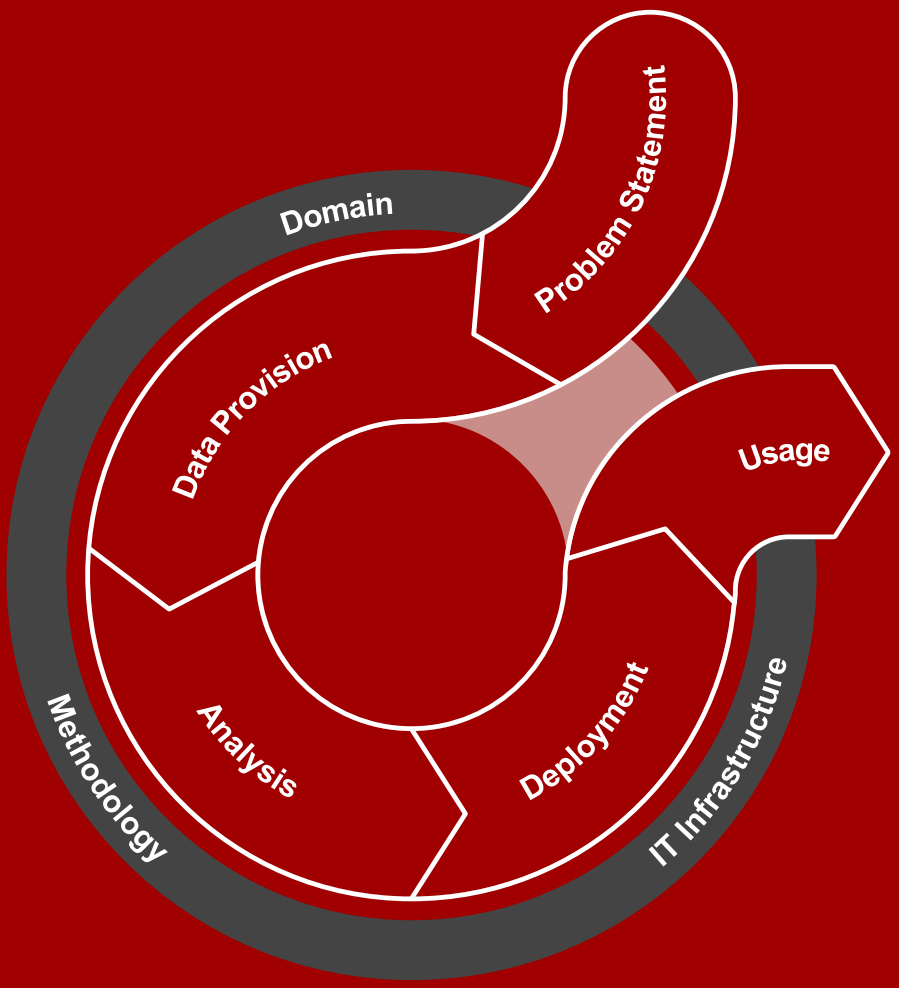
Michael Schulz ▪ Jens Kaufmann ▪ Stephan Kühnel ▪ Uwe Neuhaus

DASC-PM

A Process Model for Data Science and AI Projects

v2.0

Heiko Rohde ▪ René Theuerkauf ▪ Carsten Lanquillon ▪ Stephan Daurer ▪ Jonas Dieckmann ▪ Stefan Sackmann ▪ Felix Welter ▪ Uwe Haneke ▪ Christian Beecks ▪ Martin Böhmer ▪ Bahne Christiansen ▪ André Drews ▪ Arne Ewald ▪ Viktor Harkov ▪ Franziska Herrmann ▪ Tim Hilbig ▪ Dirk Johannßen ▪ Lutz Kretschmann ▪ Bernhard Meussen ▪ Christian Pasold ▪ Stefan Rösl ▪ Klaus Schmerler ▪ Theo Schnelle de Lourenco ▪ Martin Schultz ▪ Michael Seifert ▪ Marcus Soll ▪ Robert Stahlbock ▪ Niklas Ullmann



The current version of the work
DASC-PM - A Process Model for Data Science and AI Projects – v2.0
is based on the publication

Schulz, M., Kaufmann, J., Kühnel, S., Neuhaus, U., Rohde, H., Theuerkauf, R., Lanquillon, C., Daurer, S., Dieckmann, J., Sackmann, S., Welter, F., Haneke, U., Beecks, C., Böhmer, M., Christiansen, B., Drews, A., Ewald, A., Harkov, V., Herrmann, F., Hilbig, T., Johannßen, D., Kretschmann, L., Meussen, B., Pasold, C., Rösl, S., Schmerler, K., Schnelle de Lourenco, T., Schultz, M., Seifert, M., Soll, M., Stahlbock, R., Ullmann, N.

***DASC-PM - Ein Vorgehensmodell für Data-Science- und KI-Projekte* – v2.0**

NORDAKADEMIE gAG Hochschule der Wirtschaft, Elmshorn, 2022
ISBN: 978-3-9824465-4-7, https://doi.org/10.25673/123274.

ISBN: 978-3-9824465-5-4 (eBook)

DOI: 10.25673/124050

Elmshorn 2026

info@dasc-pm.org

Publisher

NORDAKADEMIE gAG Hochschule der Wirtschaft
Köllner Chaussee 11
25337 Elmshorn

Funding Note

Most of the substantive changes compared to Version 1.1 were developed during a workshop. The workshop itself, as well as the publication of this document, was made possible through funding provided by the NORDAKADEMIE Foundation.

The preparation of the results in the area of competencies and roles was supported by the *Digital Learning Campus* project. The Digital Learning Campus is funded by the State of Schleswig-Holstein and the European Union. Further information is available at: dlc.sh

Model graphic: With the kind support of Eike Behnke.

Table of Contents

Preface to Version 2.0 ..... **I**

Part A General Considerations and the Overall Model **1**

1 Data Science and Artificial Intelligence **2**
1.1 Characteristics of Data Science 3
1.2 The Interrelationship Between Artificial Intelligence and Data Science 6

2 Process Models for Data Science Projects **7**

3 DASC-PM as a Process Model **8**
3.1 Phases 9
3.2 Overarching Aspects of the Model 9
3.3 Iterations and Project Termination 10

4 Competencies and Roles **11**
4.1 Competencies in DASC-PM 12
4.2 Roles in DASC-PM 15

Part B DASC-PM in Detail **19**

5 Problem Statement **22**
5.1 Feature-Bearing Area "Trigger" 24
5.2 Core Task "Use Case Development" 25
5.3 Accompanying Task "Suitability Check" 29
5.4 Accompanying Task "Ensuring Realizability" 30
5.5 Core Task "Project Design" 31
5.6 Feature-Bearing Area "Project Outline" 31

6 Data Provision **32**
6.1 Feature-Bearing Area "Raw Data Sources" 34
6.2 Core Task "Data Preparation" 36
6.3 Accompanying Task "Data Management" 37
6.4 Accompanying Task "Exploratory Data Analysis" 38
6.5 Feature-Bearing Area "Analytical Data Source" 39

7 Analysis **40**
7.1 Feature-Bearing Area "Analytical Data Source" 43
7.2 Feature-Bearing Area "Requirements for Analytical Methods" 43
7.3 Core Task "Identifying Suitable Analytical Methods" 44
7.4 Core Task "Applying Analytical Methods" 45
7.5 Accompanying Task "Tool Selection" 46
7.6 Core Task "Developing Analytical Methods" 47
7.7 Accompanying Task "Evaluation" 48
7.8 Feature-Bearing Area "Analysis Results" 49

8 Deployment **50**
8.1 Feature-Bearing Area "Analysis Results" 52
8.2 Feature-Bearing Area "Analytical Data Source" 52
8.3 Core Task "Technical and Methodological Provision" 52
8.4 Accompanying Task "Ensuring Technical Feasibility" 54
8.5 Accompanying Task "Ensuring Applicability" 55
8.6 Core Task "Professional Provision" 56
8.7 Feature-Bearing Area "Analysis Artifacts" 57

9 Usage **58**
9.1 Feature-Bearing Area "Analysis Artifacts" 60
9.2 Accompanying Task "Monitoring" 60
9.3 Feature-Bearing Area "Usage Findings" 61

10 Overarching Aspects **62**
10.1 Domain 63
10.2 Methodology 64
10.3 IT Infrastructure 65

References **67**

Further Publications on DASC-PM **68**

Index of Authors **71**

List of Figures

Figure 1: Characteristics of data science 2
Figure 2: Artificial intelligence, machine learning, and deep learning 6
Figure 3: DASC-PM 8
Figure 4: Competencies required of data scientists, based on Conway (2010) 11
Figure 5: Competencies required in a data science project 12
Figure 6: Competencies needed for a data science project in an exemplary profile 14
Figure 7: Role areas and selected responsibilities within a data science project 15
Figure 8: Nomenclature used and notation in the phases 21
Figure 9: Brief overview of the “Problem Statement” phase 22
Figure 10: Competence profile for the “Problem Statement” phase 22
Figure 11: Detailed presentation of the “Problem Statement” phase 23
Figure 12: Brief overview of the “Data Provision” phase 32
Figure 13: Competence profile for the “Data Provision” phase 32
Figure 14: Detailed presentation of the “Data Provision” phase 33
Figure 15: Brief overview of the “Analysis” phase 40
Figure 16: Competence profile for the “Analysis” phase 40
Figure 17: Detailed presentation of the “Analysis” phase 41
Figure 18: Brief overview of the “Deployment” phase 50
Figure 19: Competence profile for the “Deployment” phase 50
Figure 20: Detailed presentation of the “Deployment” phase 51
Figure 21: Brief overview of the “Usage” phase 58
Figure 22: Competence profile for the “Usage” phase 58
Figure 23: Detailed presentation of the “Usage” phase 59

Preface to Version 2.0

With the present Version 2.0 of DASC-PM, the original process model has been further developed both in terms of content and structure. Building on the experience gained from applying Versions 1.0 and 1.1, the model was revised in order to improve its comprehensibility and to facilitate its use in different project contexts.

Since the publication of the first version in 2020, the environment of data-driven projects has continued to evolve dynamically. In particular, advances in the field of artificial intelligence have given rise to new requirements for methods, infrastructure, and project organization. At the same time, numerous practical applications have shown which elements of the model have proven effective and where adjustments are appropriate. The present version incorporates these experiences and brings them together in a revised form.

A central objective of Version 2.0 is to present the model with a stronger focus on results. Earlier versions of the document contained references in several places to the model's development process and to the contributions made by individual members of the working group. In the present version, greater emphasis is placed on the model itself and on the structures and tasks derived from it. In this context, the structure of the document was also revised. The key areas presented in earlier versions originally served to structure the development of the model and formed the basis for deriving the project phases. In practical application, however, it became apparent that their additional presentation alongside the model's phases could lead to ambiguity. Version 2.0 therefore places greater emphasis on the process model itself.

The fundamental phases of DASC-PM are retained. However, they were reviewed in terms of content and partly refined in order to ensure their applicability to different types of data- and analysis-oriented projects. Particular attention was paid to ensuring that activities in the context of artificial intelligence can be taken into account in all phases of the model. The aim of these adjustments is to ensure the model's broad applicability without narrowing it to a specific class of AI projects.

The presentation of the competencies required in the project context was revised. Feedback from workshops and working groups was consolidated and evaluated in a structured manner. In order to ensure that the model remains stable over the long term despite changing job titles, more emphasis is placed on working with role types. A distinction is made between roles at the core of the project, collaboration roles, and roles with steering and influencing capabilities. Specific role titles will continue to be mentioned as examples, but they will no longer be at the center of the model description.

In addition, both the presentation of the model and selected terms used in the model figure were revised. To simplify language use, this document furthermore refers to DASC-PM only by its abbreviation and no longer by its full name (Data Science Process Model).

As in the previous versions, DASC-PM continues to be understood as a structuring framework for data- and analysis-oriented projects. The model is intended to make the typical tasks, roles, and interrelationships in such projects comprehensible and to establish a common basis for communication among the various stakeholders involved. We hope that the present version will help to further improve the understanding of the structure of data science and AI projects and provide users with practical guidance for their own projects.

Elmshorn, Flensburg, Halle (Saale), and Mönchengladbach, Germany in June 2026

The DASC-PM Core Team

Part A
General Considerations and the Overall Model

1 Data Science and Artificial Intelligence

In recent years, the term *data science* has gained noticeable importance both in academia and in practice. In numerous organizations, data-driven analytical methods are used to identify relationships in large data sets, generate forecasts, or support decision-making processes. At the same time, a wide range of terms has become established, some of which describe similar or overlapping concepts.

Despite the increasing prevalence of the term *data science*, there is still no generally accepted and uniform definition. In scientific publications, the term is used across different disciplines and is consistently interpreted from the perspective of the respective field. In practical application, the range of uses is often even broader. As a result, different expectations arise with regard to the content, methods, and objectives of data science activities.

Against this background, the present document aims to establish a consistent understanding of the terminology used in the context of data science projects. In this regard, data science is understood as an interdisciplinary field that combines methods from various scientific disciplines and can be applied across different domains.

Based on this understanding, the following section presents a definition of data science that serves as the starting point for the subsequent discussion in this document. The characteristics of this definition are shown in Figure 1 and are examined in greater detail below.

Data science is a field of interdisciplinary expertise in which a scientific approach is 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.

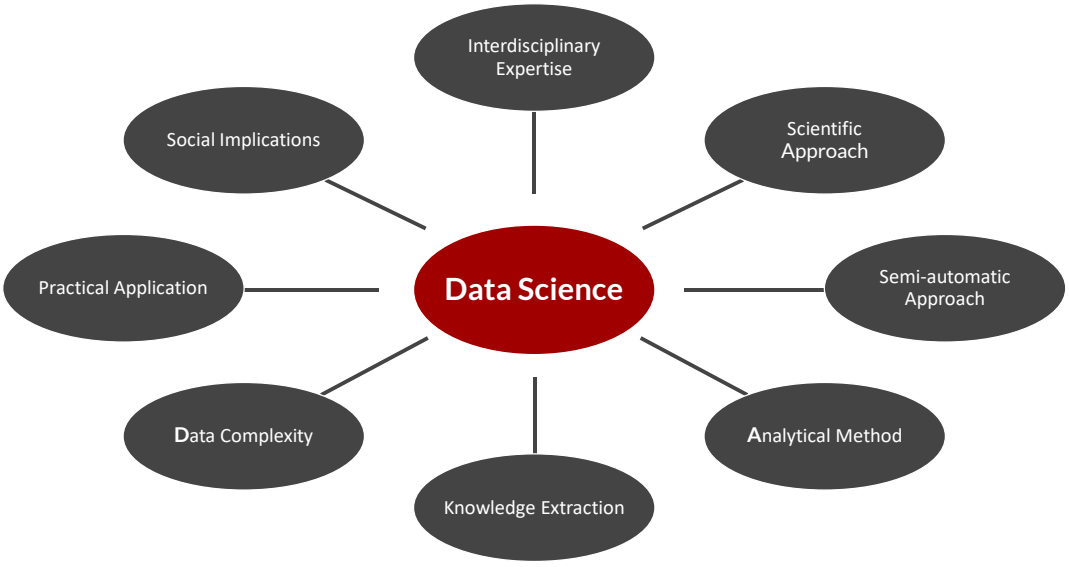


Figure 1: Characteristics of data science

1.1 Characteristics of Data Science

Interdisciplinary Expertise

Interdisciplinarity is a central characteristic of data science that arises from the necessary collaboration of different scientific disciplines, which jointly contribute to the analysis of complex data. Of particular relevance in this context are mathematics—especially statistics and numerical mathematics—and computer science. Other disciplines, such as linguistics, also contribute to the methodological and conceptual examination of data. In all of these fields, approaches to the scientific analysis and interpretation of data have existed for some time (Provost & Fawcett, 2013).

The field subsumed under the term *data science* has developed particularly in areas where the methods of individual disciplines were no longer sufficient on their own to address current challenges in dealing with data. These include, for example, rapidly growing volumes of data, unstructured data formats, or dynamically generated data streams. At the same time, technological developments—such as the increasing availability of digital information, declining costs of data storage, and growing computing power—have enabled the use of increasingly complex analytical methods (McAfee & Brynjolfsson, 2012; Donoho, 2017). These developments have made a significant contribution to the establishment of data science as an independent interdisciplinary field.

Interdisciplinarity is also reflected in the variety of application domains in which data science is used. While many accounts traditionally locate the field primarily in an economic context, its application is not limited to that area. Rather, data science is applied across numerous scientific and practical domains, including biology, medicine, physics, and astronomy. In these and many other fields, it supports the analysis of complex data sets and contributes to the generation of new insights.

Scientific Approach

The term *data science* already suggests that data analysis is not understood solely as a technical activity, but is based on a scientifically informed approach. This scientific character is reflected, among other things, in the fact that data science activities often aim at generating generalizable insights. In addition to the direct use of analysis results, the focus may therefore also be on broader questions, such as examining the suitability of particular methods for specific problem contexts, assessing their informative value in relation to the underlying data basis, or analyzing the complexity of different methods.

A scientific approach is further characterized by the fact that the problem contexts under investigation are not trivial and that the methods applied are used in a transparent, systematic, and reproducible manner. This includes, in particular, documenting analytical processes and ensuring that their results remain verifiable (in principle). These aspects help to ensure that insights gained from data analyses are not only useful in a specific situation, but can also be situated within a broader professional context.

