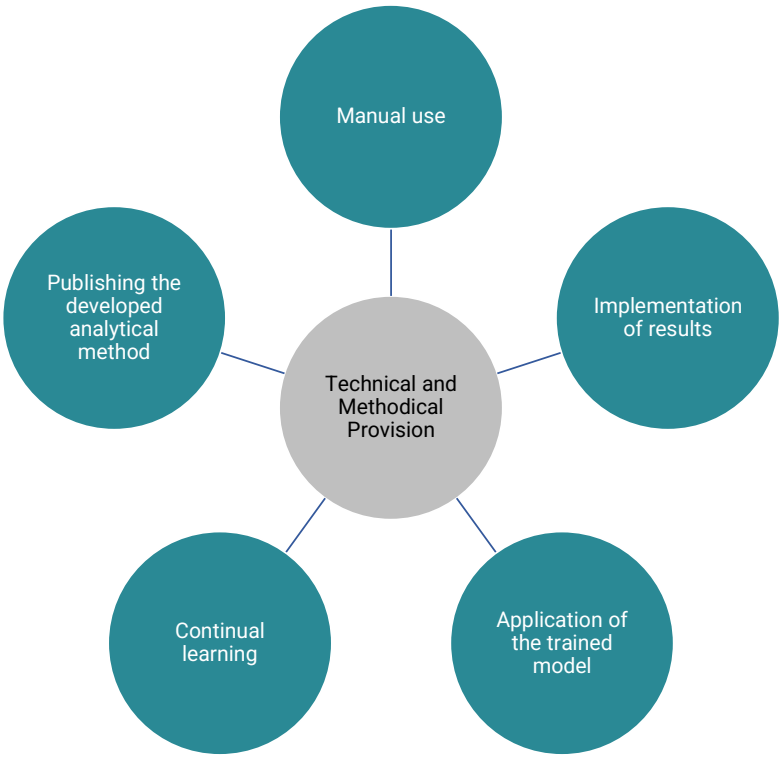


Figure 32: Forms of “Technical and Methodical Provision”

Depending on the project, selecting multiple implementation options is also conceivable.

The model must be embedded in an operative production environment. One-time results are relevant in exceptional cases (for a proof of concept, for example), but otherwise the model’s value is normally in being embedded in a product environment either continually or on request.

Table 19 lists and describes the subtasks of *technical and methodical provision* frequently named by participants.

Table 19: Frequently mentioned subtasks of the task "Technical and Methodical Provision"

Subtask	Description
Preparing the results for the recipients	Suitable technical and methodical preparation and possibility of interpretation by the user
Building the product environment	It might be necessary to build a new infrastructure in which the results can be continually updated and considered.
Transferring the results	For ongoing operations, it might be necessary to transfer the results from the analysis environment into an operative system.
Context creation	It must be easy to see how and when the results were gained.
Automating processes	Consider general challenges when processes are automated, such as: • *What happens in case of errors?* • *How should one deal with media disruptions? Can they be avoided or compensated for?* • *How can the execution be suitably logged?*
Dealing with IT resources	An efficient use of IT resources must be ensured.
Technically testing the system used	The functioning of the analysis system must be checked to make sure it is free from technical errors—especially if it is integrated into the organization's product environment and connected to real data sources.

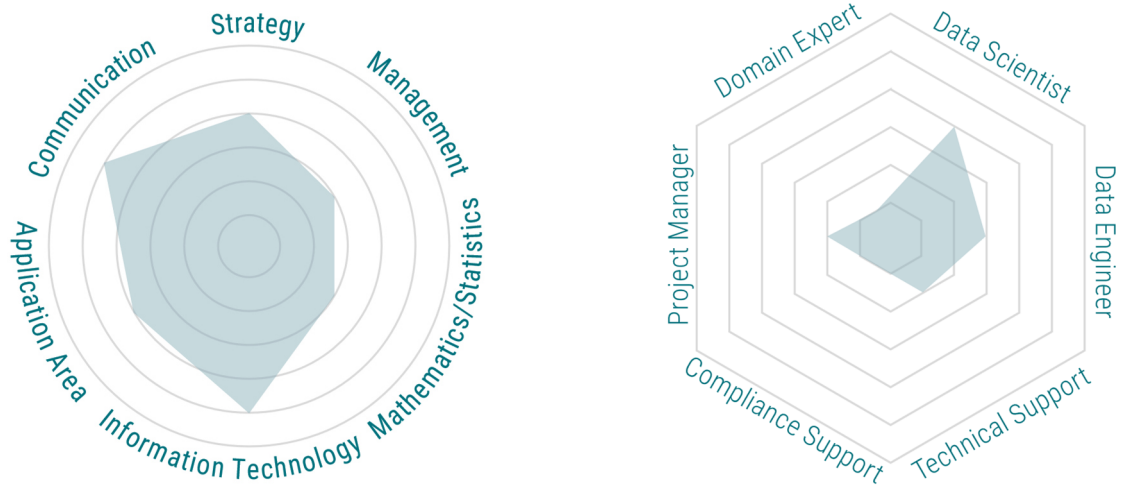


Figure 33: Competence and role profile for the "Technical and Methodical Provision" task

8.4 Accompanying task “Ensuring Technical Realizability”

Normally, technical and methodical provision means a complete automation of the procedure. In a few cases, however, it can be useful or necessary to incorporate manual steps. The accompanying task “*ensuring technical realizability*” should guarantee the initial setup and permanent operation of the analysis application under the defined economic framework conditions. This also includes ensuring the (technical) operability of the application, the performance of maintenance work, and the implementation of technical adjustments.

Table 20 (see next page) lists and describes subtasks of ensuring technical realizability frequently mentioned by participants.

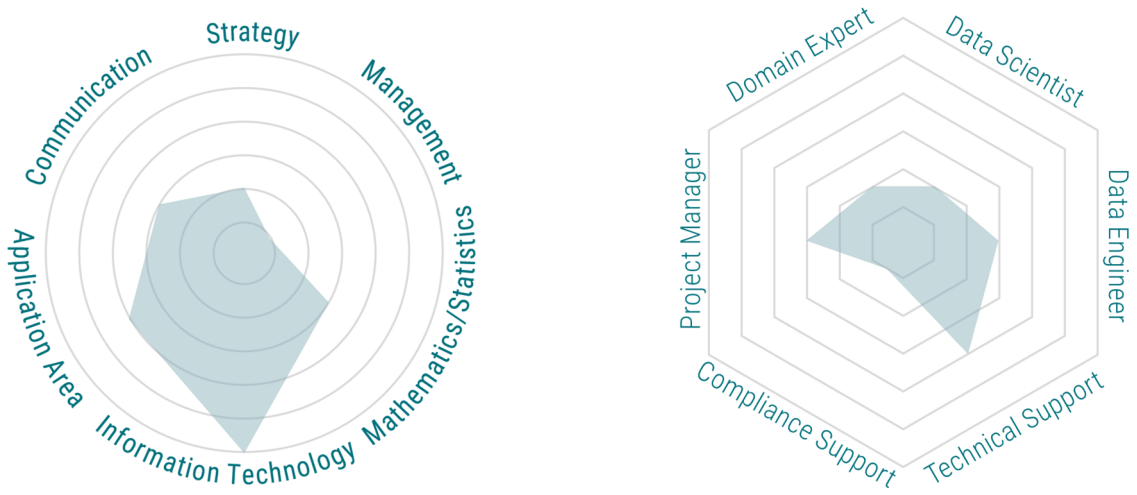


Figure 34: Competence and role profile for the “Ensuring Technical Realizability” task

Table 20: Frequently mentioned subtasks of the task "Ensuring Technical Realizability"

Subtask	Description
Considering time criticalities	Must the analysis be performed in real time, or is it not time-critical and performable overnight in batch operations, for example?
Considering durations	How CPU-intensive is the algorithm? Is it easily scalable with the data quantity, for example?
Dealing with the connected data sources	How can one react to changes in the data sources (formats, quality, right, etc.)? Who is responsible? What is the information flow?
Identifying the hardware stacks	What hardware is needed to operate the analysis solution? What realization form (on premise, Private Cloud, Cloud, IaaS, PaaS, SaaS, etc.) is suitable?
Identifying the software stacks	Is the software stack to be used by the organization already prescribed, or must it still be evaluated as part of the project? The competencies of the groups of participants should also be reflected on.
Identifying technical conditions and opportunities	A consideration of the given IT infrastructure, or the possibility of procuring one, must be checked.
Testing software licenses	Are further or additional licenses needed for the productive system?
Legal framework conditions	Were the legal framework conditions for using the analysis application (data protection, compliance, etc.) clarified, defined, and documented?
Create memory access concept	Is it possible to restrict access to analysis results to authorized user groups? Have precautions been taken to guarantee the security of all data?
Ensure operations and support	Who is responsible for the productive operations of the analysis application? Who can offer support if technical or methodical questions or problems arise?
Automation	To what extent can the evaluation of data and integration of results be automated? How often are the analyses repeated?