This scientific foundation also plays a role in business applications. Many of the analytical methods used today were originally developed in scientific contexts and subsequently transferred to practical applications. The depth of scientific engagement may vary depending on the field of application. While research projects often involve the development of new methods or the detailed examination of existing ones, many practical applications focus primarily on the appropriate use of established methods. In such contexts, the approach may take on the character of an engineering-oriented use of established methods.

Analytical Method

The explicit inclusion of individual algorithms or groups of algorithms in a definition of data science is not appropriate given the rapid pace of development in the field. Such a specification would implicitly limit the scope to certain methods and would not adequately account for future developments. Moreover, the term *algorithm* is not always appropriate in the present context, since not all analyses are based exclusively on algorithmic procedures, and many of the methods used were also developed outside the field of data science.

For this reason, the more general term *analytical method* is used in the definition. In connection with its application to data, it describes the methodological core of data science. Analytical methods can take very different forms. These include, among others, hypothesis-testing and non-hypothesis-driven analyses with descriptive, predictive, or prescriptive objectives. Such analyses are often aimed at identifying patterns, trends, and correlations in data or to support optimization processes.

The analytical methods used in a specific project depend largely on the particular use case and on the data available. In many cases, existing methods can be applied. However, it may also be necessary to further refine analytical methods or to develop new approaches if no suitable methods are available for a specific problem context.

Semi-Automatic Approach

Since analytical methods usually cannot be fully automated, their application is semi-automated and includes both human and machine-based activities. Hybrid learning approaches are also worth mentioning, as they are specifically designed to promote the interplay between expert knowledge and analytical methods (Olivotti et al., 2018). This often requires high-performance hardware and software platforms which, taken together, form a complex infrastructure. In addition, developments in the field of generative artificial intelligence, particularly with regard to large language models, are giving rise to tools that can enable a higher degree of automation for certain types of analysis in the preparation, execution, and interpretation of results.

Depending on the specific scenario, a high degree of automation may be the goal. This usually requires preparatory manual steps. Ultimately, however, the insights gained from data analysis remain dependent on the involvement of human actors.

Knowledge Extraction and Data Complexity

One objective of data science is to extract insights from data that are often complex. Such data may differ, for example, with regard to their structure, quality, completeness, size, and dimensionality. They may take the form of either static data or data streams. In addition, data can be related to one another in diverse and sometimes complex ways. The analysis of large and heterogeneous data sets once required the development of new methods, which were often grouped under the term *big data*.

Before analytical methods can be applied, the data generally first needs to be extracted from source systems, prepared, and made available in a suitable format. These steps also often require complex technical infrastructure.

Practical Application

Data science encompasses not only the extraction of insights from data, but also their practical application. This may involve, for example, providing analysis results to domain experts or other users. It may also include integrating the results into existing systems or applying analytical models to new data in an automated manner. Various authors place the creation of economic value at the center of data science projects. In the present definition, however, the more general term *practical application* is deliberately chosen to include not only economic objectives but also scientific or other forms of use.

In many cases, both the semi-automated extraction of insights and the preparation and provision of data, as well as their subsequent practical application, require an appropriate IT infrastructure. This includes hardware and software components that must be adapted to the respective project requirements. Examples include scalable system architectures, the processing of distributed data, and the use of cloud infrastructures.

Social Implications

The use of data-driven analytical methods can have a wide range of societal impacts. Data science therefore also includes addressing the ethical, legal, and societal issues that arise from the use of data and analytical methods. This applies both to the input data and to the analysis results derived from them.

A central aspect is the responsible handling of data. In particular, when dealing with legally protected or personal data, the relevant legal framework conditions must be taken into account, for example in the context of data protection or data security. Furthermore, issues of transparency and the traceability of algorithmic decisions may also play an important role, especially when analytical models are integrated into decision-making processes.

In addition to legal requirements, ethical issues are also receiving increasing attention. These include, among other things, potential biases in data sets or analytical methods that may lead to the systematic disadvantaging of specific groups of people. Another question concerns the extent to which automated analytical systems should prepare, support, or make decisions and what role human responsibility should play in such contexts.

Therefore, data science is not understood solely as a technical discipline. Rather, active participation in the societal discourse on the opportunities, risks, and limits of data-driven analyses is also part of the responsibilities of this field. The aim is to shape the use of data and analytical methods in such a way that their potential can be realized without neglecting societal or legal framework conditions.

1.2 The Interrelationship Between Artificial Intelligence and Data Science

In recent years, alongside data science, artificial intelligence (AI) has also gained significant importance. The terms are often used in similar contexts in scientific publications, public discourse, and practical application. At the same time, it becomes apparent that their meanings partly overlap, while in some respects they are interpreted differently depending on the context.

In principle, AI describes a broad field of research rooted in computer science that is concerned with the development of systems capable of imitating human abilities such as logical reasoning, learning, planning, and creativity (adapted from Europ. Parl., 2026). Examples include pattern recognition, decision-making, and the processing of natural language. Within this field of research, various subfields have emerged, including machine learning (ML). Machine learning refers to a class of methods in which models are generally trained on the basis of existing data. Deep learning represents a specific form of such methods in which multilayered neural networks are used, for example for the processing of images, text, or speech.

The relationships between these terms are often presented in the form of overlapping or nested concepts. One such simplified classification presents machine learning as a subfield of artificial intelligence, while deep learning in turn represents a specific class of machine learning methods. Figure 2 illustrates this by way of example.

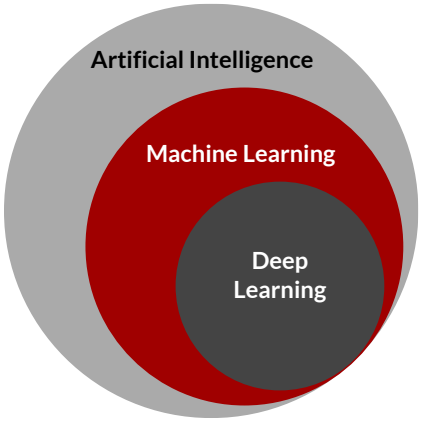


Figure 2: Artificial intelligence, machine learning, and deep learning

Data science has close points of contact with the areas described above in many respects. In particular, ML methods are frequently used to derive models from data or to generate forecasts. At the same time, data science encompasses a broad range of additional activities. In addition to the analysis itself, these include the procurement and preparation of data, the selection of suitable analytical methods, the interpretation of the results in the context of an application domain, and their practical application. Other areas of AI, such as robotics, are in turn not the focus of data science.

Given their distinct historical roots and trajectories of development, data science and AI can neither be completely separated from one another nor clearly arranged in a hierarchical relationship. In many practical applications, the two fields overlap. Depending on the perspective, it may therefore be appropriate to speak of data science projects, AI projects, ML projects, or other related variants.

To simplify language usage and account for the model's ongoing development, DASC-PM primarily uses the terms "data science" and "data science projects" to describe the activities at the center of attention. However, regardless of terminology, DASC-PM can be applied to many variants of the project types mentioned above.

2 Process Models for Data Science Projects

The execution of data science projects is generally associated with a wide range of different tasks. These range from the identification of suitable questions to the procurement and preparation of data, and on to the development, evaluation, and use of analytical models. In addition, there are organizational, technical, and domain-specific requirements that must be taken into account over the course of a project. Against this background, the question arises as to how such projects can be planned and carried out in a structured manner.

Process models provide a framework for this by describing the typical activities within a project and clarifying their interrelationships. They serve to structure complex project workflows, support communication among the different stakeholders involved, and create a shared understanding of tasks, results, and responsibilities. At the same time, they can help to increase the transparency and replicability of analytical activities.

In the context of data-driven analysis, process models were developed at an early stage to provide a systematic structure for such projects. Among the best-known models are the *Knowledge Discovery in Databases process* (KDD; Fayyad et al., 1996) and the *Cross Industry Standard Process for Data Mining* (CRISP-DM; Wirth & Hipp, 2000). Both models describe a sequence of typical activities ranging from data preparation and exploration to the application of analytical methods and the evaluation of results. Beyond that, other models developed specifically for the field of data science can also be considered.

These models have made a significant contribution to structuring data-oriented analytical projects and establishing a shared understanding of key project activities. At the same time, practical application shows that data science projects often need to take additional aspects into account. These include, for example, the integration of analysis results into operational systems, the collaboration among interdisciplinary teams, and the development of suitable technical infrastructures. Particularly in connection with CRISP-DM, efforts can be identified that seek to adapt this model to the requirements of data science projects, for example CRISP-ML(Q) (Studer et al., 2021).

DASC-PM takes up these considerations in the context of current developments and provides a process model specifically tailored to the requirements of modern data science and AI projects. The aim of the model is to present the typical phases and activities of such projects in a structured manner and to make their interrelationships comprehensible. In doing so, the model is intended both to serve as guidance for the planning and execution of projects and to provide a shared basis for communication among the individuals and organizational units involved.

3 DASC-PM as a Process Model

Building on the foundations outlined above, DASC-PM is presented below as a process model for data science projects (see Figure 3). The model describes typical phases and activities that may arise in their planning and execution. Its aim is to provide a structured presentation of these activities and to make their interrelationships comprehensible.

Data science projects are often characterized by a high degree of complexity. Different areas of expertise, technical infrastructures, and domain-specific requirements must be brought together in order to generate robust insights from data and subsequently put them to practical use. DASC-PM provides a framework for this by structuring typical areas of activity and helping to facilitate communication among the people involved.

The model is not intended to represent a rigid process. Rather, it describes a structured view of activities that may occur in different forms across different projects. Depending on the use case, the organizational context, or the technical infrastructure, individual steps may vary in scope or be carried out repeatedly. In the graphical representation of the model, this characteristic is illustrated by the light red segment of the circle indicating iterativeness. Section 3.3 returns to this aspect separately.

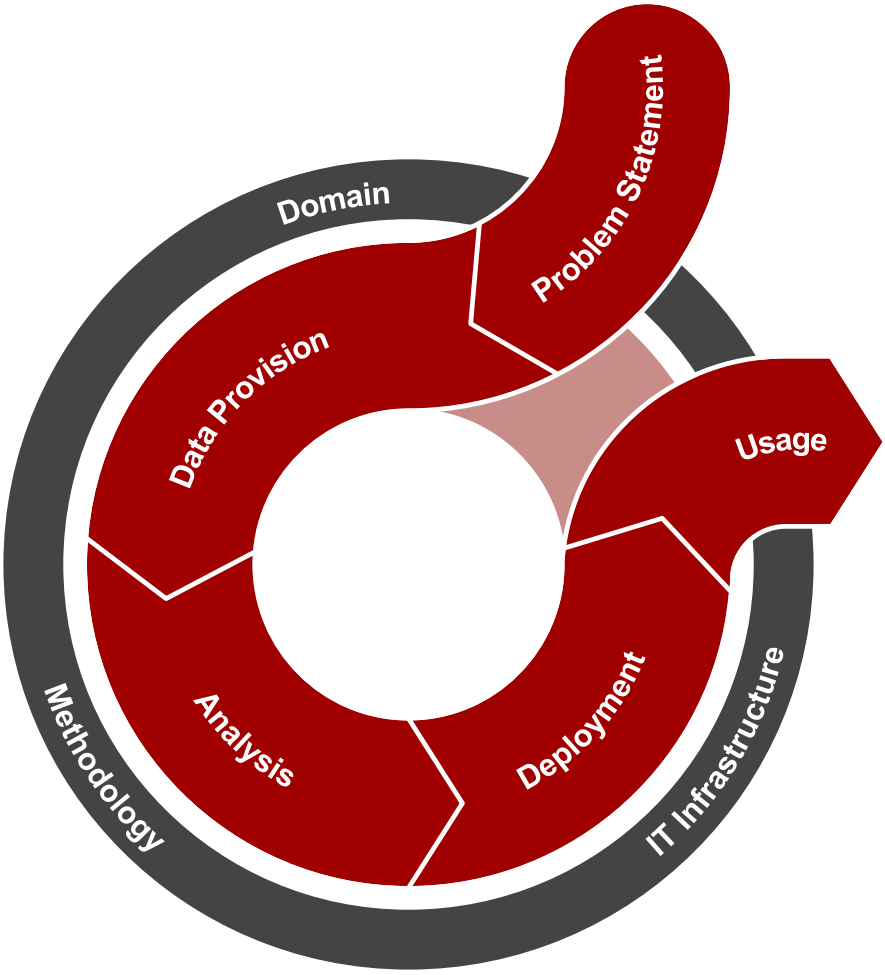


Figure 3: DASC-PM

3.1 Phases

Problem Statement

The problem statement forms the starting point of a data science project. Questions or problems arising within a domain can trigger the development of a data-driven use case. In this phase, potential use cases are identified, assessed, and further specified. The aim is to formulate a clear problem description and to produce a project outline that serves as the basis for the subsequent work.

Data Provision

The *Data Provision* phase comprises all activities associated with the procurement, preparation, and structuring of data. These include, for example, the identification of relevant data sources, the integration of different data sets, and the cleansing and transformation of data. Exploratory analyses may also form part of this phase in order to gain an initial understanding of the available data. The outcome of this phase is a data basis that is suitable for the subsequent analysis.

Analysis

In the *Analysis* phase, suitable analytical methods are selected and applied to the available data. Depending on the problem context, existing methods may be used or new methods may be developed. In addition to the actual application of analytical methods, this phase also includes activities such as the selection of suitable tools and the evaluation of the results. The aim is to derive robust insights from the data and to assess their informative value.

Deployment

In this phase, the results of the analysis are transformed into a form that enables practical application. This may take place, for example, through the integration of analytical models into existing systems, the provision of useful visualizations, or other forms of presenting the results. Depending on the respective project, both technical and domain-specific aspects may need to be taken into account.

Usage

The use of analysis artifacts often takes place outside the actual data science project. Nevertheless, important insights may arise from their practical application, for example with regard to model quality or potential improvements. Monitoring this usage can help to identify the need for adjustments or derive new project approaches.

3.2 Overarching Aspects of the Model

In addition to the project phases described above, DASC-PM includes several aspects that must be considered across all phases of the model. These aspects provide a framework for the successful execution of data science projects and shape the design of the individual activities.

Domain

Data science projects are always carried out within a specific application domain. In addition to the problem statement as the starting point of a project, domain-specific requirements or framework conditions often arise in the individual phases that influence the execution of the respective tasks. Domain knowledge of processes, data sources, and interrelationships within the domain under consideration are therefore of central importance for many project activities. The domain thus constitutes a continuous frame of reference that must be taken into account in all model phases.

Methodology

In the context of data science projects, the scientific aspect derived from the definition of data science does not necessarily imply a fully formalized or exclusively research-oriented approach, as is often the case in scientific research projects. In practical application contexts, it refers primarily to a rigorous and comprehensible methodology, which is why this term is used in the model. This includes, in particular, a structured approach, appropriate documentation, and a systematic evaluation of analysis results. The methodological effort required in each case depends on the respective project objectives, the organizational framework conditions, and the characteristics of the domain under consideration.

IT Infrastructure

The execution of data science projects depends to a considerable extent on the underlying IT infrastructure. Many project activities, such as the storage, processing, and analysis of data, require the use of suitable hardware and software components. However, the specific extent of the IT support required may vary considerably from one project to another. In many organizations, certain technical platforms or tools are already predefined. Nevertheless, both the possibilities and the limitations of the existing infrastructure should be taken into account in all phases of the project. This includes aspects such as available computing capacities, data access, and system architectures, or the possibility of extending existing infrastructures where necessary.

3.3 Iterations and Project Termination

At first glance, the presentation of DASC-PM may suggest a sequential process. This structure primarily serves to provide clarity and to describe the typical areas of activity within a data science project. In practice, however, such projects rarely proceed in a strictly linear manner. Rather, it is often necessary to revisit individual phases multiple times or to return to earlier activities.

The model should therefore be explicitly understood as an iterative approach. New insights gained in a later phase may make it necessary to adapt or repeat earlier steps. For example, the deployment and use of a developed model within a project may indicate that certain analytical methods should be further improved.

As a general principle, only those phases that are relevant to the problem statement need to be revisited in an iteration. Depending on the project context, it may therefore be sufficient to repeat individual activities on a reduced scale. It is also possible to return to various phases during the course of the project if the project team deems it necessary. In addition, a data science project can, in principle, be terminated in any phase, even if that means the objective defined in the problem statement is not fully achieved. This does not necessarily mean that the project should be considered a failure. Insights gained up to that point may still help to improve the understanding of the problem under consideration or to identify new questions.

This iterative structure allows DASC-PM to be used flexibly within various project workflows. The phases of the model describe areas of activity that can be addressed repeatedly throughout the course of a project. In this way, the model can be applied both in traditional project structures and in agile approaches involving repeated cycles of project activities.

4 Competencies and Roles

The execution of data science projects requires a wide range of different competencies. These relate not only to methodological and technical aspects of data analysis, but also to organizational, domain-specific, and communicative competencies. In public discourse, the impression is often created that these competencies are combined in the role of a data scientist. However, this view falls short, since the successful execution of data-driven projects generally requires collaboration among various disciplines and roles.

In practice, data science projects are therefore typically carried out by interdisciplinary teams. Within such teams, different competencies are brought together and jointly contribute to addressing a data-related problem.

A commonly used representation of the fundamental areas of competence in data science projects goes back to Conway (2010). In this representation, data science is described as the intersection of mathematical and statistical competencies, information technology competencies, and domain-specific competencies (see Figure 4).

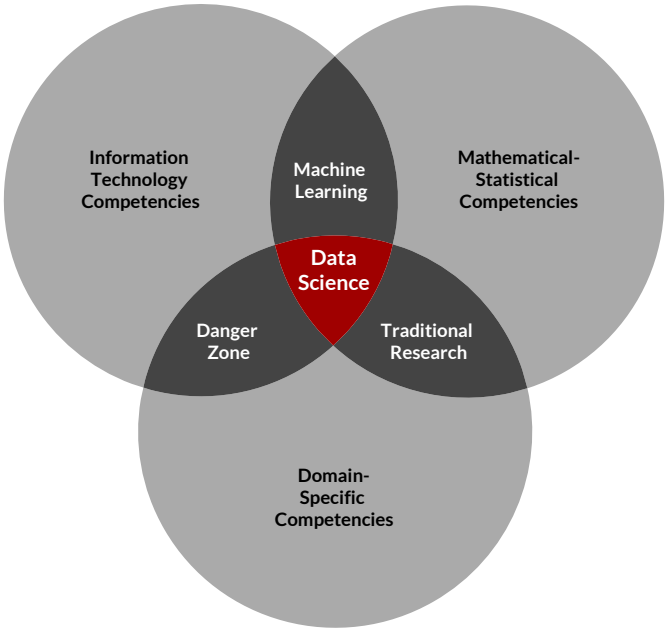


Figure 4: Competencies required of data scientists, based on Conway (2010)

Mathematical-statistical competencies include, in particular, knowledge and skills in statistics, mathematics, and modeling. These competencies form the basis for the development and application of analytical methods, as well as for assessing the informative value of analysis results.

Information technology competencies concern the processing, storage, and provision of data, as well as the technical implementation of analytical methods. These include, for example, knowledge and skills in programming languages, databases, and distributed data processing systems.

Finally, domain-specific competencies primarily describe knowledge and skills relating to the respective application context of a project. These include, for example, domain-specific processes, relevant problem contexts, and the interpretation of analysis results within the context of the respective domain.

Only through the interaction between these areas of competence is it possible to appropriately address data-driven problem contexts. In many projects, these competencies are therefore contributed by different individuals or roles.

4.1 Competencies in DASC-PM

The competencies of data scientists presented in Figure 4 provide a broad orientation with regard to the central areas that need to be taken into account in the context of data-driven projects. In practice, however, the knowledge and skills actually required are often more differentiated and vary depending on the use case. Against this background, it is crucial not only to identify all relevant areas of competence as comprehensively as possible, but also to formulate them precisely so that they can serve as a basis for the targeted composition of data science teams. In particular, this enables the targeted staffing of different roles in a complementary manner and systematically fosters collaboration within the team (Neuhaus et al., 2024; Wicht et al., 2021).

In addition, a mere list of areas of competence is often not sufficient. In order to align the specific or desired competence requirements of a project phase (or a task) with the knowledge and skills actually available within the team, a differentiation based on competence levels is required. Figure 5 illustrates this in the form of a radial diagram that combines nine areas of competence with corresponding levels of proficiency. Both the individual areas of competence and the levels are described in greater detail below.

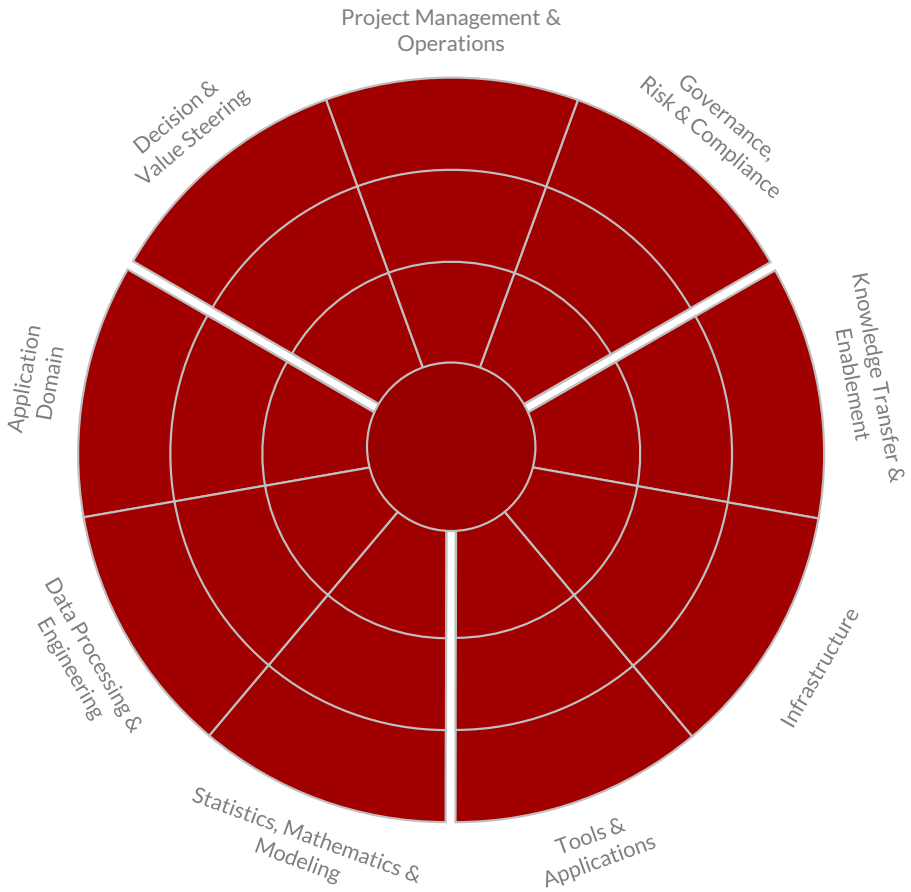


Figure 5: Competencies required in a data science project

Project Management & Operations

This area of competence describes the knowledge and skills required to plan, coordinate, and operationally implement data-driven projects in a structured manner. This includes coordinating the actors involved, guiding iterative and agile approaches, structuring work processes, and handling project artifacts. *Operations* refers to the ability to manage the transfer of the artifacts developed in the course of the project into live operation.

Decision & Value Steering

This area of competence comprises the knowledge and skills required to describe, assess, prioritize, and align data-driven initiatives with organizational objectives. This includes the identification and selection of use cases, the assessment of their value contribution under the given framework conditions, and the support of organizational change processes.

Application Domain

This area of competence focuses on understanding contexts and their specific requirements. This includes knowledge and skills such as identifying contextual relationships and business processes, as well as identifying related problem contexts and potential solutions.

Data Processing & Engineering

This area of competence describes the knowledge and skills required for the structured preparation, integration, and organization of data. This includes the design and implementation of data architectures, data management, and the assurance of an adequate level of data quality. The focus lies on transforming raw data into a usable form for analysis and modeling. The competencies of this area explicitly do not include the ability to create or adapt infrastructure from a technical perspective.

Statistics, Mathematics & Modeling

This area of competence encompasses conceptual and methodological knowledge and skills in the context of data-driven analyses. In particular, it integrates sound mathematical and statistical knowledge with proficiency in analytical methods and modern model architectures. This includes both the design and evaluation of complex analytical models and their targeted application and further development.

Tools & Applications

This area of competence describes the knowledge and skills required to implement and operate data-driven solutions and systems. It includes the selection and use of software, programming languages, frameworks, and technical stacks, as well as the integration, deployment, and monitoring of models in applications.

Infrastructure

This area of competence describes the knowledge and skills required to design and provide a suitable technical foundation for data-driven systems. This includes technical infrastructures such as cloud platforms, distributed systems, or high-performance computing systems, along with aspects such as scalability and availability. The aim is to create stable and high-performance conditions under which data processing and applications can be executed reliably and operated over the long term.

Knowledge Transfer & Enablement

This area of competence describes the knowledge and skills required to communicate knowledge, present it in a manner appropriate for the target audience, and thereby enable its use within the organization. This includes not only communicating and presenting analytical results, but also fostering collaboration among different stakeholders and empowering users to work with data and AI.

Governance, Risk & Compliance

This area of competence describes the knowledge and skills required to take legal, ethical, and regulatory requirements into account and to ensure compliance when working with data and AI. This includes topics such as information security, data protection, compliance, transparency, and risk management. The aim is to design data-driven systems responsibly, identify potential impacts, and comply with applicable requirements.

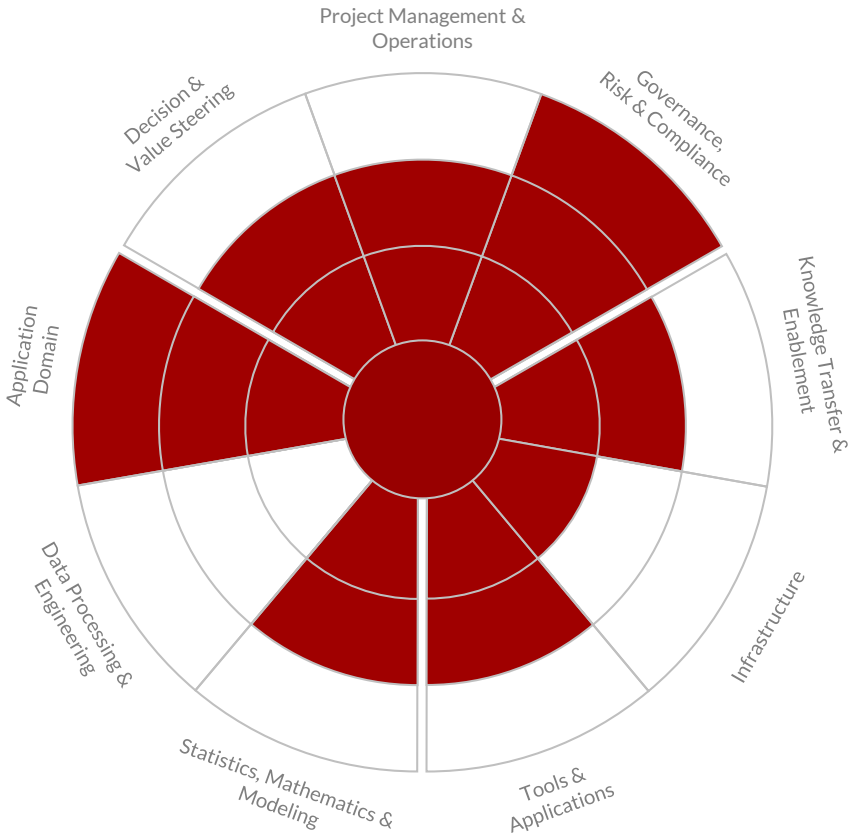


Figure 6: Competencies needed for a data science project in an exemplary profile

Figure 6 shows the competencies in an exemplary profile, as might be found based on the expertise of an individual person. A greater degree of filling from the center outward indicates a deeper level of competence, which can be described in terms of four levels. In this model, it is generally assumed that all employees involved in data science projects reach Level 1, while Level 4 describes expert status. In detail, the individual levels can be characterized as follows:

Level 1: Basic Competencies

At this level, individuals possess a basic understanding of key concepts and their interrelationships. They are able to follow clearly structured instructions and to apply routines reliably. However, they rarely make independent adaptations or assessments.

Level 2: Intermediate Competencies

Individuals at this level act with increasing independence and are able to adapt their approach to a limited extent. They are capable of assessing results and insights and interpreting them within the respective application context.

Level 3: Advanced Competencies

At this level, individuals are able to assess more complex situations independently. In doing so, they take relevant contextual factors, conflicting objectives, and uncertainties into account and make well-founded decisions. In addition, they are able to evaluate results and insights in a manner appropriate to the situation.

Level 4: Expert-Level Competencies

Individuals at the expert level are able to critically question established approaches, independently develop new solutions to new problems, and reflect critically on the added value and limitations of these solutions. They are able to advise others on a well-founded basis and actively contribute to the further development of best practices and standards.

4.2 Roles in DASC-PM

The successful implementation of data science projects requires close collaboration among different roles that collectively address technical and domain-specific challenges on the one hand, and legal, strategic, and process-related challenges on the other. Many representations work with fixed role titles, which, however, may vary considerably from one organization to another and may also change over time. DASC-PM therefore does not describe specific role names, but rather three overarching role areas (see Figure 7).

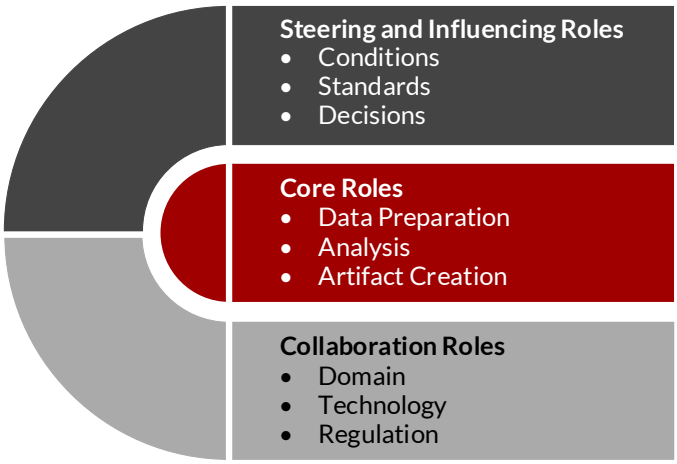


Figure 7: Role areas and selected responsibilities within a data science project

As a rule, it is not possible for a single individual to develop extensive competencies in all of the areas mentioned (Zschech et al., 2018). Data scientists may therefore either specialize in selected disciplines or take on more overarching and less data-oriented roles.

Core Roles

Core roles include functions that are responsible for the analytical and data-related implementation of a data science project and for developing its central artifacts. In practice, the term *data scientist* is used in different ways: in some cases, as an umbrella term for all persons involved in a data science project, and in others in a narrower sense for those who specialize in the actual analysis of data. In this document, the term is used in this narrower sense:

> ***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 larger and more complex data science projects, the role of the data scientist may be divided into additional sub-roles. These include, among others, data analysts, who are concerned with data preparation, analysis, and evaluation and are particularly active in exploration and forecasting (often using traditional analytical methods). In contrast, methodology specialists focus on the research and further development of data science methods and are more oriented toward theory and science. Machine learning engineers represent a specialization closer to software and infrastructure and are responsible for the technical implementation as well as the productive and scalable deployment of analytical artifacts.

An equally central function within the core team is performed by data engineers, who are responsible for the procurement, storage, preparation, structuring, and provision of data. They are particularly active in the stages preceding the actual analysis. Compared with data scientists, they have a stronger technical focus and also deal with the IT infrastructure required for the data science project. The term *data architect* is also occasionally used for this role.

A sub-role of the data engineer, which is often assigned separately (particularly in larger data science projects), is that of the *data steward* (also referred to as a *data manager* or *data quality engineer*). This role is continuously concerned with access to the data and their protection, as well as with ensuring a consistently high level of data quality over time. A data steward therefore has strong points of contact with the domain-specific area of application.

Collaboration Roles

Collaboration roles comprise functions that operate at the interfaces between the domain, technology, and regulatory requirements, thereby ensuring the integration and applicability of the solutions. A central function within this group is performed by *domain experts*. These are often the domain users themselves or their representatives. They possess specific knowledge of the application domain and have a substantive understanding of the problem context or use case. Their deep domain knowledge enables them to define the domain-relevant priorities for analysis and modeling.

Within the group of domain experts, there may in turn be further sub-roles. In business contexts, these often include *business developers*, who develop the domain-specific use case of a project and thus form the link between business objectives and data analyses, and *business analysts*, who later use the developed analytical models as part of their domain-specific tasks.

In addition to these domain-related interfaces, various organizational and technical functions also play a role. *Governance, risk, and compliance experts* support the project by interpreting regulatory requirements, accompanying their implementation, and promoting awareness of responsible use.

Platform and infrastructure functions comprise all tasks required to establish the technical prerequisites for a data science project. Typical roles in this area include, for example, the *IT infrastructure architect*, who is responsible for designing a suitable IT infrastructure for the project, and *IT technicians* or *IT administrators*, who provide the necessary hardware and software and configure the underlying systems. *Application developers*, who are concerned with implementing application software and tools for the productive use of analysis results, are also assigned to this area of responsibility.

Solution engineers and *solution architects* are responsible for ensuring that the developed solutions can be integrated into existing systems and embedded in a viable overall architecture. *Platform engineers* provide cloud or data platforms on which data processing, training, and model operation can take place. The role of the product owner ensures that requirements are prioritized and that project results are aligned with the expected benefits. *Data owners* make decisions regarding data quality, access rights, and the domain-specific interpretation of central data objects. *UI/UX designers* help ensure that analysis artifacts are embedded in user-friendly, comprehensible systems that can be used effectively in everyday work.

While other roles primarily implement or integrate analysis results, the *MLOps engineer* ensures that the analytical artifacts can be operated reliably, efficiently, and in a controlled manner over the long term. This includes establishing stable operating environments, continuously monitoring utility, costs, and performance, and automating key processes such as deployment, retraining, and monitoring.

Steering and Influencing Roles

Roles with steering and influencing capabilities include positions that significantly shape the course of a data science project through decisions, requirements, or the design of legal, security-related, and organizational conditions. *Project managers* plan, direct, and coordinate the overall course of a data science project. In addition to traditional project management skills, they require a sound understanding of the methodological and technical aspects of data science, knowledge of suitable process models, and insight into the application domain.