8.5 Accompanying task “Ensuring Applicability”

The analysis results must be available in a form that can be used by or conveyed to the target group. Applicability should be ensured through interactions between people with methodical expertise and people from the domain.

Table 21 describes subtasks for ensuring applicability frequently named by participants.

Table 21: Frequently mentioned subtasks of the task “Ensure Applicability”

Subtask	Description
Identify target recipients	To ensure applicability, the target recipients of the analysis must be known.
Establish UI/UX design	The surface should be simple for all user groups to understand and use while offering flexibility and covering the complexity of the topic. Analysis results should be prepared in a way that is easy to understand, especially by using visual elements.
Ensure memory access	Authorization structures and accesses must be defined. Guaranteeing realizability is part of the accompanying task “*ensuring technical realizability*.”
Involve users	Before the analysis results are applied, workshops can be held, for example, to obtain feedback on ensuring applicability.
Create a documentation concept	Besides technical and methodical documentation, suitable user documentation must be planned, especially as an interpretive aid or to describe the key figures used.
Create a training concept	Depending on the scope of the analysis artifact developed and the form of the deployment, a suitable training concept must be developed.

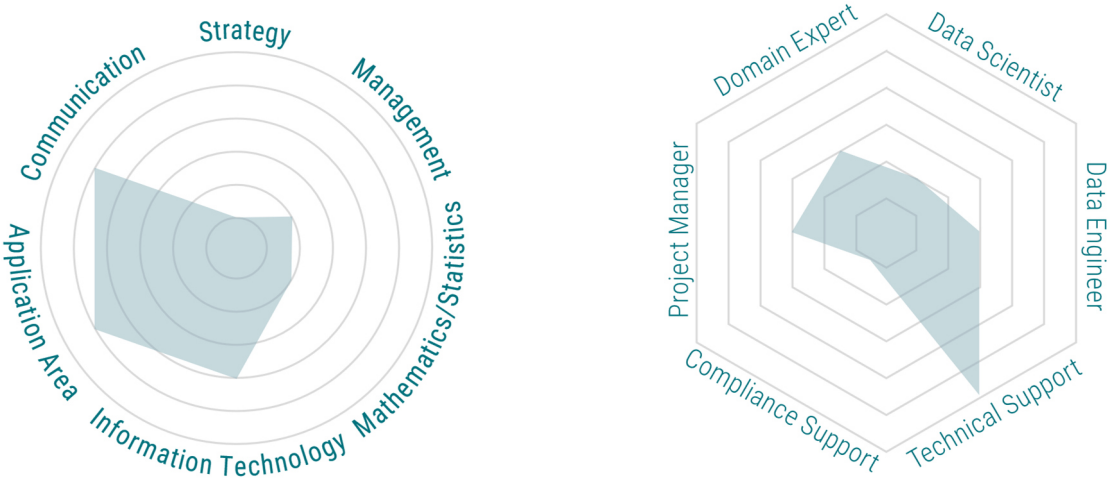


Figure 35: Competence and role profile for the task “Ensuring Applicability”

8.6 Core task “Professional Provision”

The subtasks of “professional provision” strongly depend on the form of provision and the domain in which the project will be implemented. Therefore, only general tasks will be shown. Table 22 describes subtasks for professional provision frequently named by participants.

Table 22: Frequently mentioned subtasks of the task “Professional Provision”

Subtask	Description
Ensure sustainability	Sustainability means ensuring permanent use or relevance.
Consider reach and impact	Before the results are published outside the project teams, their potential impact should be assessed from moral and economic points of view (among others).
Consider legal issues	Before the analysis results are used, data protection and legal issues must be assessed.
Establish points of contact	Points of contact must be established for questions that arise during ongoing use. To that end, a defined possibility for establishing contact must also be established.
Integration into existing processes	The analysis artifacts must be functionally integrated into existing processes.
Internal cost calculation model	Personnel and IT costs for operating the analysis artifacts must be determined and allocated to the user if applicable.
Hold trainings	The trainings as part of ensuring applicability must be carried out in a suitable form (in-person trainings, online trainings, webinars, etc.).
Create a user handbook	The user documentation conceived as part of ensuring applicability must be created.
Determine troubleshooting	Testing mechanisms and behavior patterns must be determined for the eventuality that the analysis artifact delivers or stops delivering useful results.

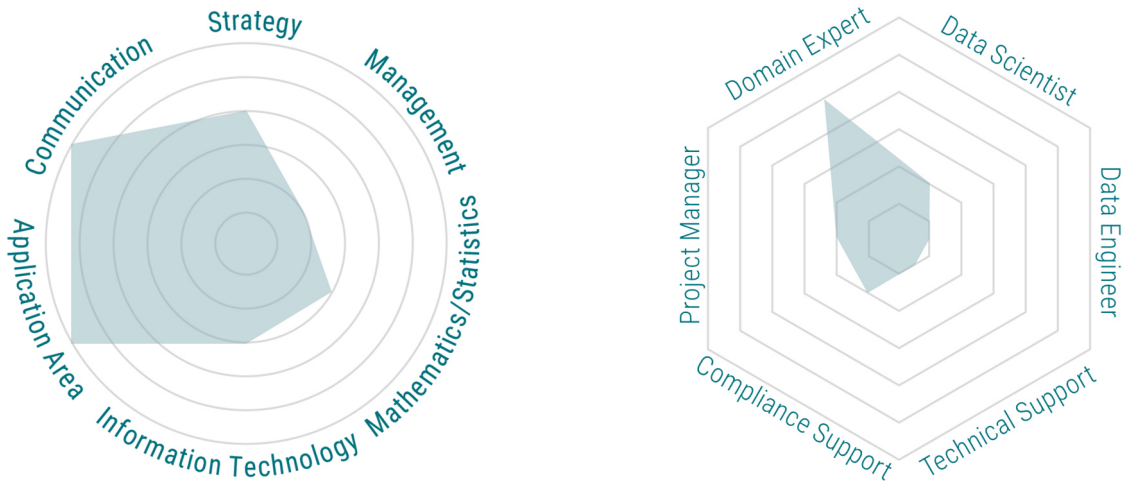


Figure 36: Competence and role profile for the “Professional Provision” task

8.7 Feature-bearing area “Analysis Artifacts”

The characteristics of the analysis artifacts depend on the form in which the results are presented (cf. Chapter 8.1). Table 23 describes characteristics of analysis artifacts frequently mentioned by participants.

Table 23: Frequently mentioned characteristics of the area “Analysis Artifacts”

Characteristic	Description
User documentation	The users of the analysis system must be given a user guide or user handbook that describes the available reports, dashboards, database, etc., along with their memory access rights. Furthermore, specialized points of contact must be named.
Technical documentation	To maintain and enhance the analysis system, a detailed description of the software used (code base, input and output, interim steps performed, and dependencies on other components) must be available. Moreover, the technical infrastructure that was created for the analysis system or in which it is embedded must be documented. Technical points of contact must be named here as well.
Model documentation	The analysis model must be described in detail (including the premises for using it) so it can be adjusted and enhanced in the future.
Recommended actions	Recommended actions must be defined for the recipients of the analysis artifacts, at least if results are used manually.
Models	The models emerging from the analysis can be applied to new data.
Reports	The data emerging from the analysis must be presented to the appropriate target groups in the form of reports.
Analysis infrastructure	To ensure long-term use of the analysis model, a specific analysis infrastructure must frequently be provided that itself is embedded in the organization’s infrastructure.
Support	Defined professional and technical support is needed both to support operations and to rectify problem cases.

9 Application

Using artifacts after the project performance is not thought of as a primary part of a data science project. Monitoring is necessary, however (depending on the form of deployment), to check the model's continuing suitability in the application and obtain findings for ongoing and new developments (including developments for the purposes of iterative approaches).

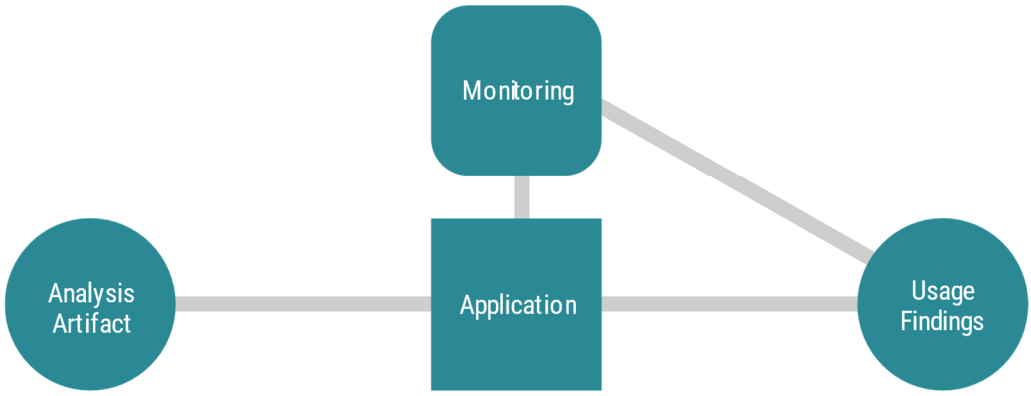


Figure 37: Brief overview of the "Application" phase

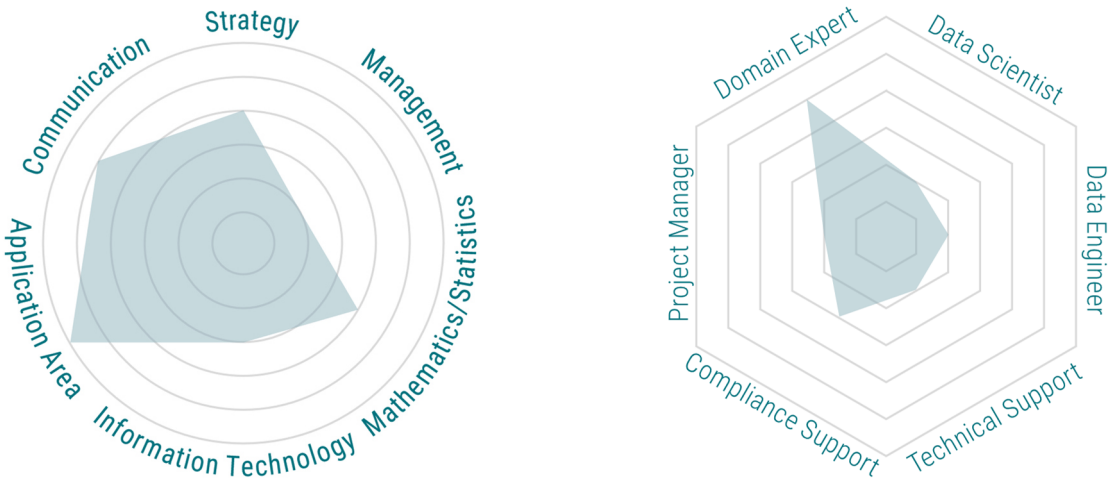
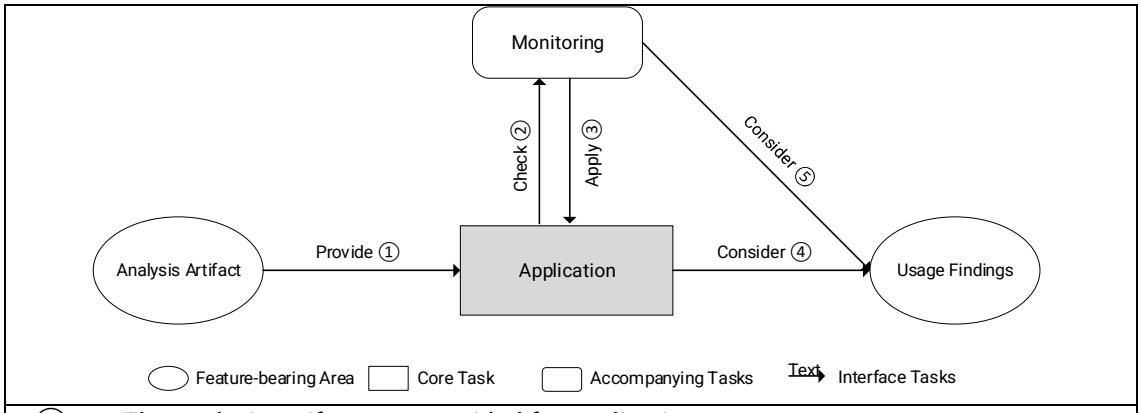


Figure 38: Competence and role profile for the "Application" phase

Detailed presentation of the “Application” phase



① The analysis artifacts are provided for application.
② The application must be repeatedly monitored.
③ Results of the monitoring should be contemplated during application.
④ Domain-specific findings arising from application must be considered during new and further developments.
⑤ Data science-specific findings arising from applications should be taken into account during new and further developments.

Figure 39: Detailed presentation of the “Application” phase

9.1 Feature-bearing area “Analysis Artifacts”

Application of the analysis artifacts is based on their characteristics (cf. Section 8.7). No additional special features should be recorded during this phase.

9.2 Accompanying task “Monitoring”

Monitoring must include the regular operations for which the analysis artifact is designed over the long term. Therefore, the quality of the analysis results must be continually examined and the constant applicability of the model must be verified. Table 24 lists and describes the subtasks of monitoring the application frequently named by participants.

Table 24: Frequently mentioned subtasks of the task “Monitoring”

Subtask	Description
Analysis artifacts in general	
Ensuring the correct application domain	The analysis artifacts are created for a certain domain. That specialization must be preserved.
Evaluating the analysis artifacts	The results of the analysis should be repeatedly evaluated regarding their informative power and predictive qualities.
Check the sustainability of the analysis artifacts	A check must be made to determine whether the analysis artifacts are maintained and how quickly results become outdated.
Applying analysis artifacts	
Checking the data	The model might be applied to data that do not yet exist at the time of creation. It should be ensured as much as possible that the application delivers correct results. This should be verified by both data experts and domain experts.
Monitoring errors	Error reports must be collected and evaluated; they include, among other things, the unexpected behavior or models or new forms of data errors.
Metadata of the application	
Recognizing performance challenges	Identifying performance challenges is limited during deployment. Therefore, this aspect should also be monitored during the application.
Evaluating usage data	A check must be made to determine whether the analysis artifacts should still be used. The usage data must be recorded for this purpose.

One artifact of this task area is an evaluation report that allows analysis artifacts to be evaluated in terms of their usefulness. Since the key area “application” contains only one subtask that is a data scientist’s responsibility (monitoring), no specific graphics for competencies and roles is available, but only the general graphics or the phase overview.

9.3 Feature-bearing area “Usage Findings”

Based on the usage findings, it can be determined whether analysis artifacts should be reworked or discontinued. This can become necessary because circumstances have changed or the developed solution has not proven itself productive.

Table 25 describes characteristics frequently mentioned by participants according to which usage findings can be subdivided.

Table 25: Frequently mentioned characteristics of the area “Usage Findings”

Characteristic	Description
Error reports	The reports enable an evaluation of whether the analysis artifacts can be operated adequately in a stable manner.
Usage frequency	If analysis artifacts are not used by the domain experts, operations can be unnecessary and enhancements might be dispensed with.
Performance of the analysis artifacts	Observing the performance enables an evaluation of the suitability of the technical infrastructure used.
Type of application	The type of application can influence a possible enhancement from a domain perspective.

Part C
Overarching key areas

10 Domain

Besides the tasks addressed explicitly in chapter 5, the domain influences all other key areas, although to different extents. What follows is a presentation of the subareas, shown according to the key area that participants most frequently named as relevant regarding the domain.