Particularly in smaller projects, project management is often carried out by individuals who also serve as data scientists or data engineers. However, project management may also be performed by individuals without specific data science expertise, provided that appropriate experts support them. Such experts, also referred to as *methodology leads* or *technical leads*, possess advanced methodological and technical expertise and support the

domain-related and technical orientation of the project. Together with domain experts, they define the scope of the analysis and implementation.

At the strategic level, *managers* specializing in data science and AI take on responsibilities for long-term alignment, identify potential, assess risks, and develop a data science strategy that connects the project environment with the organization's objectives. *Sponsors* and *decision-makers* influence the project through budget responsibility, approvals, and the setting of key priorities. *Security*, *privacy*, *and legal specialists* shape the legal, security-related, and ethical conditions within which data science projects may be carried out. They ensure data protection, information security, and compliance with legal requirements and thus act not only as a control function, but also as a shaping force within the project.

Part B
DASC-PM in Detail

Notes on Part B

The following chapters examine the individual phases of DASC-PM in detail. To structure the individual phases, a uniform nomenclature is used that focuses on four central terms: *core tasks*, *accompanying tasks*, *feature-bearing areas*, and *interface tasks*.

Core tasks comprise those activities that contribute directly to achieving the objectives of the phase. Their execution is essential to the success of the project. They therefore form the substantive focus of each phase.

Accompanying tasks support the execution of the core tasks. They describe activities such as quality assurance, risk management, or preparatory activities without which the core tasks cannot fully achieve their objectives or cannot do so with the desired quality.

Feature-bearing areas describe the points of reference for the tasks within the respective phase. They are often described as objects, artifacts, or results, but may also refer to more abstract elements such as events. The tasks of the phase draw on them, are influenced by them, or produce them as outcomes.

Core tasks, accompanying tasks, and feature-bearing areas are grouped together under the term *areas*. They are presented in the chapter descriptions of this part of the document and discussed in greater depth. The extent of the discussion varies from one area to another, just as the areas in most projects do not require exactly the same degree of detail.

Interface tasks relate to the linkage of the elements described above. They usually arise implicitly and describe, for example, the handover of results, coordination between the individuals carrying out the work, or the integration of different elements.

Each chapter begins with both a simplified and a detailed presentation of the areas and interface tasks of the phase described. The simplified presentation serves to provide an overview, while the detailed presentation shows the individual interface tasks and possible sequences. Figure 8 illustrates the nomenclature used in an exemplary notation for the detailed presentation that is employed in the individual chapters to show the interrelationships.

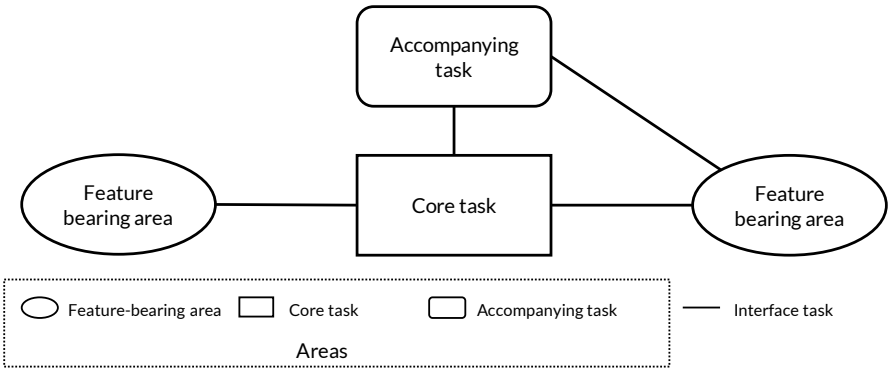


Figure 8: Nomenclature used and notation in the phases

The introductory overview for each phase is complemented by a presentation of the competencies required. This presentation is based on the discussion in Section 4.1 and, in particular, on Figure 6. It represents an aggregation of the assessments provided by the working group and is intended to enable an approximate assessment of the competence profile for a typical data science project. However, the actual profiles may differ considerably depending on the individual project and its context. Especially when considering competencies at the team level, it is also conceivable to mark not only fully attained levels, but also levels that have been only partially attained by team members.

Part B concludes with a consideration of the overarching aspects of Domain, Methodology, and IT Infrastructure.

5 Problem Statement

Problems arising within a domain trigger use case development. The most promising use cases are then elaborated into a data science project outline. All related tasks are reflected in the *Problem Statement* phase. The competencies typically required in this phase focus primarily on domain expertise and overarching project steering capabilities.

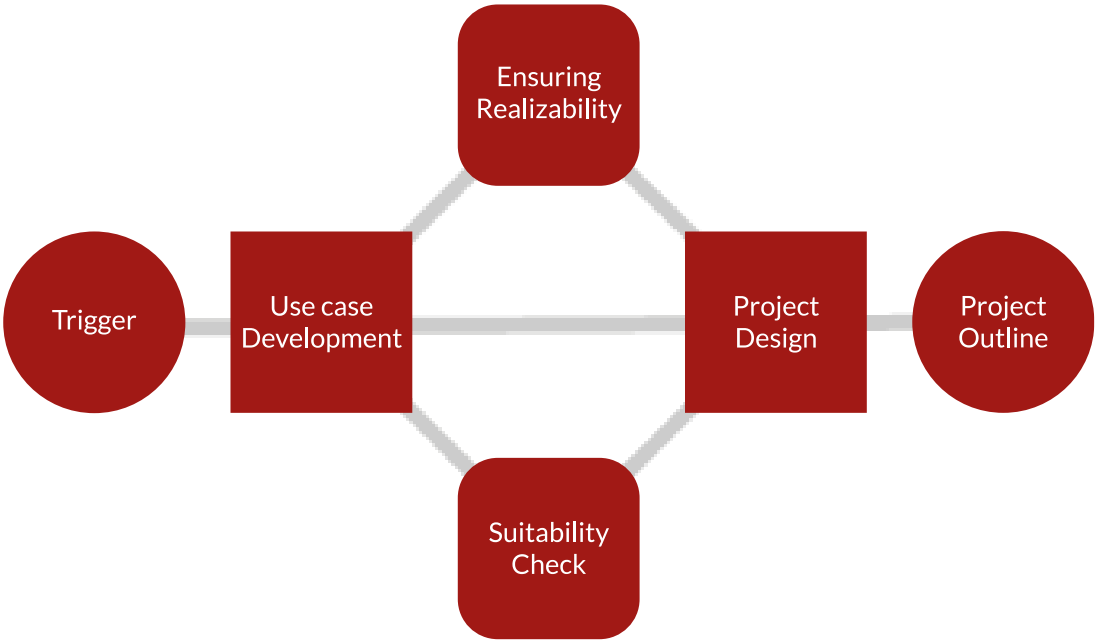


Figure 9: Brief overview of the "Problem Statement" phase

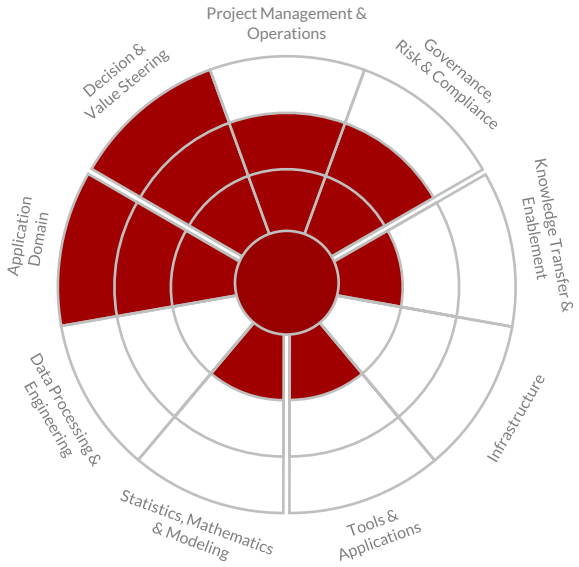


Figure 10: Competence profile for the "Problem Statement" phase

Detailed Presentation of the "Problem Statement" Phase

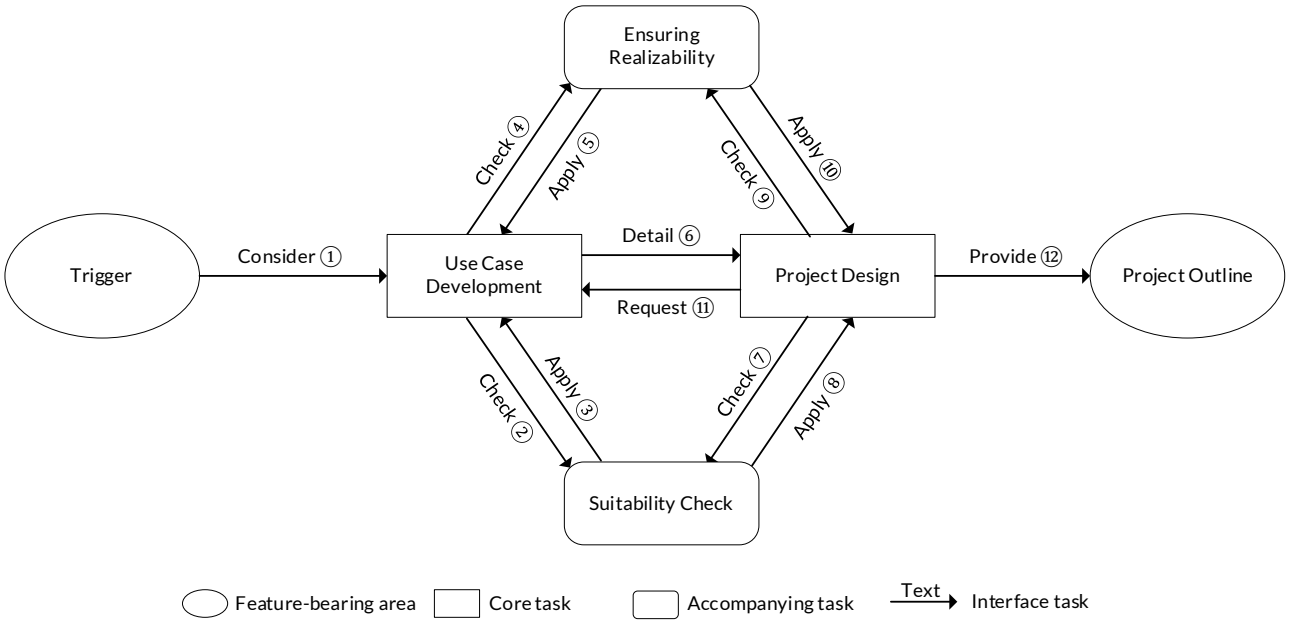


Figure 11: Detailed presentation of the "Problem Statement" phase

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 feasibility 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 designed 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 feasibility 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 design 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.

5.1 Feature-Bearing Area "Trigger"

The initialization of a data science project is triggered by an event—usually a problem—within the domain. In a research-oriented or exploratory context, however, it may also involve open questions that do not, in the usual sense of the term, constitute a "problem" as such.

As a rule, it is only in the subsequent steps that the triggering problems can be sharpened and specified into one or more use cases. For this reason, no particular formal requirements are imposed at this stage with regard to wording or forms of documentation. Nor does a specific level of abstraction need to be prescribed for the descriptions, since triggers may differ considerably from one another.

Characteristic	Description
Objective	A decision must be made as to how the results of a data science project triggered by a problem are to be used at a later stage, for example whether the project objective is to generate insights or to operate models in production.
Functional purpose	Defining the domain-specific purpose of the solution to be developed can help establish the scope of the project. It also makes it possible to determine the relevance of a given solution to the problem.
Requirements	A description must be provided of the requirements that the solutions to be developed need to fulfill.
Participating areas	The areas on the domain side that are involved in carrying out the project must be specified.
Functional domain	A description must be provided of the application domain within which the problems are to be addressed.
Application area	The level of abstraction must be defined. For example, is the solution to be developed intended merely to provide recommendations for action for a single department, or is it concerned with strategic decisions for an entire corporation?
Complexity	An appropriate classification is possible only once the complexity of the problems under consideration has been assessed.
Alternate courses of action	This concerns both alternative ways of carrying out the data science project and alternatives to carrying it out.

5.2 Core Task "Use Case Development"

A frequently cited challenge when organizations engage with the discipline of data science is the identification and selection of feasible use cases. For the purposes of DASC-PM, the following applies:

---

*A **use case** is a clearly defined, self-contained problem or sub-problem within a domain under consideration, with a formulated objective and a clearly identifiable domain-specific benefit, the solution of which may require the completion of several tasks.*
*A **project** may address one or more use cases.*

There is no clear answer to the question of which groups of people or departments within an organization should drive the development of problems into use cases. Domain experts and data scientists, as well as other relevant groups, can and should be involved.

Subtask	Description
Develop an understanding of the discipline	Expectations of the discipline of data science often do not align with its actual capabilities. In addition, data science initiatives frequently lack a clear focus. A sound understanding of the discipline therefore needs to be developed.
Use case identification	On the one hand, breaking down organizational objectives into use cases that can be addressed within a project may be challenging. On the other hand, use cases must arise from requirements and problems within an organization. It often remains unclear what results can be expected from implementing the respective use cases. The task is therefore to develop ideas that are both creative and feasible.
Prioritizing use cases	In some cases, selecting suitable use cases for implementation is not possible, as the potential for the organization may not be directly apparent before the project is carried out. Input data and interim metrics for evaluating alternative use cases—for example on the basis of returns or return on capital—may be unavailable or only imprecise. It is therefore possible that efforts to prioritize use cases will not prove worthwhile, as they may later turn out to have been misguided.
Coordinating groups of participants	The groups within an organization that could benefit from the implementation of use cases in the form of projects are often unaware of the value that data science can provide for them. Data specialists, by contrast, may not recognize the most relevant use cases for the organization. Communication and coordination between these two groups are therefore of great importance.

Problem Statement

When selecting suitable use cases, it is important to establish a basis of trust among the groups involved. Domain experts should be able to speak openly about difficult, resource-intensive, or challenging tasks without being influenced by premature solution proposals. Preparatory bilateral discussions are advisable in order to sound out the framework conditions, build good relationships with the relevant contacts, and prepare interviews or workshops. The methods chosen must be adapted to the specific organizational context.

Frequently mentioned formats for identifying suitable use cases are interviews and workshops.

Interviews are more suitable for smaller areas with fewer participants. They should be conducted by experienced interviewers as conversations among equals.

Workshops are also suitable for larger groups and areas. Workshops with many participants are used particularly for generating ideas. The further specification and prioritization of use cases can then be carried out in smaller workshops. In order to illuminate as many facets as possible, the group of participants should be as heterogeneous as possible, covering different areas (for example domain experts, data scientists, and decision-makers) as well as different levels of seniority. Possible items on the agenda of such a workshop include the presentation of exemplary use cases, joint brainstorming, and the evaluation of proposals in terms of relevance and feasibility. Where appropriate, selected data science theory may also be used to provide a conceptual foundation. The outcome of the workshops should be the selection and elaboration of a use case that is as concrete as possible. It should be clear what is to be achieved by addressing the use case within a project and what business value or impact may result from it.

Various methods can be used to structure workshops. In general, the methods employed should provide the workshop with a basic structure and a common thread, while at the same time leaving room for flexible and creative elements. Such a basic structure can, for example, help bring workshop participants to a similar level of understanding or enable the questions under discussion to be examined on the basis of a shared scientific foundation. The creative elements can increase motivation, stimulate different ways of thinking and approaches, and foster group interaction, even among participants with very different backgrounds.

General methods such as focus groups, fishbowls, design thinking, and hackathons have proven effective for identifying and selecting suitable data science use cases. In addition, there are also more specialized methods tailored to the context of data science, artificial intelligence, machine learning, or big data, such as the Enterprise AI Canvas (Kerzel, 2021), the Machine Learning Canvas (Dorard, 2015), or the approach described by Bill Schmarzo (2015). It is important to ensure that the methods selected are appropriate both to the problem context and to the group of participants involved (the key consideration here is acceptance, especially in the case of more “exotic” 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).
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 following best practices may be used 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 a concrete business benefit with relatively little effort and a realistic prospect of success. Quick wins often help to convince initial skeptics and secure resources for subsequent, more demanding projects.

Lighthouse projects are showcase projects developed with considerable effort and commitment. They are intended to demonstrate what can be achieved through a well-planned and well-executed data science project. Lighthouse projects should achieve a high level of visibility, convince initial skeptics, for example through testimonials, and thereby encourage other departments or organizational units to follow suit. Ideally, parts of the lighthouse project can be adapted and reused for subsequent projects.

Data science use cases for which technical feasibility is uncertain may begin with a *proof of concept*. In such cases, an attempt is made, using a predefined and generally limited amount of resources (for example time and personnel), to demonstrate that the desired objective can in principle be achieved under the given conditions (existing data, available IT infrastructure, and so forth).

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 aim of the *suitability check* is to determine whether the identified use cases, and those subsequently selected, can be implemented successfully within a project. To this end, it must be examined whether the defined requirements can in principle be met using the available resources. The actual assessment of whether the necessary technical, domain-specific, and organizational prerequisites are in place is carried out as part of ensuring feasibility (see Section 5.4).

If the use case under consideration is selected for implementation in a project, a more detailed assessment is carried out as part of *project design* (see Section 5.5). At this stage, use cases may also be prioritized.