Key area “Data”

- *Raw data sources: The data relevance need only be evaluated specifically for the domain; likewise, an understanding of the data can be developed only while considering the domain.*
- *Data preparation: Data standards within the domain (data protection, for example) must be considered. Transformations (such as making measurement values comprehensible, or identifying data errors or outliers that lie far from the center of data distribution but are correct values and not measurement errors) should also be viewed in a domain context.*
- *Explorative data analysis: The domain specifics are incorporated into the explorative data analysis through the problem to be solved. The goal is to create added value regarding the domain.*

Key area “Analytical Methods”

- *Requirements for analytical methods: Domain-specific framework conditions (such as legal or regulatory ones) might exclude entire categories of procedures for the application. Moreover, desirable but not absolutely necessary requirements often exist that are significant for the domain experts in the application. For example, there are many areas in which explainable models are desirable or causal dependencies should be considered.*
- *Identifying suitable analytical methods: Many domains entail frequently used analytical methods that can be used as a benchmark. The form of the desired analysis results influences the selection.*
- *Evaluation: Results must be shown as classified in the domain context and in a form suitable for domain experts. Depending on the application, more extensive requirement results for the evaluation or the relevant evaluation metrics are defined from the domain.*

Key area “Deployment”

- *Ensuring applicability: The user’s domain background needs to be taken into account.*
- Professional *provision: Analysis results are applied in a domain context.*
- *Technical and methodical provision: The framework conditions for the deployment need to be considered.*
- *Ensuring technical realizability: Existing nonfunctional requirements of the domains must be known.*

11 Scientificity

When data science projects are carried out, a scientific approach is needed for every single phase: This can be seen in the use of a process model. The degree of scientificity can vary, but the minimum requirements of complete replicability and statistical validity must be guaranteed. The degree of variation particularly affects the theoretical anchoring of the research question, which can be cut short during projects that are strongly oriented toward practice. Therefore, a cost-benefit analysis must be performed considering the risks, such as the possibility that the solution space might be limited because the literature has not been reappraised thoroughly or at all, causing important earlier findings or methods to be neglected. Regardless of characteristics of the individual project, a structured approach, suitable documentation, and an evaluation or validation of the results are indispensable.

The scientific requirements that affect the entire data science project should be briefly illuminated. What follows is a presentation of the subareas, shown according to the key areas that participants most frequently named as relevant regarding the scientificity.

Scientific requirements that affect all key areas

Basically, the same fundamental standards apply to a data science project that must satisfy other scientific work that is oriented toward practice. There are primarily four points:

1. *The research object (the project order, in this case) must be outlined precisely enough so that third parties can recognize it. This is important for demarcating and classifying the statement of the scientific contribution and selecting the suitable methods.*
2. *The project must result in a statement that could not have been made before (from this viewpoint). Otherwise, the project would be obsolete.*
3. *The result must be usable, although this is frequently already prescribed by the project order.*
4. *The project must be documented to make it possible for a "scientific public" to use existing information to verify the statements or hypotheses made. It is possible, however, for the "scientific public" to exist only within a company—especially in a corporate context. This last item goes hand in hand with the principle of replicability, which ensures that the method is described well enough that another party with access to the same infrastructure and data will arrive at the same result. This also has implications for the statistical durability of the results, which must be rigorously evaluated.*

Furthermore, the three requirements of objectivity, reliability and validity must be observed.

Key area "Domain"

- *Definition of the project: Whether the project exists in an economic or scientific field, it typically has a scientific context that can necessitate a reappraisal of existing procedures and scientific publications, depending on the project.*

Key area "Data"

- *Explorative data analysis: Data must be understood as extensively as possible, and a well-founded, statistical evaluation or validation of the results and the formation of potential errors in the data must be performed. The data's suitability for investigating the problem to be solved must be checked, and data cleansing requirements must be identified. Evidence must be provided to show that a suitable explorative data analysis was applied correctly.*
- *Data preparation: The data transformation must be transparent and replicable, correct procedures must be used, and preparation steps must be documented in a suitable manner. The raw data must be archived for the long term so the data preparation can be replicated.*

Key area "Analytical Methods"

- *Identifying suitable analytical methods: The requirements for the analytical methods to be developed must be checked, and goals and framework parameters must be transparently established. An overview of available procedures that meet these criteria must be created (and must include the consideration of current scientific publications), and the selection of the analytical methods must be justified. Even the realization that no suitable procedure exists must be suitably presented.*
- *Applying analytical methods: When analytical methods are parameterized, the process must be goal-oriented. The specific application of the method must be guaranteed as well as the objectivity. Even with correct application, suitable scientific literature must be consulted. The basic assumptions for the analytical method must be ensured. Documentation of analysis results must be prepared that includes their interpretation. The evaluation must be kept in mind from the beginning: Testing and validation approaches must be kept in suitable form. Only this can prevent the analysis results from reflecting statistical artifacts of the data considered rather than generally valid contexts.*
- *Developing analytical methods: When existing procedures are integrated, the locations at which previous methods are installed and previous methods feature weaknesses, and how these will be remedied by changing the procedures are needed. In addition, it must be proven that no procedure currently exists. This should show that a new development is needed. In so doing, the results from the identification must be consulted. Interaction with the specialist community is useful when an analytical method is being developed to develop suitable procedures according to defined standards and have them examined if applicable.*
- *Evaluation: Evaluations and tests must be systematically prepared, and a correct application of suitable evaluation procedures must be ensured by using a constant test environment. The correct functioning of the procedures must be verified, the results must be evaluated critically, and the evaluation must be completely documented.*

Key area "Deployment"

- *Technical and methodical provision: After the analytical method has been concluded, it will be made available both technically and methodically. This may take the form of self-contained software modules or packets or a usable service within an IT infrastructure. The latter contains the provision of a programming interface or (web) service, for example. A complete description and documentation should be provided so users can employ the analytical method.*

Key area “Application”

- *Monitoring and evaluating the usage findings: Handing the analysis artifacts over to the users is connected with hypotheses on the contribution of the artifacts. The extent to which the analysis artifacts deserve these hypotheses must be recorded and evaluated with scientific accuracy. One essential requirement is to document differences between the original environment of training data and the actual usage environment of the analysis artifacts. Changes in the usage environment must also be recorded.*

12 IT Infrastructure

The IT infrastructure should be well-planned in all key areas, although to different extents. The type and size of the project are important factors in terms of the role the infrastructure actually plays. Thus, for example, the form of the planned deployment or the complexity of the data and analytical methods must be considered. The organization's existing infrastructure should typically be studied during an observation. This applies mainly to systems that are directly connected with the data science project, but also for infrastructure that can be used for the project management or team cooperation.

What follows is a presentation of the subareas, shown according to the key areas that participants most frequently named as relevant regarding the IT infrastructure.

Key area "Data"

- *Analytical data source: The question of which interfaces and access possibilities there are for the analytical data sources should be well thought out. The target system might restrict the applicable technologies.*
- *Data preparation: The available computing power should also be determined, while keeping in mind the software that will be used for the data preparation.*
- *Explorative data analysis: The computing power of the IT infrastructure must be considered, as well as whether data can be analyzed directly in the database or whether they must be extracted first.*
- *Raw data sources: The question of which interfaces and memory access possibilities there are for the data sources must be considered. The source systems might restrict the applicable technologies.*

Key area "Analytical Methods"

- *Evaluation: Possible analytical methods should be evaluated taking into account the available or procurable technologies.*
- *Identifying suitable analytical methods: The question of which IT infrastructure is needed must be evaluated to perform the necessary analyses. An additional check should be made to determine whether the data can be examined in the analytical data source or whether they must be downloaded first.*

Key area "Deployment"

- *Ensuring technical realizability: Within this accompanying task, the requirements for the IT infrastructure must be reasonably detailed.*
- *Technical and methodical provision: The IT infrastructure must be suitable for operating the model in the intended form. To that end, the possibility of updates, backup, and memory access must be considered.*

Key area "Application"

- *Monitoring: Checks must be performed repeatedly to determine whether the IT infrastructure and its dimensioning are suitable and efficient for operations.*

Key area “Domain”

- *Ensuring realizability: A check must be made to determine whether the project can be implemented with the IT infrastructure that is available or can be procured within the project budget.*

Part D
Closing Remarks and Appendix

Closing remarks

Structuring a vast topic so it can be grasped in its entirety, and then purposefully using individual parts of it, is a widespread approach in both science and practice alike. It is, therefore, no wonder that people who deal with structures, patterns, and analytical preparation as part of their job feel a particular urge to delve into a complex topic like data science and make it accessible to a broader readership. The results of this project mark the end of such an endeavor and make it clear on various levels and in many facets how practitioners and researchers perceive the topic of data science and anchor themselves in their everyday lives. Readers of this work are, therefore, gaining a catalog of knowledge that has been reappraised in a structured manner while remaining directly applicable to their own professional context.

To attain this, a more comprehensive image of data science and the associated topic areas and related terminology was provided than was initially shown. The surveys within the working groups corroborated what the available literature suggests: Data science is a multilayered and strongly interdisciplinary area of work and research. The definition developed presents an extensive, yet precise description of the essential characteristics.