Subtask	Description
Suitability of the use case	A check must be performed to determine whether the use case under consideration is in fact one for which the use of data science appears appropriate.
Suitability of methods	A check must be performed to determine whether analytical methods exist, or can be developed, that are likely to produce a suitable result with an adequate degree of probability. If necessary, initial tests must be carried out for this purpose.
Assessing the data basis	It is often unclear which data are available or can be procured for data science projects, what quality they have, and to what extent they are suitable for use in analyses. The effort required for data preparation and the actual value of the data can also often only be assessed with difficulty in advance. An initial assessment of the data foundation at this early stage is essential.
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.

5.4 Accompanying Task "Ensuring Realizability"

At this stage, it must be examined which project ideas can in fact be implemented. This often involves an iterative process with all stakeholder groups. In some cases, the implementation of use cases is characterized by a high degree of uncertainty as to whether the investigations will yield insights and what form those insights may take.

While the *suitability check* (Section 5.3) assesses whether a use case is, in principle, suitable for data science, ensuring feasibility focuses on the specific technical, organizational, and domain-specific conditions under which it can actually be implemented.

Organizations often possess a high proportion of tacit knowledge, and at the beginning of a project it is often unclear how this knowledge can be represented in analysis artifacts. In addition, legal aspects may restrict possible use cases. In some cases, the responsible contacts are not sufficiently familiar with these issues and may therefore be overly cautious. The effort required to implement use cases, as well as the resources needed to carry out the data science project successfully, is also frequently underestimated.

If the use case under consideration is selected for implementation in a project, a more detailed assessment is carried out as part of *project design* (see Section 5.5).

Subtask	Description
Checking the IT infrastructure	A check must be performed to determine whether the available IT infrastructure is suitable for implementing the use case under consideration. Alternatively, it must be examined whether other technical options exist and whether additional infrastructure can be procured if necessary.
Evaluating the expertise	The suitability of the expertise of the individuals involved for implementing the use case under consideration must be assessed.
Risk assessment	The risk involved in implementing the use case as part of a project must be assessed, including both the likelihood of occurrence and the severity of the consequences.
Cost-benefit analysis	Although an analysis of the expected benefits is often very difficult to carry out, the costs should nevertheless be assessed as a matter of principle.

5.5 Core Task “Project Design”

The aim of *project design* is to determine the necessary work steps that lead to fulfilling the requirements specified by the respective use case. This must be done on the basis of the available information about the data basis and while taking the specific characteristics of the domain into account. Since the fundamental characteristics of project management differ only marginally from those of other types of projects, reference is made here to the relevant standard literature.

Data science projects do, however, also exhibit highly specific project characteristics. Since their success is often more difficult to assess in advance than that of projects in other areas, they require particularly careful consideration. In some cases, this requires a distinction to be made between exploratory research and development projects and those projects that are specifically aimed at implementation or regular operation.

5.6 Feature-Bearing Area “Project Outline”

Unlike in many other disciplines, it is generally not possible to fully plan and describe the course of data science projects in advance. As a result of this subprocess, only a *project outline* can be produced, which must be further elaborated over the course of the project. In particular, in agile process models such as Scrum or Kanban, this may initially amount to the creation of a backlog or a comparable collection of intended functionalities and structures for the desired solutions. Accordingly, the term project outline as used here should be transferred to, or adapted for, the respective approach.

In any case, it is important that the project be described at a level of abstraction that allows all relevant requirements and information from the perspectives of data, the domain, and analysis to be presented concisely.

In addition, the project description should identify the work steps and sequences that can already be determined at this stage and that lead to the fulfillment of the defined requirements. Should any changes arise during project execution, the project outline must be adapted accordingly.

6 Data Provision

The *Data Provision* phase comprises data preparation (from acquisition to storage), data management, and exploratory data analysis. The outcome of this phase is a data source that is suitable for further analysis.

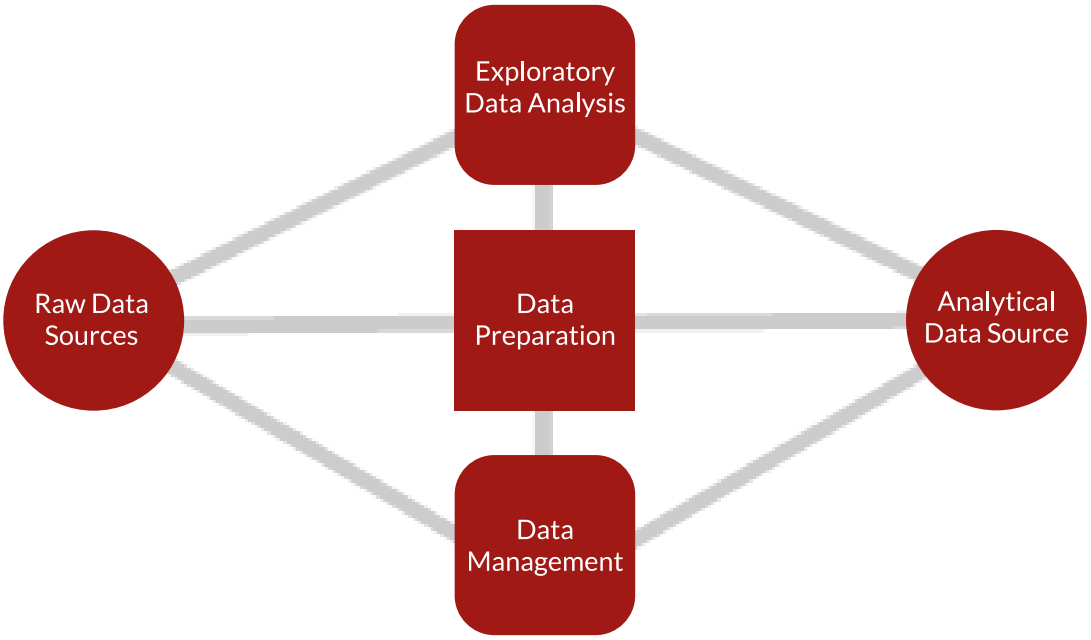


Figure 12: Brief overview of the "Data Provision" phase

Figure 13: Competence profile for the "Data Provision" phase

Detailed Presentation of the "Data Provision" Phase

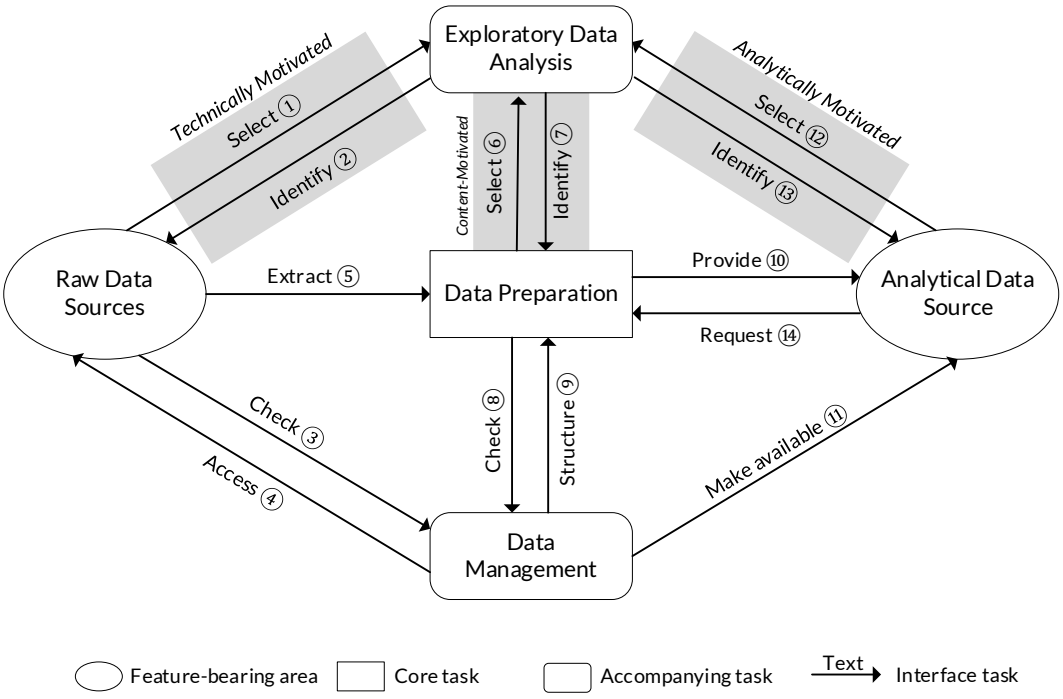


Figure 14: Detailed presentation of the "Data Provision" phase

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 defined 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 exploratory 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 is adjusted.

6.1 Feature-Bearing Area "Raw Data Sources"

Data are usually not collected for analytical purposes. If existing data sources are used, an understanding must first be developed of the processes, recording procedures, and conditions under which these sources came into being. This metadata must be documented in an appropriate form and assigned to the data set. Ideally, metadata from multiple data sources should be managed in a metadata repository in order to enable the sustainable use of these data sources in other projects as well.

Characteristics of data (sources) can be divided into four categories that may be relevant for a data science project. The purpose of this presentation is not to provide a comprehensive checklist of all conceivable characteristics, but rather to facilitate a structured approach to a basic inventory.

Characteristic category	Description
Procurement effort	The availability of data can have a major impact on which analyses are carried out. For example, if data are already available within the organization and can be loaded automatically, this involves significantly less effort than the use of external data, which must first be collected, purchased, or located.
Administrative effort	Depending on the volume, rate of change, and confidentiality of the data, different forms of data storage may be required. Another relevant characteristic is whether the data need to be accessed only once or repeatedly.
Processing effort	The way in which data must be transformed in order to make them usable for analysis is influenced, among other things, by its granularity, redundancy, and structure, as well as by any preprocessing already carried out in the source systems.
Data quality	The quality of the data depends, among other things, on their timeliness, the proportion of missing or erroneous values, and their relevance to the data science project. In order to assess data quality, it is necessary not only to understand their origin and the process by which they were collected, but also to conduct an exploratory data analysis before applying complex analytical methods.

A more detailed discussion of the possible characteristics of data sources can be found, for example, in Jayawardene et al. (2013). Despite its breadth, the list of data quality criteria does not claim to be exhaustive. The relevance of the individual characteristics must be assessed on a project-by-project basis.

Data quality criteria (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/Redun-dancy	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"

The aim of *data preparation* is to transform the available raw data into a consistent, structured form of adequate quality so that they can be used efficiently and reliably in subsequent analytical methods. A key sub-objective of this task is therefore to improve data quality.

The processing of large volumes of data requires the use of high-performance hardware and software, and in some cases also innovative methods.

Possible artifacts of data preparation include scripts that may also be automated, but in any case document the process and make it repeatable. The result of executing these scripts is a prepared data basis that is suitable for the data science project and takes the tasks outlined above into account.

Documentation of the preparation steps is just as necessary as documentation of the characteristics in a data catalog.

Subtask	Description
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 anonymization	If confidential data (such as personal data) are needed in data science projects, they must first be anonymized or pseudonymized as applicable.
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.
Dimensional reduction	Irrelevant or redundant material should be removed from the database.
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 includes both the need for transformation identified during the exploratory data analysis and transformations from a more technical viewpoint driven by data management.
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.
Generating characteristics	Additional or alternative characteristics can be derived from existing data.
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"

In *data management*, the focus is on making the required data available without, at this stage, formulating requirements for an IT infrastructure.

Unlike the other areas of data provision, which focus on the identification, assessment, and preparation of specific data sets, data management comprises the overarching organization, governance, and safeguarding of the sustainable use of data throughout the entire course of the project.

As an artifact, the data catalog is extended in order to ensure the traceability of data management.

Subtask	Description
Data archiving	If analytical methods are to be reproducible and this cannot be ensured by the source systems, the data used must be archived. In addition to technical challenges, issues such as copyright must also be taken into account, as they may make permanent storage impossible.
Data protection	Depending on the data used, and in particular on their sensitivity and whether they relate to identifiable individuals, data protection requirements must be examined and complied with. These include, for example, obligations regarding anonymization and pseudonymization. Role and permission concepts must be designed in compliance with data protection requirements.
Data security	The aspect of confidentiality in particular must be addressed through appropriate technical and organizational measures. Depending on the required level of protection for the data, this may include, for example, the implementation of access controls, encryption, logging, and recovery procedures in an appropriate form.
Backing up prepared data	A check must be performed to determine whether the prepared data need to be backed up during the execution of the data science project or whether they can be restored automatically by means of the scripts developed.
Storing raw data	It must be examined whether the source data are to be backed up separately for the project. If data volumes increase over the course of the project or new data are continuously added, appropriate processes and infrastructures must be put in place.
Data access	Data can be loaded—and processed—either as a one-time transfer, at defined intervals by means of batch processing, or as a stream in (near) real time. In the context of open science, access to the data may also be granted to third parties where appropriate.
Meta data management	Metadata extracted from the sources, or supplemented or identified in the course of the tasks carried out, should be managed appropriately.

6.4 Accompanying Task "Exploratory Data Analysis"

The aim of *exploratory data analysis* is to develop both a structural and a substantive understanding of the data, identify anomalies, and derive hypotheses for further analysis. Unlike the Analysis phase, the focus here is not on the application and evaluation of specific analytical methods, but on the preparatory examination of the data as a basis for their targeted use.

It should also be clarified whether the quantity and quality of the available data are sufficient for the chosen problem context and whether the planned analysis requires any additional data preparation steps.

Since the purpose of exploratory data analysis is precisely to uncover aspects that are not yet known, there is no fixed sequence of methods to be applied. In addition to data visualization, various statistical methods are usually employed, such as correlation, factor, and cluster analyses, as well as statistical and, where appropriate, causal modeling. The resulting visualizations and models therefore constitute the artifacts.

Identified problems relating to data quality, as well as any necessary modifications to the data material, must be documented. However, since exploratory data analysis typically involves examining—and in some cases discarding—a large number of hypotheses in rapid succession, the documentation usually does not go into a high level of detail.

Subtask	Description
Identifying outliers	Outliers can strongly influence subsequent analysis results. A decision must be made as to whether the identified outliers correspond to actual data points or have arisen as a result of other effects. Depending on this assessment, these values may need to be filtered out or replaced.
Data validation	Drawing on domain knowledge, values can be identified in data sets that may be formally correct but are substantively incorrect or not plausible.
Data visualization	Simple charts (for example histograms, line charts, or scatter plots) can be used to illustrate the distribution of the available data and to reveal simple relationships between attributes.
Identifying central attributes	Subsequent data analysis can be carried out more efficiently if the data sets contain fewer attributes. The aim is therefore to identify those attributes that are as central and informative as possible and to exclude those that are irrelevant. Domain knowledge and statistical knowledge are often drawn upon for this purpose.
Understanding content	The data must be assessed with regard to their suitability for the specific domain and in light of the objectives of the current data science project.
Statistical analysis	Simple statistical measures such as the median, mean, standard deviation, or correlation help to quickly develop a better understanding of the available data and to detect unexpected deviations.
Examining the necessity of data transformations	To ensure the comparability of attributes, it is often necessary to normalize the data. Another reason for transformations lies in the analytical methods to be applied later, which often require a specific data format. Identifying the necessary transformation tasks is part of exploratory data analysis, while their implementation belongs to the area of data preparation.
Examining missing values	If attribute values are missing from data sets, a decision must be made as to whether the affected data sets or attributes can be removed. Since this may affect the volume, representativeness, and informative value of the underlying data, it may also be appropriate to replace the missing values. Identifying suitable methods for handling missing values is part of exploratory data analysis, while the implementation of corresponding measures belongs to the area of data preparation.

6.5 Feature-Bearing Area “Analytical Data Source”

Raw data sources and analytical data sources correspond in their essential characteristics. However, the preparation of data for data science applications gives rise to differences in content, scope, structure, and format at a more detailed level.

For example, with regard to the analytical objective, the aim is for the attributes to be cleansed and as free from redundancy as possible. In addition, the attributes should be of particular relevance to the analytical objective. However, especially at an early stage of a data science project, it is not always possible to assess relevance unambiguously. An assessment by domain experts is therefore advisable.

Depending on the analytical methods to be applied, the data formats and scale levels must be adapted. Many learning methods, for example, process only numerical attributes.

Analytical data sources can generally be worked with independently by the groups involved in the project. Access to the data may take place in real time, continuously, or on a one-time basis. In addition, metadata relating to the data sources may be made available.

7 Analysis

In a data science project, existing methods may either be applied or new methods may first need to be developed—the decision between these alternatives constitutes a challenge in its own right. The Analysis phase therefore encompasses not only the execution of the analysis itself, but also related activities. The artifact of this phase is an analysis result that has undergone both methodological and domain-specific evaluation.

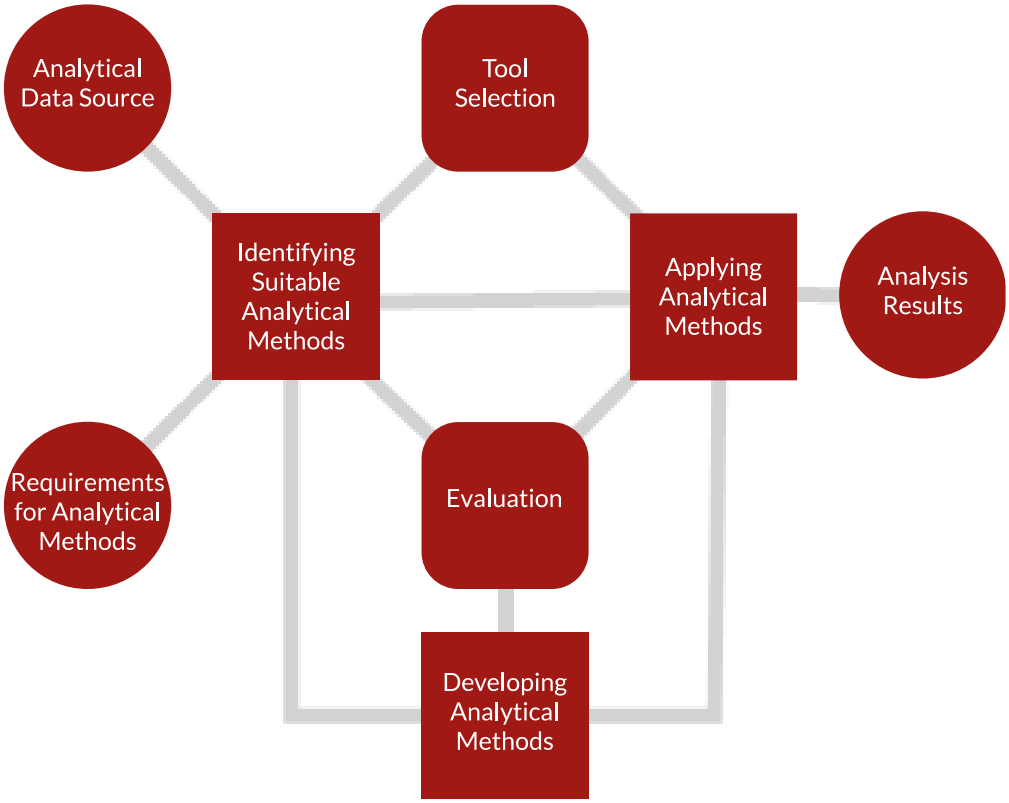


Figure 15: Brief overview of the "Analysis" phase

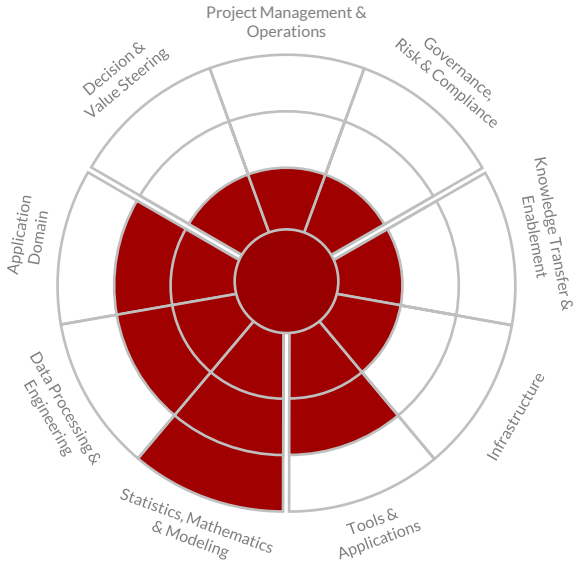


Figure 16: Competence profile for the "Analysis" phase

Detailed Presentation of the "Analysis" Phase

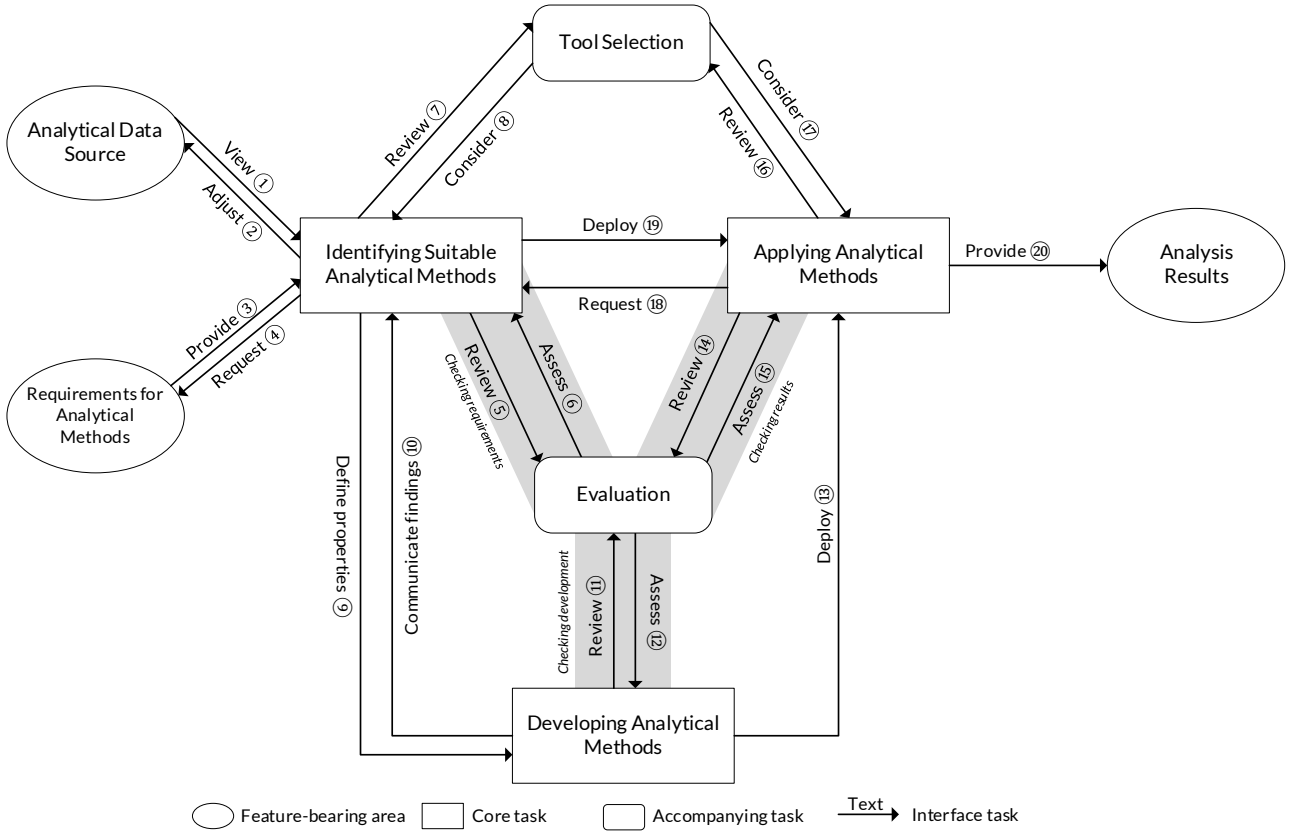


Figure 17: Detailed presentation of the "Analysis" phase

①	The analytical data source is created by handling the tasks described in the Data Provision phase. 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	After potentially suitable analytical methods are identified, the analytical data source might need to be adjusted to ensure applicability.
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 must be considered during the (further) development of the analytical methods.
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 deployment.

7.1 Feature-Bearing Area "Analytical Data Source"

The identification of suitable analytical methods is based on the characteristics of the available analytical data source (see Section 6.5).

The analytical question under consideration, or the required analysis result, often gives rise to specific requirements for the data source. Conversely, unchangeable characteristics of the analytical data source may also limit the range of questions that can be answered.

Within the Analysis phase, no additional specific characteristics need to be identified beyond those already described.

7.2 Feature-Bearing Area "Requirements for Analytical Methods"

The characteristics considered in this section constitute the non-functional requirements for analytical methods. In an individual project, they may already be defined with explicit threshold values and used as specification requirements.

They thus form a central basis for the systematic selection, evaluation, and, where appropriate, further development of suitable analytical methods over the course of the project.

Characteristic	Description
Requirement coverage	The selected analytical methods will not always be able to fully meet all application requirements. Nevertheless, the aim should be to achieve the highest possible degree of coverage.
Efficiency	The method must be applicable within an appropriate time frame using the available IT infrastructure. The less data and computing time it requires, the more easily it can be integrated into the organization's ongoing operations and the more economically it can be applied.
Innovative solution	The method must solve a problem that existing methods do not yet address to the same extent or with the same level of quality.
Replicability	In order for the result to be reproducible by others, and ideally for the method used to be applicable across different scenarios, technologies and algorithms must be employed that are thoroughly documented and widely available.
Robustness	The methods used should be as robust as possible to errors. For example, it is helpful if erroneous data or outliers are detected automatically or have only a limited effect on the result.
Scalability	In practice, the volume and/or dimensionality of newly analyzed data often increases considerably over time. It is therefore advantageous if the chosen method is also capable of processing a growing volume of data with a reasonable amount of additional effort.
Implementability	The method must be implementable using the available resources (e.g. technical infrastructure and qualified personnel). In addition, it should require as little implementation effort as possible.
Validity	The predictions or derived structures should reflect the reality of the problem context as accurately and reliably as possible. The acceptable margin of error depends on the specific problem context.
Comprehensibility	Where possible, the results of the methods should be comprehensible and easy to communicate and/or visualize.

7.3 Core Task "Identifying Suitable Analytical Methods"

Before this task begins, it should already have been clearly established that the problem at hand can in fact be addressed by means of data science—that is, that it represents a problem which is potentially solvable, while at the same time not being so trivial that it could, for example, be solved by means of a standard report.

Identifying suitable analytical methods often presents a major challenge. Although a very large number of analytical methods exist, it is possible that none of them is suitable for the problem at hand. In that case, it must be examined whether selected project framework conditions can be modified, whether the development of a new analytical method is conceivable, or whether the project must, if necessary, be terminated.

Accordingly, the focus of this task lies on gaining an overview of existing methods and identifying the most suitable methods for application. Since no final selection can be made without further evaluation, several methods may initially be taken forward for subsequent assessment. The decision to develop a new method should be made with due regard to the effort involved and the uncertainty that remains.

An artifact of this task is a list of analytical methods that also includes the reasons why each method is suitable for the problem context. If no suitable analytical methods can be identified, methods may be selected for further development. Where appropriate, a prototype may even already be created in order to verify the suitability of the selected methods.

The insights gained from this task should be documented in such a way that they not only justify the selection made for the current project, but also record the decisions in a form that can be applied to future problem contexts.

Subtask	Description
Identifying requirements	Before different methods are assessed, clarity must be established as to which problems they are intended to solve.
Determining the problem class	Based on the identified requirements, the problem can usually be assigned 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 is helpful to investigate whether there are publications on similar use cases.
Determining potentially suitable procedures	Against the background of the problem class and on the basis of research into comparable problem contexts, analytical methods or variants of analytical methods that appear promising in principle can now be identified.
Selection	Once the candidate methods have been identified, those that best match the project-specific criteria and available resources should be selected.

7.4 Core Task "Applying Analytical Methods"

The correct *application of analytical methods* requires detailed knowledge of existing methods. If methods are applied incorrectly, this leads to arbitrary results and, consequently, to inaccurate or false conclusions.

It must be ensured that the methods to be applied fulfill the respective tasks in an appropriate manner. This already plays a central role in the identification stage (see the previous section), but can only be conclusively verified in the actual application to the data to be analyzed. The aim is to achieve the best possible analysis result. In detail, this depends on the method selected as well as on the specific requirements of the respective domain. In the case of some methods, for example, a trade-off must be made between striving for the highest possible level of accuracy and developing a model that can be applied across as many scenarios as possible.

A large proportion of the artifacts produced and the documentation required depends on the individual project and is therefore inseparably linked to the problem context, the data used, and the analytical methods applied. In general, the following artifacts are produced:

- documentation of the execution of the analysis and of the evaluation results (including intermediate results and figures),
- a justification for the selection of the final model,
- a backup of the development environment,
- the trained models,
- interface documentation, and
- the parameter configurations.

Domain-related information should be prepared in a way that is easy for domain experts to understand, including indications of any errors or anomalies that occurred and of further problem contexts that could be investigated using the analytical methods. Depending on the tools used, basic documentation is already generated during the analysis process itself.

Subtask	Description
Setting up a development environment	Especially where multiple users are involved, there should be a powerful and easily accessible development environment with version control in order to ensure the smooth long-term execution of the data science project.
Constructing the processes	The individual components of the processes must be established and arranged in the correct sequence.
Reducing dimensions	Since many algorithms do not perform well on high-dimensional data, it should be examined whether data dimensions can be removed or combined.
Ensuring validity	Even during model construction, the probability of overfitting can, for example, be reduced by splitting the data into training and test partitions and by using cross-validation.
Considering multiple analytical methods	Where appropriate, several analytical methods may need to be tested or combined by forming ensembles.
Selecting the best parameter configuration	Systematic testing of different combinations is necessary in order to identify suitable or desired parameter settings.
Weighing time against benefit	The quality of the result must be appropriate for the problem context. At the same time, the total computational cost of the analysis must not exceed the value provided by the model.
Ensuring replicability and transparency	Reproducibility and transparency must be ensured, among other things, by storing the transformed data and all configurations of the training process (e.g. the random seeds used).

7.5 Accompanying Task "Tool Selection"

The aim of *tool selection* is to identify a suitable implementation infrastructure for the selected methods. This relates to both hardware and software. This area therefore also overlaps in part with the overarching aspect of *IT Infrastructure* (see Section 10.3), which, however, normally does not form part of the core tasks of data scientists and is defined much more broadly.

The term *tool selection* should therefore be understood more as the selection of individual components of the IT landscape that directly contribute to solving the problem at hand. Depending on the organization, it may be the case that the hardware and software are already predefined and that their selection therefore no longer falls within the scope of the project, although their required use must still be taken into account as a constraint.

In contrast to the requirements relating to the implementation infrastructure, detailed documentation of the selection process is generally necessary only in more extensive projects.

Subtask	Description
Researching suitable software	As soon as it becomes possible to estimate which analytical methods are suitable, it should be clarified which software is to be used to implement those methods and how that software can be procured or developed if it is not yet available.
Researching suitable hardware	Depending on the amount of computing power required and on whether the application is run locally or in a cloud environment, different hardware may be needed.
Comparison with the abilities available in the project team	If a tool cannot be used, or can only be used inadequately, either a different tool must be selected, further training must be initiated, or external resources must be brought in.
Evaluating the tool's suitability	If a tool is not fully compatible with the rest of the project workflow, a compromise must be found between fully implementing the intended method and integrating it into the remaining infrastructure.
Quality assurance during implementation	The quality of the implementation must be ensured, for example through software validation, peer review, or similar measures.

7.6 Core Task "Developing Analytical Methods"

If no suitable analytical method exists, existing methods must—provided this is feasible within the scope of the project—be adapted or combined.

Alternatively, entirely new solutions may be developed. In doing so, it must be determined whether the method should be designed for the broadest possible range of applications or optimized for the specific use case and the data available. The efficiency of developing the method in-house must also be taken into consideration. Unnecessary work—for example, resulting from a failure to make use of existing auxiliary methods—should be avoided. The newly developed method must be integrated into the implementation infrastructure. Time and budget constraints must be taken into account.

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 documented carefully and comprehensively. This documentation may include, for example:

- 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

7.7 Accompanying Task “Evaluation”

Evaluation is a multifaceted task, as it is carried out at three different points:

- during the selection of analytical methods that appear potentially suitable for the problem context,
- during the development of new analytical methods, and
- during the application of the selected or newly developed analytical method to the specific problem context.

In all three cases, the aim is to provide a comprehensible evaluation and classification of the results. In each case, the evaluation is based on the selection of a suitable metric. In addition to technical metrics, the central criteria of the application domain in particular must also be taken into account, since only this perspective makes it possible to determine the actual value of the analysis performed.

As part of method selection, in particular, a comparison of the advantages and disadvantages of the analytical methods under consideration, as well as a description of suitable use cases, must be documented.

For the evaluation of results, the relevant aspects include, in particular, the presentation of the evaluation criteria and their levels, the approach chosen, the test setup, configuration tables, a list of the parameter combinations examined, and the specific test results (including details of execution time).

The experience gained during the evaluation and any potential weaknesses of the method should also be documented.

Finally, the decisions made on the basis of the evaluation must be justified in a comprehensible manner and with reference to the problem context under investigation.

Subtask	Description
Determining the evaluation criteria	The criteria according to which the evaluation is carried out must be selected in a manner appropriate to the domain and with regard to the project objective.
Estimating added value	The benefit expected to result from the analysis must be estimated in advance. This can only be done in the context of the domain-specific problem. Estimating the added value establishes a framework for the level of effort that can reasonably be expended on the analyses.
Reviewing feasibility	The feasibility of the analysis must be assessed with regard to whether the objective set can be achieved, whether the available data are suitable, and whether the resources available are appropriate.
Benchmarking	To assess the results at a later stage, a suitable benchmark must be selected. This may, for example, be an existing method that is to be replaced, or a very simple comparative method that can be used with relatively little effort.
Cost estimate	The effort required to carry out the analytical methods must be estimated. The estimated effort must be significantly lower than the added value expected from the analysis.
Comparing procedures	The fundamental characteristics of the methods under consideration must be identified and compared. On that basis, the fit between the methods and the problem to be addressed must then be assessed.
Evaluating results	The results of the analyses carried out must be assessed. This typically includes a plausibility check, various statistical evaluations, the validation of the results, and an examination of the robustness of the method. In addition, the applicability of the method from the perspective of the domain must be reviewed.
Performance tests	If the analytical method developed is to be transferred into regular operation at a later stage, the performance of the method must be assessed, in particular with regard to the hardware required and the volume of data that can be processed.