DASC-PM was developed to be a work that is relevant to practice: a process model that presents the relevant steps in the project-driven application of data science and describes the implementation of data science activities in detail. Experienced users in the data analysis field will find a model whose structure is similar to models that have been tried and tested for years—such as CRISP-DM—so that established activities can be successfully transferred to DASC-PM with reasonable effort. Newcomers to the topic will gain a model that reduces and gradually reformulates the complexity of data science initiatives into their core topics, so that when data science projects are first performed the main points can be explored in depth wherever this seems necessary. In both cases, as a result of an intensive exchange between scientists and practitioners, DASC-PM holds up a scientific approach as a key characteristic and supports the model's users by proceeding methodically and in a way that is easy to understand, so the results add value while remaining reliable.

For every core competence defined, the process model lists the activities and results that are relevant as part of data science initiatives and explains how they can be configured. Thus, the most important tasks are described and defined for each development. The extensive enumerations of characteristics for the individual tasks allow all users of the model to critically consider their own approach or seek out the characteristics that are the most relevant for their company based on the possibilities shown. A conscious selection can be performed to that end, obviating the tiring process of gathering information from different sources and offering a more comprehensive reference, so that the approach used seems not to have been chosen randomly but based on the experiences and exchange of a large group of experts. Supplementally, DASC-PM also provides various references to follow-up works or lists of criteria—on the topic of data quality, for example.

Besides the methodical considerations, DASC-PM also uses "data scientists" to place the most important components of successful data science in the foreground. The triumph of data science (as a term) began not with "data science" as such, but with the analysts who used it. The "data scientist" is the most prominent figure in the topic area discussed (Davenport and Patil, 2012). The activities and specific delineation of the topic areas have already been the topic of a dynamic

discussion for a number of years, one which has brought about publications to match (Harris et al., 2013; Zschech et al., 2018).

The observation of core competencies made here shows that there is no such thing as “the” data scientist, and there does not need to be. Although primarily knowledge of mathematics, statistics, and information technology must be available for the initial provision, preparation, and explorative analysis of data, the necessary competence profile shifts slightly when analytical methods are considered and primarily specify a greater understanding of the application area as a requirement. The fact that the core profile of the key area “analytical methods” also demands fundamental communicative abilities and an understanding of information technology, making it more comprehensive than that of most other fields, is explained by the fact that data science, as presented here, essentially describes data analysis. Accordingly, the most complex requirement profile is found in that area.

The more strongly the key areas affect the experts’ perspective, the more mathematics, statistics, and information technology move into the background as competencies. Instead, what is decisive for the deployment and general appearance in the domain are communicative abilities, strategic understanding, and, crucially, a higher comprehension of the application area. Personnel managers can also read from the extensive presentations on the individual key areas that project management as a competence actually affects only a few people. As in many other considerations concerning this topic, it is evident that a data science project requires an orchestrating element, but not every data scientist must be a project manager or even possess extensive knowledge in all areas.

And this is good news for companies. The high demand for data scientists in the labor market combined with the huge number of relevant competencies makes it almost impossible to find a jack-of-all-trades or the “all-round talent.” If DASC-PM is used to structure the analysis process, however, individual positions can be purposefully filled with people who have great expertise in their area and can implement that know-how at the appropriate points. Many such experts are available in companies or can be sought after or trained.

This process model, like all model, is a simplified version of reality. It need not be followed to the letter and makes no claim to present every variant and eventuality of a process or method. It provides no instructions on how to completely process every individual building block shown. Instead, the model is a solid foundation for performing data science initiatives because it represents more than the experiences of only one company or one research group. Therefore, DASC-PM is more than a best-practice approach. It is a structured, well-founded, and implementable reappraisal of one of the most relevant topics in business and science: the systematic, results-oriented deployment of data—data science.

Version 1.1 of DASC-PM also contains a supplement with practical value: an extensive appendix that particularly supports the project order phase and considers it holistically.

References

Alekozai E. M., Kaufmann J., Kühnel S., Neuhaus U., Schulz M. (2021). Data-Science-Projekte mit dem Vorgehensmodell "DASC-PM" durchführen: Kompetenzen, Rollen und Abläufe. In: Barton T., Müller C. (eds.) Data Science anwenden. Angewandte Wirtschaftsinformatik. Springer Vieweg, Wiesbaden.

Chatfield, A. T., Shlemoon, V. N., Redublado, W., Rahman, F. (2014). Data scientists as game changers in big data environments. In: *Australasian Conference on Information Systems (ACIS), Faculty of Engineering and Information Sciences - Papers: Part A*, 5646.

Conway, D. (2010). The data science Venn diagram. Dataists, drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, retrieved on February 3, 2020.

Davenport, T. H., Patil, D. J. (2012). Data scientist. *Harvard Business Review*, 90 (5), 70-76.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56 (12), 64-73.

Dorard, L. (2015). Machine Learning Canvas, https://www.ownml.co/machine-learning-canvas, retrieved on December 20, 2021.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17 (3), 37.

Harris, H., Murphy, S., Vaisman, M. (2013). Analyzing the Analyzers: An Introspective Survey of Data Scientists and Their Work. O'Reilly.

Jayawardene, V., Sadiq, S., Indulska, M. (2013). The Curse of Dimensionality in Data Quality. *Australasian Conference on Information Systems (ACIS) 2013 Proceedings*, paper 165.

Kerzel, U. (2021). Enterprise AI canvas integrating artificial intelligence into business. In: *Applied Artificial Intelligence*, 35 (1), 1-12.

McAfee, A., Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, 90 (10), 60-66.

Microsoft (2021): What is the team data science process?, https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview, retrieved on February 6, 2022.

Olivotti, D., Passlick, J., Axjonow, A., Eilers, D., Breitner, M. H. (2018). Combining machine learning and domain experience: A hybrid-learning monitor approach for industrial machines. In: *International Conference on Exploring Service Science*, 261-273. Springer, Cham.

Palmer, M. (2006). Data is the new oil, https://ana.blogs.com/maestros/2006/11/data_is_the_new.html, retrieved on December 9, 2019.

Provost, F., Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data*, 1 (1), 51-59.

Schmarzo, B. (2015). Big Data MBA: Driving Business Strategies with Data Science. Wiley.

Schulz, M., Neuhaus, U., Kaufmann, J., Badura, D., Kuehnel, S., Badwitz, W., Dann, D., Kloker, S., Alekozai, E. M., Lanquillon, C. (2020). Introducing DASC-PM: A data science process model. In: *Australasian Conference on Information Systems (ACIS) 2020 Proceedings, paper 45*.

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., Müller, K. R. (2021). Towards CRISP-ML (Q): A machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, 3(2), 392-413.

van der Aalst, W. (2016). Data science in action. In: *Process Mining*, 3-23. Springer, Berlin, Heidelberg.

Wirth, R., Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In: *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data mining*, 29-39.

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

Index of authors

The list of authors includes everyone who actively participated in revising and editing Version 1.1 and consented to being listed.

They would like to thank everyone who participated in Version 1.0 for their work on the previous draft.

Prof. Dr. Michael Schulz, NORDAKADEMIE Hochschule der Wirtschaft

Dipl.-Inform. Uwe Neuhaus, NORDAKADEMIE Hochschule der Wirtschaft

Prof. Dr. Jens Kaufmann, Hochschule Niederrhein

Dr. Stephan Kühnel, Martin-Luther-Universität Halle-Wittenberg

Dr. Emal M. Alekozai, Robert Bosch GmbH

Heiko Rohde (M.Sc.), valantic

Sayed Hoseini (M.Sc.), Hochschule Niederrhein

René Theuerkauf (M.Sc.), Martin-Luther-Universität Halle-Wittenberg

Daniel Badura, valantic

Prof. Dr. Ulrich Kerzel, IU Internationale Hochschule

Prof. Dr. Carsten Lanquillon, Hochschule Heilbronn

Prof. Dr. Stephan Daurer, DHBW Ravensburg

Prof. Dr. Maik Günther, IU Internationale Hochschule

Dr. Lukas Huber, FH Kufstein Tirol

Lukas-Walter Thiée, Universität Lüneburg

Philipp zur Heiden (M.Sc.), Universität Paderborn

Dr. Jens Passlick, VHV Gruppe

Jonas Dieckmann (B.Sc.), Philips

Dr. Florian Schwade, Universität Koblenz

Dr. Tobias Seyffarth, Martin-Luther-Universität Halle-Wittenberg

Wolfgang Badewitz, FZI Forschungszentrum Informatik

Dr. Raphael Rissler, SAP SE

Prof. Dr. Stefan Sackmann, Martin-Luther-Universität Halle-Wittenberg

Prof. Dr.-Ing. Philipp Gölzer, TH Nürnberg

Felix Welter, Universität Hamburg

Jochen Röth (M.Sc.), Shopfloor Management Systems GmbH

Julian Seidelmann (M.Sc.), Hapag-Lloyd

Prof. Dr. Uwe Haneke, Hochschule Karlsruhe

Appendix

The following pages present a survey that was given as part of preparing DASC-PM v1.1. When the model is applied to the “project order” phase, the survey helps to recognize and clarify the project’s key points, goals, and possible pitfalls. Answer possibilities are given for each question and/or a reference is made to a freely reworded version.