7.8 Feature-Bearing Area "Analysis Results"

The results of the analysis process can take very different forms, depending on the problem context, objective, methods used, and the available data basis. The spectrum ranges from descriptive and diagnostic analyses to predictive and prescriptive models, and extends to (partially) automated and adaptive systems.

Characteristic	Description
Meaningfulness	What conclusions can be drawn from the analysis result? Does it amount more to a rough estimate, or to a precise statement? Is it to be expected that the results will remain valid in the future and not only for the current state?
Form of presentation	How are the analysis results communicated? Are they described in a way that is easy to understand? Are they visualized in order to enhance clarity? Are the results presented in detail or in aggregated form?
Type of result	What kind of analysis result is it (e.g. a description of a relationship, an explanation of a relationship, a forecast of future behavior, the derivation of a recommended course of action, or the optimization of a system)?
Ease of generalization	To what extent can the results be generalized to additional data?
Limits	What are the limitations of the developed model? What is the reason for these limitations (e.g. limited data volume, missing attributes, limitations of the analytical method)? How might they be overcome, where appropriate?
Feasibility	Can and should the analytical model be further developed into software that provides the analytical functionality on a continuing basis for new data?
Complexity	How easy are the results to understand, and how readily can measures be derived from them?
Novelty	Were insights gained that would not otherwise have come to light or that were not previously available?
Quantitative evaluation	Which quantitative evaluation metrics are available (significance level, error rate, etc.)?
Relevance	Do the results contribute to solving the original problem, or do they answer a different question / provide additional value? Are the results trivial, or do they generate new insights? Can concrete courses of action be derived from them?
Transparency	Is the process by which the analysis results were produced transparent and comprehensible?
Comparability	Can the analysis results be compared with the results of other methods that are already known?
Comprehensibility	Are the results understandable on their own? Are interpretive aids required?
Completeness	How complete are the results available? Were only individual aspects examined, or was a comprehensive analysis carried out? 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 specific project, this may entail a comprehensive consideration of technical, methodological, and domain-specific tasks, or it may be handled pragmatically. The analysis artifacts may comprise results as well as models or methods themselves and are made available to their intended recipients in different forms.

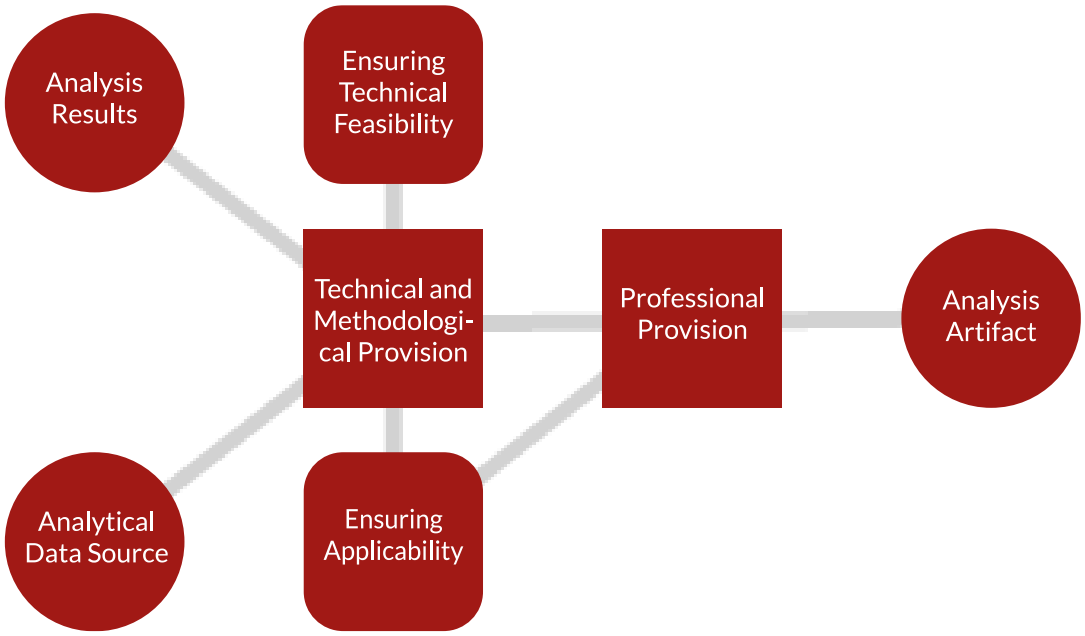


Figure 18: Brief overview of the “Deployment” phase

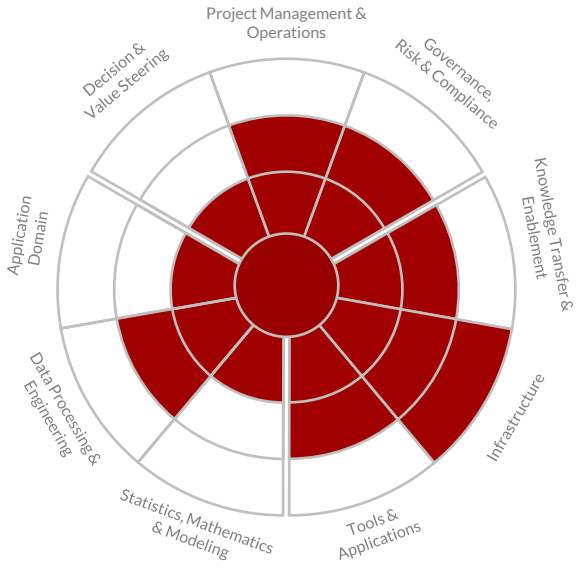


Figure 19: Competence profile for the “Deployment” phase

Detailed Presentation of the "Deployment" Phase

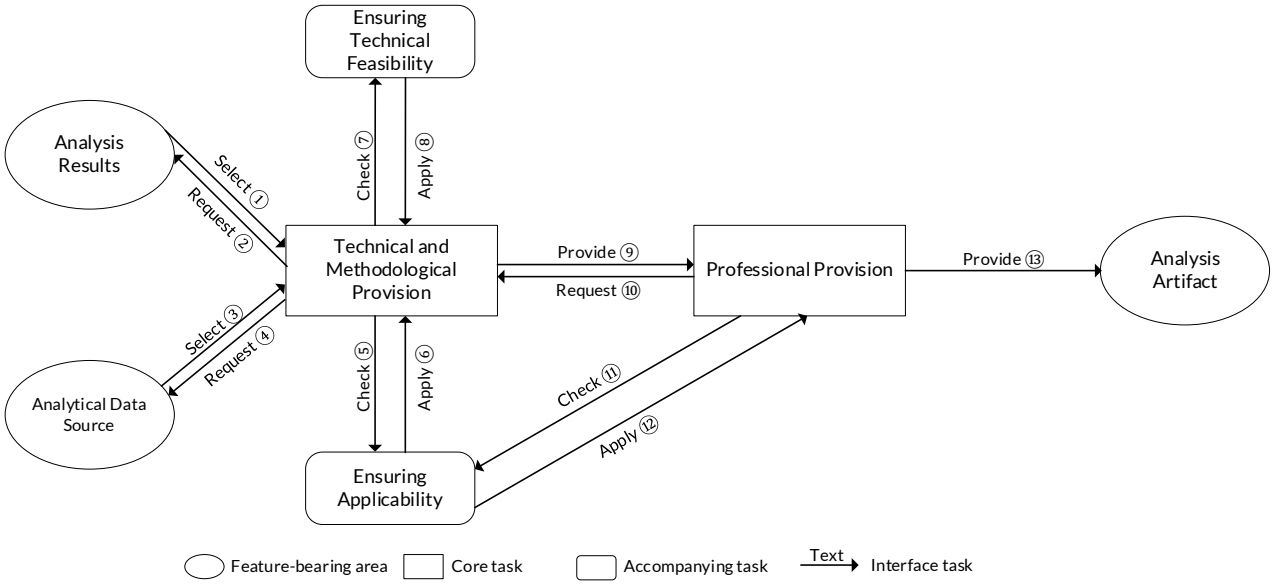


Figure 20: Detailed presentation of the "Deployment" phase

①	The analysis results are selected for the technical and methodological provision.
②	If the analysis results do not suit the intended technical and methodological provision, changes can be requested.
③	If access to data is required for technical and methodological 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 methodological provision, changes can be requested.
⑤	During the technical and methodological 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 methodological provision.
⑦	During the technical and methodological provision, technical requirements must be examined.
⑧	Identified technical possibilities for implementation must be considered during the technical and methodological provision.
⑨	The results of the technical and methodological 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 methodological 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 must be considered during the professional provision.
⑬	Analysis artifacts that arise from the professional provision must be carried over into practical application in a suitable form.

8.1 Feature-Bearing Area “Analysis Results”

Deployment is based on the characteristics of the analysis results described in Section 7.8. No additional special features should be recorded at this point.

8.2 Feature-Bearing Area “Analytical Data Source”

For the deployment of 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 Methodological Provision”

The results of the analysis must be prepared in a form suitable for implementation. Depending on the project, it may also be appropriate to consider several implementation options.

A distinction can be made between the following forms:

- the *manual use* of the results, in which the results are prepared for the target group and communicated, for example, in seminars or workshops,
- the *implementation of the results*, for example in the form of a report in which the results are prepared on a one-time basis,
- the *application of the trained model* so that it can also be used on previously unseen data,
- *continuous learning*, in which the model can adapt itself autonomously through repeated application to previously unseen data, and
- the *publication of the analytical method developed* (possibly only within the organization) in order to enable its use by third parties. In this way, model results can be reviewed independently and weaknesses identified at an early stage.

The model must be embedded in an operational production environment. One-time results are relevant only in exceptional cases (e.g. for a proof of concept); otherwise, the value of the models is generally realized by embedding them in a production environment either continuously or on demand.

The outcome of *technical and methodological provision* therefore consists of implemented and operational solutions in which the analysis results are made available in an appropriate form so that they can be used in existing systems or integrated into operational processes.

Subtask	Description
Preparing the results for the recipients	The results must be prepared in technical and methodological terms so that users can interpret them.
Building the production environment	Where appropriate, it may be necessary to build a new infrastructure in which the results can be continuously updated and taken into account.
Transferring the results	For ongoing operation, it may be necessary to transfer the results from the analytical environment into an operational system.
Context creation	The manner and the time period in which the results were obtained should be evident.
Automating processes	General challenges of process automation must also be taken into account, for example: ▪ What happens in the event of an error? ▪ How should discontinuities between media or systems be handled, and can they be avoided or compensated for? ▪ How can execution be logged in an appropriate manner?
Dealing with IT resources	Efficient use of IT resources must be ensured.
Technically testing the system used	The technically error-free functioning of the analytical system must be verified, especially once it has been integrated into the organization's production environment and connected to real data sources.

8.4 Accompanying Task "Ensuring Technical Feasibility"

The accompanying task of *ensuring technical feasibility* focuses—beyond mere provision—on the sustainable operability of the analytical application. It encompasses ensuring stable and economically viable operation, including (technical) usability, maintainability, and the ability to implement technical adaptations efficiently.

Subtask	Description
Automation	To what extent can data evaluation and the integration of results be automated? At what intervals are the analyses repeated?
Consideration of execution times	How computationally intensive is the algorithm? For example, does it scale well with the volume of data?
Considering time criticalities	Must the analysis be carried out in real time, or is it a non-time-critical analysis that can, for example, be performed overnight in batch mode?
Ensuring operations and support	Who is responsible for the production operation of the analytical application? Who can provide support in the event of technical or methodological questions and problems?
Identifying the hardware stacks	What hardware is required to operate the analytical solution? Which form of implementation (on-premises, private cloud, cloud, IaaS, PaaS, SaaS, etc.) is suitable?
Identifying the software stacks	Is the software stack to be used already predefined by the organization, or does it still need to be evaluated as part of the project? The competencies of the groups involved must also be taken into account here.
Identifying technical conditions and opportunities	It must be examined whether the existing IT infrastructure can be taken into account or, where necessary, procured.
Testing software licenses	Are any further or additional licenses required for the production system?
Legal framework conditions	Have the legal framework conditions for the use of the analytical application (data protection, compliance, etc.) been clarified, defined, and documented?
Dealing with the connected data sources	How can changes in the data sources (formats, quality, access rights, etc.) be addressed? Who is responsible? How is the flow of information organized?
Defining an access concept	Is it possible to restrict access to analysis results to authorized groups of users? Have measures been taken to ensure the security of all data?

8.5 Accompanying Task “Ensuring Applicability”

The aim of *ensuring applicability* is to prepare and provide the analysis results in such a way that they can be understood and used by the intended target group. In particular, consideration must be given to the form in which results are interpreted, further processed, or incorporated into decision-making processes.

Its design takes place in close interaction between methodological expertise and domain knowledge in order to ensure both the domain relevance and the practical usability of the results.

Subtask	Description
Identifying target recipients	In order to ensure applicability, the recipients of the analysis must be known.
Establishing UI/UX design	The interface should be easy for all groups of users to understand and use, while still offering flexibility and covering the complexity of the subject matter. Analysis results should be presented in a comprehensible manner, for example through visualizations.
Ensuring access	Authorization structures and access rights must be defined. Ensuring their implementability forms part of the accompanying task of Ensuring Technical Feasibility.
Involving users	Before the analysis results are put to use, workshops may, for example, be held in order to obtain feedback aimed at ensuring their applicability.
Creating a documentation concept	In addition to technical and methodological documentation, suitable user documentation must also be planned, for example as an aid to interpretation or to describe the indicators used.
Creating a training concept	Depending on the scope of the analysis artifacts developed and the form of their practical application, an appropriate training concept must be developed.

8.6 Core Task "Professional Provision"

Professional provision comprises the purposeful preparation and embedding of the analysis results within the respective application context. The aim is to make the results available in such a way that they can be used in domain-specific processes and integrated into existing decision-making or work routines.

The specific manifestations depend heavily on the form of provision chosen as well as on the requirements of the respective domain. Accordingly, the following section describes generally applicable tasks that support the meaningful use of the results from a domain-specific perspective.

Subtask	Description
Ensuring sustainability	Sustainability means ensuring continued use and ongoing relevance.
Considering reach and impact	Before the results are published beyond the project team, their potential impacts must be assessed, including from ethical and economic perspectives.
Considering legal issues	Data protection and legal issues must be assessed before the analysis results are used.
Establishing points of contact	There must be domain-specific contacts available to answer questions during ongoing use. A defined means of establishing contact must also be specified.
Integration into existing processes	The domain-specific integration of the analysis artifacts into existing processes must be ensured.
Internal cost calculation model	The personnel and IT costs associated with operating the analysis artifacts must be determined and, where appropriate, allocated to the users.
Conducting training sessions	The training measures designed as part of ensuring applicability must be carried out in an appropriate form (in-person training sessions, online training, webinars, etc.).
Creating a user handbook	The user documentation conceived as part of ensuring applicability must be created.
Determining troubleshooting	Review mechanisms and procedures must be defined for cases in which the analysis artifact no longer produces meaningful results.

8.7 Feature-Bearing Area “Analysis Artifacts”

The characteristics of the *analysis artifacts* depend on the form in which the results are presented (cf. Section 8.1).

Characteristic	Description
User documentation	Users of the analytical system must be provided with a user manual describing the available reports, dashboards, databases, etc., including the corresponding access rights. In addition, domain-specific contacts must be designated.
Technical documentation	For the maintenance and further development of the analytical system, a detailed description of the software used or developed must be available, including the code base, inputs and outputs, intermediate steps performed, and dependencies on other components. In addition, the technical infrastructure created for the analytical system, or into which it has been integrated, must be documented. Technical contacts must also be designated in this context.
Model documentation	The analytical models must be described in detail in order to enable their adaptation and future further development, including the assumptions underlying their use.
Recommended actions	At least in cases where results are used manually, recommendations for action must be defined for the recipients of the analysis artifacts.
Models	The models resulting from the analysis can be applied to new data.
Reports	The data resulting from the analysis must be presented in the form of reports tailored to the target audience.
Analysis infrastructure	For the sustained use of analytical models, it is often necessary to provide a specific analytical infrastructure, which in turn is itself embedded in the organization’s IT infrastructure.
Support	Defined domain-specific and technical support is required both to support ongoing operations and to resolve incidents.

9 Usage

The *usage* of analysis artifacts is not to be regarded as a primary part of a data science project. However, depending on the form of deployment, monitoring is necessary in order to assess the model's continued suitability in use and, where appropriate, to derive insights from its use for further development and new development, including in the context of iterative approaches.