The final part of the document is the poster for DASC-PM v1.1. The corresponding page can also be provided in high quality for large format printouts. We also gladly provide you with the questionnaire for more comfortable use as a file for your spreadsheet program. The same applies to a set of presentation slides or individual graphics. Please contact info@dasc-pm.org in all cases.

Key area	Subtopic	Question		MC Items	Free text (possibly shown here: sample answers)
Instructions on using the questionnaire **(72 questions per key area / subtopic, ranked by average expected influence on project success)**				The values shown in this column are to be viewed line-by-line as multiple-choice answers. A "_____" signals that, when a selection is made, a specification must be indicated as well.	A white field in this column signals the option for a free text answer, possibly in addition to multiple-choice answers. If the fields contain text, this is to be viewed exclusively as a sample answer to clarify the intention of the question and can be deleted in the application. Grayed-out cells are not to be considered.
Domain	**Problem and objectives**	*What problem is the project supposed to solve?*			We wish to find out which of our customers will probably terminate their contract in the next year. We wish to find out what the conversation probability is for a visit to the online shop and what the main influence factors are.
Domain	**Problem and objectives**	*What objectives does the project pursue?*		Gain new findings about the main topic Increase skills Solve a new type of problem Open up new business sectors / target groups Save costs Save time Reduce workload	
Domain	**Problem and objectives**	*What results are expected?*		Manual application of the results (such as a seminar or workshop) Implementing the results (preparation in the form of a one-time report, for example) Implementing the model (applying the practiced model to new, unknown data) Continual learning (independently adjusting the model through repeated applications to unknown data) Publishing the developed procedure (possibly only within the organization)	
Domain	**Problem and objectives**	*How will success be measured?*		Relevant KPI: _____ Relevant technical metrics: _____ Compare to a baseline model Compare to prior status	With technical metrics, name examples
Domain	**Problem and objectives**	*What's the motivation for addressing the problem with a data science project?*		Classical data science problem, such as segmentation, classifying, regression Complexity of the topic / non-obvious connections Good prior experience with data science approaches Extensive / suitable database Failure of prior methods Curiosity about whether data science will deliver new findings	
Domain	**Problem and objectives**	*What project-related goals should definitely not be pursued?*			Replacing standard reporting Completely automating the decision processes

This work is licensed under a Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/

Key area	Subtopic	Question		MC Items	Free text (possibly shown here: sample answers)
Domain	**Participants and stakeholders**	*What organizational units are specifically involved?*		Management Division: _____ IT Data science team External: _____	
Domain	**Participants and stakeholders**	*What organizational units are specifically responsible?*		Management Division: _____ IT Data science team External: _____	
Domain	**Participants and stakeholders**	*Who commissioned the project?*		Management Division: _____ IT Data science team External: _____	
Domain	**Participants and stakeholders**	*Besides the project participants, what stakeholder groups provide input for the specialized aspects?*			Customers Legal department
Domain	**Participants and stakeholders**	*Who supports / promotes the project?*		Management Division: _____ IT Data science team External: _____ (such as science / politics)	
Domain	**Participants and stakeholders**	*Are there possible "troublemakers" for the project?*		Departments / Organizational units: _____ Individuals: _____ External: _____	
Domain	**Participants and stakeholders**	*What are the task fields of an external service provider?*		Project management IT infrastructure Data preparation Data analysis / model creation Operations / enhancement	

This work is licensed under a Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/

Key area	Subtopic	Question		MC Items	Free text (possibly shown here: sample answers)
Domain	**Project organization**	*What project management methods are envisaged?*		Agile project approach Waterfall model Continuous Integration DevOps approach No method / mixture of various methods	
Domain	**Project organization**	*Which roles participate in the project?*		Data scientist Data engineer Domain expert: _____ Project manager Technical support Compliance support	
Domain	**Project organization**	*What does the project's organizational form look like?*			Virtual team Hierarchical team structure
Domain	**Resources**	*What temporal framework conditions exist during the project execution through the submission of the results?*		Short-term deadline: _____weeks Long-term deadline Development in agile sprints _____ weeks	
Domain	**Resources**	*What skills do the project members have?*		Mathematics / statistics Information technology (such as programming knowledge, databases) Scope of application: _____ Communication Strategy Management	
Domain	**Resources**	*What financial framework conditions exist?*		Persons: _____ IT infrastructure: _____ External: _____	
Domain	**Resources**	*How much lead time exists until the project has to start?*		None Only a little: _____weeks Somewhat adequate / a lot: _____weeks Flexible start date	
Domain	**Prior experiences**	*What solution approaches already exist?*			Descriptive approach Models Automation Reporting
Domain	**Prior experiences**	*What experiences were collected through previous similar projects?*		None Partially transferable experiences Exactly transferable experiences Positive: _____ Negative: _____	
Domain	**Prior experiences**	*Where did difficulties exist in past projects?*		Complexity of the topic Database Project organization Project environment Personnel structure Quality of the models Interpreting the results	
Domain	**Prior experiences**	*What organizational units have prior experience with data science?*		Management Division: _____ IT	

This work is licensed under a Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/

Key area	Subtopic	Question		MC Items	Free text (possibly shown here: sample answers)
Data		*What data should be used (specialty / type)?*			Master data Sensor data Transaction data
Data		*What data are available in principle (specialty / type)?*			Master data Sensor data Transaction data
Data		*What is the status of data quality (completeness, freedom from errors, etc.)?*		High: _____ Medium: _____ Low: _____	
Data		*What data sources are relevant to the project?*		Operative data sources (such as ERP system, CRM system) Analytical data sources (such as data warehouse, data lake) Streaming data sources (such as sensor data) External data sources: _____	
Data		*Who is the data owner?*		Division: _____ IT Data science team	
Data		*Who is providing and preparing the data?*		Division: _____ IT Data science team External: _____	
Data		*Are the required data sources accessible?*		Yes No, the following must be done: _____	
Data		*How significant are data protection and data security?*		High: _____ Medium: _____ Low: _____	
Data		*Must new data be collected?*		Yes, completely Yes, predominantly Yes, to some extent Yes, a little No Still must be checked	
Data		*How high is the resource percentage of the data provision and preparation?*		High: _____ Medium: _____ Low: _____	
Data		*What is the structure of the data?*		Structured data: _____ Semi-structured data: _____ Unstructured data: _____	
Data		*How are the data accessed?*			Database API CSV files Crawler

This work is licensed under a Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/

Key area	Subtopic	Question		MC Items	Free text (possibly shown here: sample answers)
Analytical Methods		*Is it clear whether the problem can / should be answered with data science analyses?*		No, it's not yet clear No, testing the suitability of DS methods is the project objective Yes, because: _____	
Analytical Methods		*Is it already clear what type of analytical method is needed (classification, regression, clustering, identifying outliers, etc.)?*		No, work is still being done toward understanding the project No, no illustration of task types has been made yet Yes, it is clear; namely (specify task type): _____	
Analytical Methods		*Is it assumed that an established analytical method can be used, or must a new one be developed?*		Established procedures have not yet been evaluated Established procedures are being evaluated now No satisfactory solution has yet been found using established procedures	
Analytical Methods		*What special requirements are made of the analytical methods?*		None Term Scalability Robustness Data availability Explainability	

This work is licensed under a Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/

Key area	Subtopic	Question		MC Items	Free text (possibly shown here: sample answers)
Deployment		*Should the analysis models also be applied to future data?*		Yes, this has been planned, for (such as rolling forecasts): _____ No, because (knowledge is gained only once, for example): _____	
Deployment		*If the analysis models are to be used permanently: Is a stand-alone application planned, or should the models be integrated into existing operative systems?*		A stand-alone application is planned Integration in existing systems is sought	
Deployment		*Can it can be assumed that the analysis models must be continually adapted to new data?*		Yes, important Yes, in principle No, because... Data stable Data sources stable _____ Currently unknown	
Deployment		*How should the project result be conveyed to the stakeholders (report, workshops, seminars, etc.)?*		Summary / project report (written, presentation document, etc.) Digital report (in the sense of a "reporting") Final presentation Live demonstration (PoC, MVP, Final) Workshop Training	
Deployment		*What is necessary for subsequent users to be able to use the analysis models efficiently and correctly?*			Handbook Training Technical documentation
Deployment		*What operating concept is sought?*			Operation is taken on by a special MLOps team. The assurance of technical enhancement should be primarily achieved through subsequent projects.
Deployment		*Who is subsequently responsible for maintaining the data and analysis models and serves as a contact person regarding their content?*		Data scientist (also: team, etc.) Data engineer Range User IT External	
Deployment		*Who is responsible for enhancing the analysis models' content?*		Data scientist (also: team, etc.) Data engineer Range User IT External	
Deployment		*Who is taking on the integration of the analysis models into the organization's operative IT infrastructure?*		Data scientist (also: team, etc.) Data engineer Range User IT External	
Deployment		*Who will take on the technical operation of the analysis application later?*		Data scientist (also: team, etc.) Data engineer Range User IT External	

This work is licensed under a Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/

Key area	Subtopic	Question		MC Items	Free text (possibly shown here: sample answers)
Application		*How should the results be used?*		For concrete improvements (such as saving costs) For research and development As a basis for more extensive analyses To be determined	
Application		*What target group is mainly interested in the results?*		Management Division: _____ Data science team External: _____	
Application		*Who will use the results?*		Management Division: _____ Data science team External: _____	
Application		*How will the maintenance of the results be guaranteed?*			Determining threshold values as measurement criteria Regular content-related examinations of the results Tracking changes in distribution of the data
Application		*Is the usage potential limited by compliance requirements?*		Yes, as follows: _____ No	

This work is licensed under a Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/

Key area	Subtopic	Question		MC Items	Free text (possibly shown here: sample answers)
IT infrastructure		*What type of software will probably be needed to carry out the project?*			Database software Analysis software Visualization software
IT infrastructure		*What project-related software is already used in the company?*			Database software Analysis software Visualization software
IT infrastructure		*Is a close interlocking with the IT infrastructure necessary for the project?*		Yes No Depends on: _____	
IT infrastructure		*How high are the anticipated requirements for the capability of the hardware and software?*		High: _____ Medium: _____ Low: _____	
IT infrastructure		*What are the requirements for procuring/using certain software products for the project?*		Open source desired Software-as-a-Service solution Data privacy / compliance: _____ Software with official support/trainings	
IT infrastructure		*What hardware is available for implementing the project?*		Laptops/desktop PCs Server environment Comprehensive resources (for example, in the cloud)	
IT infrastructure		*What must be provided for additional hardware?*			Affected capacity Storage capacity

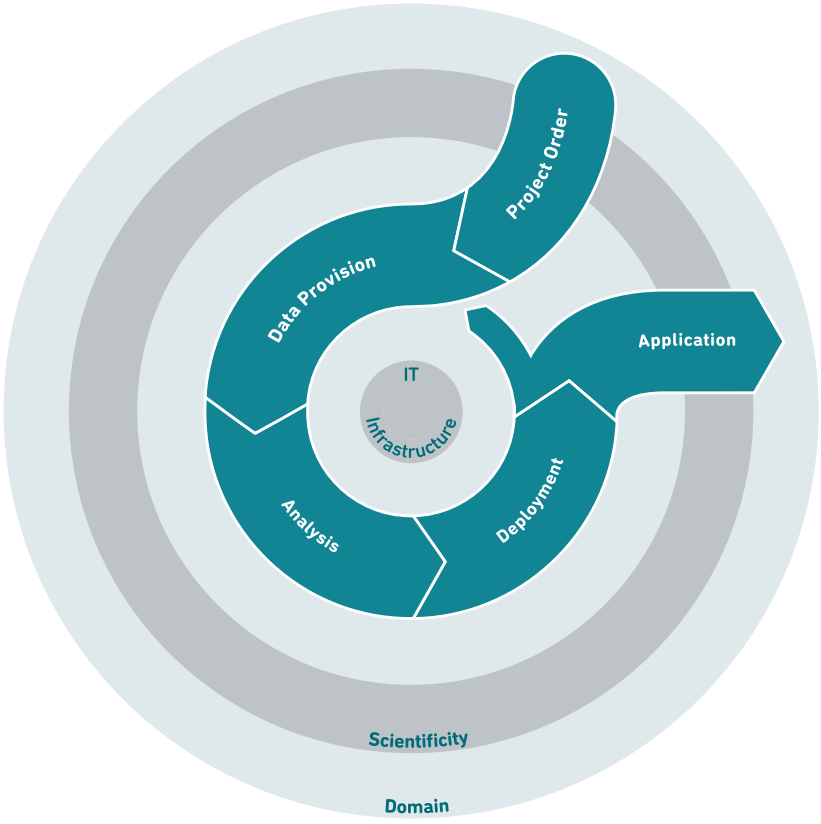
This work is licensed under a Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/

Key area	Subtopic	Question		MC Items	Free text (possibly shown here: sample answers)
Scientificity		*What process model is used?*			CRISP-DM DASC-PM KDD
Scientificity		*What methods are used to evaluate results?*			Target-actual comparison Comparison with baselines Expert survey
Scientificity		*Are there plans to provide the data used in the project to others as well?*		Yes, as follows: _____ Yes, to some extent: _____ No	
Scientificity		*Will representative results most likely be generated (meaning, can the results be generalized and used beyond the particular application context)?*		Yes No	
Scientificity		*What research paradigms underlie the project?*			Empirical quantitative Empirical qualitative Design oriented
Scientificity		*How will state-of-the-art scientific literature be considered?*		Extensive research Using standard literature Not at all, because: _____	
Scientificity		*Will the selected procedure be documented in detail?*		Yes, the entire process, including all interim steps, can be reproduced by third parties Yes, but only the results can be reproduced by third parties No, detailed documentation is unnecessary	
Scientificity		*How will the project findings most likely be shared with others?*		Forwarded to selected third parties Free publication of the results	
Scientificity		*Which evaluation criteria (quality levels) will be consulted?*			
Scientificity		*Does the project make its own contribution to research?*		Yes, a research contribution: _____ No, a standard approach is being replicated	

This work is licensed under a Creative Commons Attribution 4.0 International License. https://creativecommons.org/licenses/by/4.0/

dasc°pm v1.1

DASC-PM is a process model for data science projects. It describes the key areas relevant to the project and the phases to be completed. It explains the typical tasks within the phases and depicts the project roles involved and the required competencies.



Overarching key areas

Domain

At many points in a data science process, broad background knowledge of the domain is needed. Examples are the identification of the analysis target or the correct understanding of data, its origin, quality, and connections. Other examples include assessment and classification of analysis results in the application as well as subsequent practical use. The area "Domain" also encompasses rating strengths and weaknesses of existing solutions, conducting requirements analyses, supporting parametrization of models, and finally evaluating the success of the project. Legal, social, and ethical aspects of data science projects will also be discussed here

Scientificity

Just because data science projects are scientific in nature does not mean they claim to be complete, formalized, academic, or consistently research-oriented in general. Although this might certainly be the case for research projects, the aspect of their scientificity within a business context primarily refers to a solid methodology: a typically expected characteristic or minimum requirement of scientific work.
The defined project order must be processed methodically in every project phase. Special mention must be made here of the project management and a structured processing that is placed in the foreground by using a process model. Details on the degree of scientificity required must be established while considering the project situation and domain specifics.

IT Infrastructure

All the steps that a data science project traverses depend on the underlying IT infrastructure; the actual extent of IT support, however, should be individually assessed for each project. Even if the use of specific hardware and software is frequently determined within the organization, the limiting and empowering characteristics of the IT infrastructure (as well as the possibility of expanding the infrastructure, if applicable) must be considered in all project phases.

Roles

Core role „Data Scientist"

Data Scientists are specialists for the analysis area of a data science project.

Core role „Data Engineer"

Data Engineers take care of the procurement, storage, preparation, structuring and dissemination of data.

Core role „Domain Expert"

Domain experts are business users or representatives of business users.

Core role „Project Manager"

Project managers plan, control and coordinate the overall course of a data science project.

Supplementary role „Technical Support"

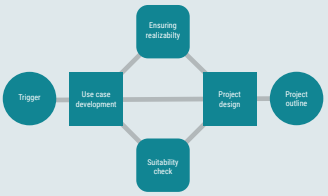
Technical support includes all tasks that need to be performed to ensure the technical prerequisites for the implementation of the data science project.

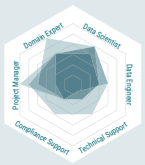
Supplementary role „Compliance Support"

Compliance support is responsible for ensuring legal requirements, compatibility with internal rules and regulations, and the correct behavior of project staff.

Project Order

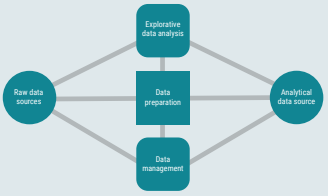
Problems existing within a domain trigger a use-case development. The promising use cases are subsequently configured to a data science project outline. All associated tasks are reflected in the project order phase. Through the early, relatively comprehensive consideration of the project, comprehensive abilities in almost all skill areas are also frequently required here.





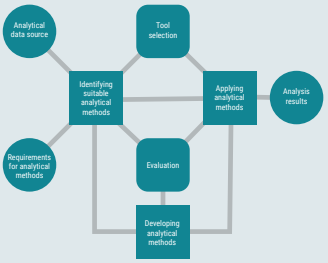
Data Provision

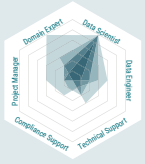
Within the data provision phase, all activities that are allocable to the data key area are summarized, which is why the term used is broadly formulated. The phase contains the data preparation (from recording to storage), data management, and an explorative analysis. This phase results in a data source that is suited to further analysis.



Analysis

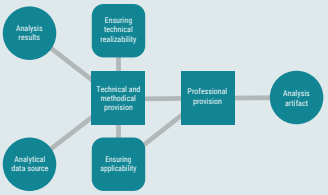
In a data science project, either existing procedures can be used or a new procedure developed—the decision in question is a separate challenge. The phase, therefore, includes not only performing the analysis, but also related activities. The artifact of the phase is an analysis result that has traversed a methodical and technical evaluation.





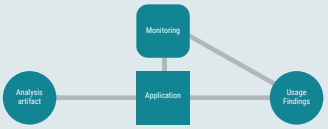
Deployment

In the "deployment" phase, an applicable form of the analysis results is created. Depending on the project, this can entail comprehensively considering technical, methodological, and professional tasks, or it can be handled pragmatically. The analysis artifact can include results as well as models or procedures and is provided to its target recipients in various forms.



Application

Using artifacts after the project performance is not considered a primary part of a data science project. Monitoring is necessary, however (depending on the form of Deployment), to check the model's continuing suitability in the application and obtain findings from the application for ongoing and new developments (including developments for the purposes of iterative approaches).



This work is licensed under a Creative Commons Attribution 4.0 International License.
https://creativecommons.org/licenses/by/4.0/

Published by
NORDAKADEMIE gAG Hochschule der Wirtschaft
Köllner Chaussee 11, 25337 Elmshorn

Supported by the NORDAKADEMIE foundation

Authors:
Schulz, M., Neuhaus, U., Kaufmann, J., Kühnel, S., Alekozai, E.M., Rohde, H., Hoxeini, S., Theuerkauf, R., Badura, D., Kanzel, U., Langallon, C., Daane, S., Günther, M., Huber, L., Thole, S.-W., zur Heiden, P., Pavsicek, J., Dieckmann, J., Schwade, F., Seyffarth, T., Badewitz, W., Rissler, R., Sackmann, S., Gölzer, P., Weller, F., Roth, J., Seidensaum, J., Hemeke, G. (2022): DASC-PM v1.1 – A Process Model for Data Science Projects.

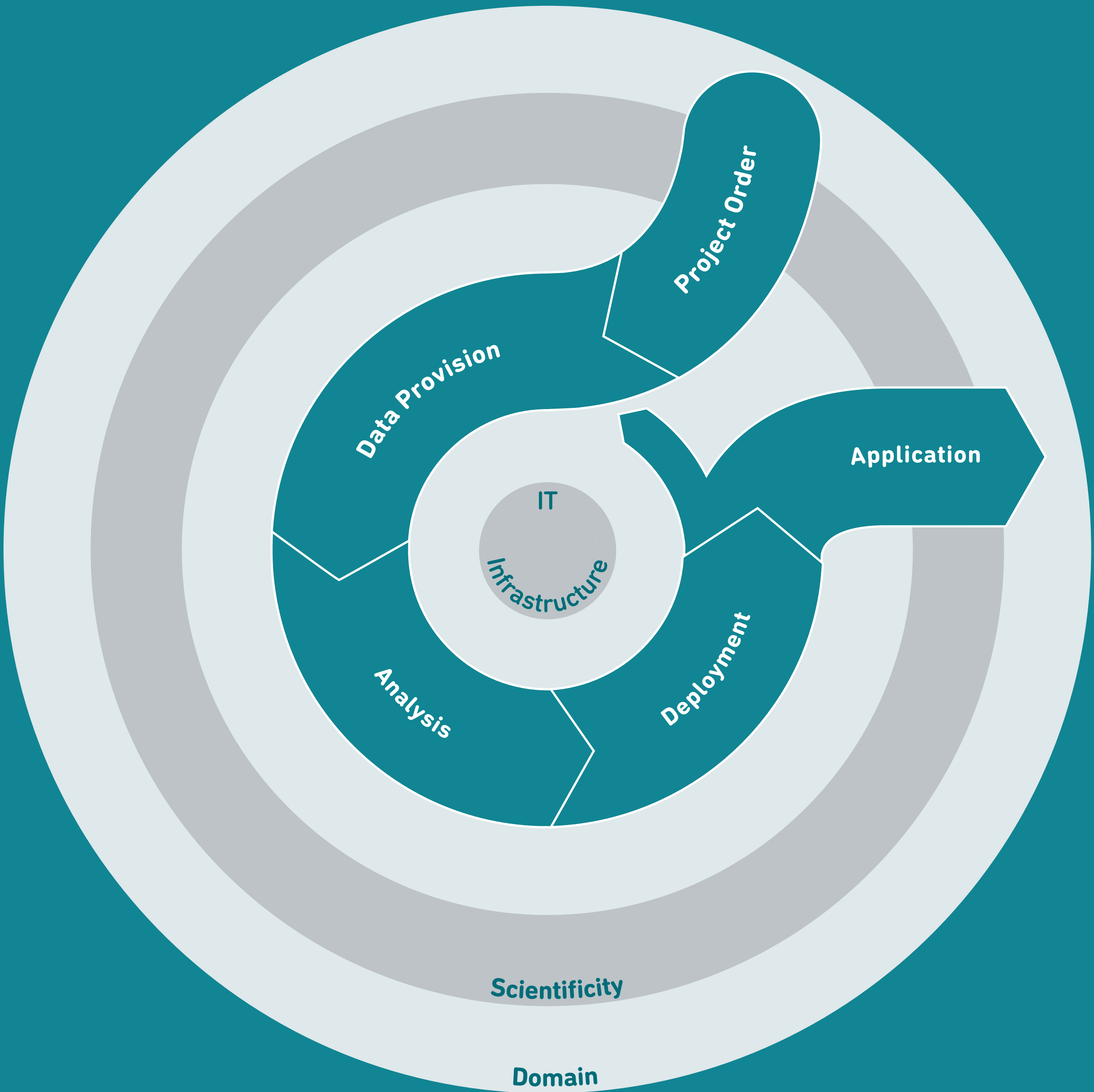
Graphic design: Fritjof Wild

Feature-bearing Area
Core Task
Accompanying Tasks
Interface Tasks

Competencies
Roles

dasc°pm v1.1

DASC-PM is a process model for data science projects. It describes the key areas relevant to the project and the phases to be completed. It explains the typical tasks within the phases and depicts the project roles involved and the required competencies.



This work is licensed under a Creative Commons Attribution 4.0 International License.
https://creativecommons.org/licenses/by/4.0/

Published by:
NORDAKADEMIE gAG Hochschule der Wirtschaft
Köllner Chaussee 11, 25337 Elmshorn

Supported by the NORDAKADEMIE foundation

Authors:
Schulz, M., Neuhaus, U., Kaufmann, J., Kühnel, S., Alekozai, E.M., Rohde, H., Hoseini, S., Theuerkauf, R., Badura, D., Kerzel, U., Lanquillon, C., Daurer, S., Günther, M., Huber, L., Thiée, L.-W., zur Heiden, P., Passlick, J., Dieckmann, J., Schwade, F., Seyffarth, T., Badewitz, W., Rissler, R., Sackmann, S., Gölzer, P., Welter, F., Röth, J., Seidelmann, J., Haneke, U. (2022): DASC-PM v1.1 – A Process Model for Data Science Projects.

Graphic design: Fritjof Wild

dasc°pm v1.1

ISBN: 978-3-9824465-1-6