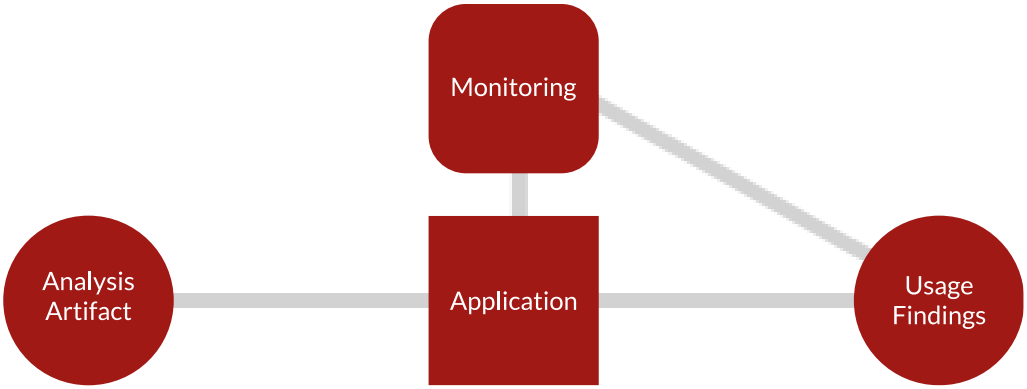


Figure 21: Brief overview of the "Usage" phase

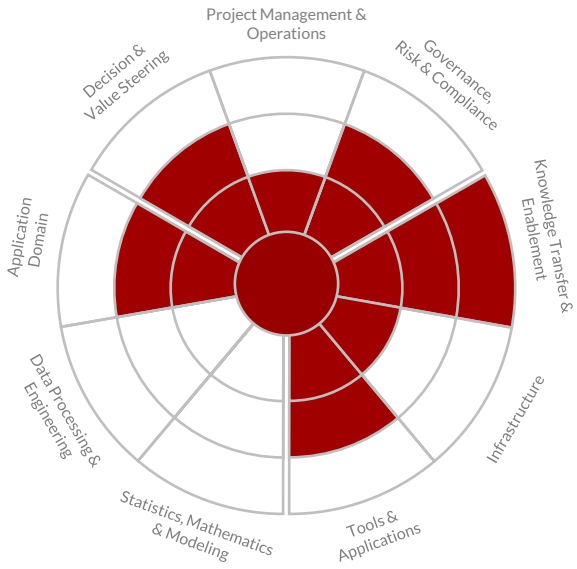


Figure 22: Competence profile for the "Usage" phase

Detailed Presentation of the “Usage” Phase

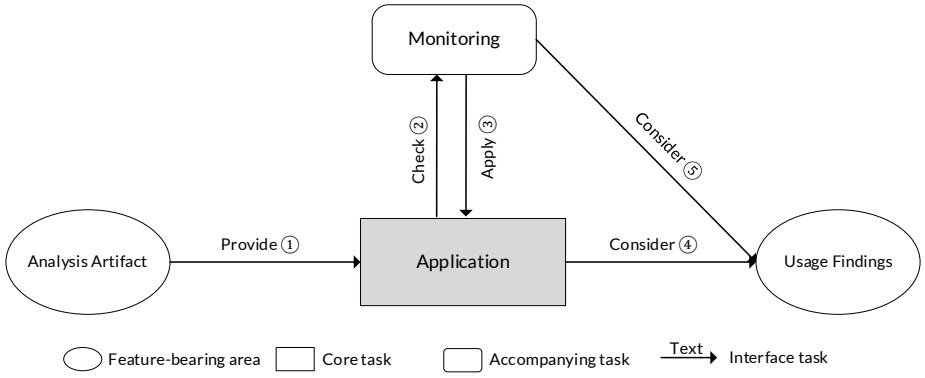


Figure 23: Detailed presentation of the “Usage” phase

①	The analysis artifacts are provided for use.
②	Usage must be repeatedly monitored.
③	Results of the monitoring must be considered during use.
④	Domain-specific findings arising from use must be considered during new and further developments.
⑤	Data science-specific findings arising from use must be considered during new and further developments.

9.1 Feature-Bearing Area “Analysis Artifacts”

The use 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”

Within *monitoring*, the regular operation for which the analysis artifact is designed over the long term must be supervised. In particular, the quality of the analysis results must be reviewed continuously and the model’s continued applicability must be verified.

An outcome of this area of responsibility is a report that enables an assessment of the usefulness of analysis artifacts.

Subtask	Description
Analysis artifacts in general	
Ensuring the correct application domain	The analysis artifacts were created for a specific domain. This specialization must be preserved.
Evaluating the analysis artifacts	The results of the analysis should be evaluated on a recurring basis with regard to their informative value and predictive performance.
Checking the sustainability of the analysis artifacts	It must be examined whether the analysis artifacts are maintained and how quickly the results become outdated.
Using analysis artifacts	
Checking the data	The model may be applied to data that did not yet exist at the time it was created. It must be ensured, as far as possible, that its use yields correct results. This should be verified by both data experts and domain experts.
Monitoring errors	Error reports must be collected and evaluated; these include, among other things, unexpected model behavior or new forms of data errors.
Metadata of the usage	
Recognizing performance challenges	The identification of performance-related challenges during deployment is limited. This aspect should therefore also be monitored during use.
Evaluating usage data	It must be examined whether the analysis artifacts should continue to be used. For this purpose, usage data must be recorded.

9.3 Feature-Bearing Area “Usage Findings”

On the basis of the usage findings, a decision can be made as to whether the use of analysis artifacts should be discontinued or whether they need to be revised. This may become necessary either because conditions have changed or because the solution developed has not proven effective in productive use.

Characteristic	Description
Error reports	The reports make it possible to assess whether the analysis artifacts can be operated with sufficient stability.
Usage frequency	If analysis artifacts are not used by domain experts, their continued operation may be unnecessary and further development may, where appropriate, be dispensed with.
Performance of the analysis artifacts	A consideration of performance makes it possible to assess the suitability of the technical infrastructure used.
Type of use	The nature of use may influence potential further development from the perspective of the domain.

10 Overarching Aspects

In addition to the phases of DASC-PM, there are aspects that must be taken into account in all phases of a data science project and that significantly influence its design. These include, in particular, the domain, methodology, and the IT infrastructure.

These overarching aspects do not operate in isolation; rather, they shape the individual phases in different ways—from the formulation of the problem statement through the processing and analysis of data to the provision and use of the results. Taking them into account is therefore a central prerequisite for the successful execution of data science projects.

The overview below relates the overarching aspects to the individual phases and provides compact guidance. The following subsections describe these aspects and their significance for project practice in greater detail.

	Domain	Methodology	IT Infrastructure
Problem Statement	Domain-specific objectives and problem definition shape the selection and design of use cases	Clear definition of the object of investigation and a comprehensible objective	Assessment of fundamental technical feasibility
Data Provision	Data relevance and data quality are assessed in the context of the domain	Transparent, reproducible data preparation and well-founded data analysis	Available data access, interfaces, and computing resources determine data processing
Analysis	The selection, application, and evaluation of methods are guided by domain-specific requirements	Methodologically sound selection, application, parametrization, and evaluation of methods	Infrastructure influences the selection and feasibility of the analytical methods
Deployment	Preparation and provision are guided by the domain-specific requirements of the target group	Traceable provision and documentation of the results	Technical integration, operation, and scalability of solutions
Usage	The use and evaluation of the results take place in the application context	Systematic evaluation of use and review of model assumptions	Ongoing review of operational performance and efficiency

10.1 Domain

The domain constitutes the substantive context in which a data science project is carried out. It largely determines which problem contexts are relevant, how data are interpreted, and what requirements are placed on analytical methods and their results. The domain thus influences all phases of the project.

Problem Statement

The domain shapes the identification and elaboration of use cases. Domain-specific objectives, conditions, and requirements arise from the application context and determine which problem contexts can be addressed in a meaningful way.

Data Provision

The relevance of data can only be assessed within the context of the domain. A sound understanding of the data can likewise only be developed by taking the domain-specific interrelationships into account.

Domain-specific requirements—such as regulatory provisions or established data standards—must be taken into account during data preparation. Transformations, such as making measurements comparable or classifying outliers, must also be carried out in the context of the domain.

In exploratory data analysis, the domain-specific problem context is specified in greater detail and translated into the examination of the data, with the aim of generating insights that are relevant to the domain.

Analysis

Domain-specific conditions may restrict the selection of analytical methods or rule out certain methods altogether. At the same time, the domain gives rise to requirements regarding the design of the methods, for example with respect to interpretability or the consideration of causal relationships.

In many domains, there are established methods that can serve as a reference for the selection of suitable approaches. In addition, the desired form of the results influences the selection of methods.

The evaluation of the analysis results requires that they be situated within the domain-specific context. Relevant metrics, as well as requirements regarding the presentation of the results, likewise arise from the domain.

Deployment

The preparation and provision of the analysis results must be aligned with the domain-specific requirements of the target group. Both the applicability of the results and the chosen form of provision must therefore be designed within the context of the domain.

In addition, domain-specific conditions, in particular non-functional requirements, must be taken into account in the technical implementation.

Usage

The actual use of the analysis results takes place within the domain-specific application context. Their evaluation in terms of usefulness, as well as the derivation of further requirements or improvements, is therefore closely tied to the domain.

10.2 Methodology

The execution of data science projects requires a structured and methodologically sound approach guided by fundamental scientific principles. The degree of methodological rigor may vary depending on the project context, particularly with regard to the depth of theoretical grounding or the systematic review of existing literature. In practice-oriented projects, this may deliberately be reduced, provided that an appropriate cost–benefit assessment is carried out and the associated risks are taken into account. These risks include, in particular, the possibility that an insufficient engagement with existing methods or findings may lead to relevant solution approaches being overlooked and/or the solution space being unnecessarily restricted.

In principle, data science projects are subject to the same requirements as other scientifically oriented undertakings. The object of investigation must be clearly defined so that the objective and context are comprehensible to third parties. The results achieved must go beyond existing findings or offer a new perspective. At the same time, it must be ensured that the results provide a recognizable benefit and are documented in an appropriate manner so that they can be reviewed and reproduced. In particular, the requirements of objectivity, reliability, and validity must be taken into account.

Problem Statement

The problem statement must be formulated in such a way that it is clearly delimited and comprehensible. A precise definition of the object of investigation forms the basis for the selection of suitable methods and for the interpretation of the results.

Data Provision

The suitability of the data for addressing the problem context must be examined systematically. Data preparation steps must be designed in a transparent, comprehensible, and reproducible manner. This includes, in particular, documenting all transformations and ensuring the long-term availability of the raw data. In exploratory data analysis, a sound understanding of the data must be established. Potential sources of error must be identified, and suitable statistical methods must be applied to assess the data.

Analysis

The following applies both to the selection and to the application of analytical methods: requirements and objectives must be defined in a comprehensible manner, and the selection of suitable methods must be made with due regard to existing approaches and current scientific knowledge. Where new or adapted methods are developed, it must be explained to what extent existing approaches are being used, extended, or replaced, and why the development of a new method is necessary. The parametrization of analytical methods must be carried out in a purposeful manner. In doing so, it must be ensured that the underlying assumptions of the methods are satisfied. Suitable scientific sources must be consulted in order to ensure correct application.

Analysis processes and results must be documented comprehensively and interpreted in a comprehensible manner. Evaluation must be considered from the outset, and suitable testing and validation approaches must be applied. This is particularly necessary in order to ensure that the results do not merely constitute statistical artifacts of the data under consideration.

Deployment

Analytical methods and their results must be provided in such a way that their functioning remains comprehensible. This includes, in particular, appropriate documentation as well as a transparent description of the methods used and their limitations.

Usage

The use of analysis artifacts must be systematically monitored and evaluated. Hypotheses regarding the performance of the models must be examined in the context of their use. In doing so, differences between the development and deployment environments, as well as changes in the usage environment, must be documented.

10.3 IT Infrastructure

IT infrastructure forms the technical foundation for the execution of data science projects. It influences all phases of the project, for example through constraints relating to available technologies, data access, or computing capacities.

The specific significance of the IT infrastructure depends heavily on the type and scope of the project. In particular, the complexity of the data, the requirements imposed by the analytical methods, and the intended form of deployment determine which infrastructural prerequisites are necessary. At the same time, existing systems and organizational conditions must be taken into account, as they have a significant influence on the design and implementation of solutions. This applies primarily to those systems that are directly connected with the data science project, but also to infrastructure that may be used for project management or collaboration within the team.

Problem Statement

At an early stage, it must be examined whether the planned project can be implemented using the existing IT infrastructure or an IT infrastructure that is feasible under the given economic conditions. This includes, in particular, evaluating the available systems, interfaces, and technical resources.

Data Provision

The available access options and interfaces to data sources play a decisive role in determining how data can be accessed and processed. Constraints imposed by source or target systems may influence the choice of technologies.

In addition, the available computing capacity must be taken into account during data preparation and exploratory data analysis. In particular, when dealing with large or complex data sets, it may be necessary to process data directly within existing systems or to carry out appropriate data extractions.

Analysis

The selection and application of analytical methods are closely linked to the technical possibilities available. It must be evaluated which infrastructure is required to carry out the analyses and whether it is available or can be procured.

In addition, it must be examined whether analyses can be carried out directly on the available data or whether separate processing is necessary.

Deployment

The IT infrastructure must be suitable for providing and operating the analytical methods developed in the intended form. This includes, in particular, aspects such as integration into existing systems, access options, and requirements relating to availability, scalability, and security.

In addition, possibilities for the maintenance, updating, and further development of the solutions provided must be taken into account.

Usage

During ongoing operation, it must be reviewed whether the chosen IT infrastructure continues to meet the requirements. This concerns both the performance and the efficiency of the systems in use. Adjustments may become necessary if usage requirements or operating conditions change.

References

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.

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

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

Europ. Parl. (2026). What is artificial intelligence and how is it used? https://www.europarl.europa.eu/topics/en/article/20200827STO85804/what-is-artificial-intelligence-and-how-is-it-used, retrieved on July 8, 2026.

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

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.

Neuhaus, U., Schulz, M., Schröder, H. et al. Kompetenzfelder künftiger Beschäftigter im Bereich Künstlicher Intelligenz. HMD 61, 471–484 (2024). https://doi.org/10.1365/s40702-024-01046-7.

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.

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.

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.

Wicht, P.; Hausmann, A.; Hoseini, S.; Kaufmann, J. (2021): Aufbau eines Data-Science-Teams – „Lessons learned“. In: *Wirtschaftsinformatik & Management* 13, 266–275. https://doi.org/10.1365/s35764-021-00350-x.

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.

Further Publications on DASC-PM

English-Language Publications

Kuehnel, S., Neuhaus, U., Kaufmann, J., Schulz, M., & Alekozai, E. M. (2023). Using the Data Science Process Model Version 1.1 (DASC-PM v1.1) for Executing Data Science Projects: Procedures, Competencies, and Roles. In: Barton, T. & Müller, C. (eds.): Apply Data Science: Introduction, Applications and Projects, 119-134, ISBN: 978-3-658-38797-6, https://doi.org/10.1007/978-3-658-38798-3_8.

Schulz, M., Neuhaus, U., Kaufmann, J., Badura, D., Kühnel, S., Badewitz, W., Dann, D., Kloker, S., Alekozai, E. M., & Lanquillon, C. (2020). Introducing DASC-PM: A Data Science Process Model, Australasian Conference on Information Systems, ACIS Proceedings, paper 45, 2020, Wellington, New Zealand, https://aisel.aisnet.org/acis2020/45, http://dx.doi.org/10.25673/92266.

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, NORDAKADEMIE gAG Hochschule der Wirtschaft, Elmshorn 2022, ISBN: 978-3-9824465-1-6, http://dx.doi.org/10.25673/91094.

Schulz, M., Neuhaus, U., Kühnel, S., Rohde, H., Hoseini, S., & Theuerkauf, R. (eds.) (2023). DASC-PM v1.1 Case Studies, NORDAKADEMIE gAG Hochschule der Wirtschaft, Hamburg 2023.

German-Language Publications

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 – Einführung, Anwendungen und Projekte, ISBN: 978-3-658-33813-8, 127-144, https://doi.org/10.1007/978-3-658-33813-8_8.

Kaufmann, J., Kühnel, S., Theuerkauf, R., Alekozai, E. M., Hoseini, S., Neuhaus, U., & Schulz, M. (2021). Where is the Science in Data Science Projects? Online-Workshop über die Wissenschaftlichkeit von Vorgehensmodellen für Data-Science-Projekte (WISDAP), Gesellschaft für Informatik e.V. (GI) (eds.): INFORMATIK 2021, Lecture Notes in Informatics (LNI), Bonn 2021, 1729-1741, https://doi.org/10.18420/informatik2021-150.

Kaufmann, J., Schulz, M., & Kühnel, S. (2025). Data Science – Projekte mit DASC-PM. In: WISU - Das Wirtschaftsstudium, Sektion Wirtschaftsinformatik, 54 Jg., Heft 3/2025, 263.

Schulz, M. (2020). Data-Science-Projekte und ihre Besonderheiten. In: Wirtschaftsinformatik & Management, 12, 376-381, https://doi.org/10.1365/s35764-020-00281-z.

Schulz, M., Neuhaus, U., Kaufmann, J., Badura, D., Kerzel, U., Welter, F., Prothmann, M., Kühnel, S., Passlick, J., Rissler, R., Badewitz, W., Dann, D., Gröschel, A., Kloker, S., Alekozai, E. M., Felderer, M., Lanquillon, C., Brauner, D., Gölzer, P., Binder, H., Rohde, H., & Gehrke, N. (2020). DASC-PM v1.0 - Ein Vorgehensmodell für Data-Science-Projekte, NORDAKADEMIE gAG Hochschule der Wirtschaft, Hamburg, Elmshorn 2020, ISBN: 978-3-00-064898-4, https://doi.org/10.25673/32872.

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 - Ein Vorgehensmodell für Data-Science-Projekte, NORDAKADEMIE gAG Hochschule der Wirtschaft, Elmshorn 2022, ISBN: 978-3-9824465-0-9, https://doi.org/10.25673/85296.

Schulz, M., Neuhaus, U., & Kühnel, S. (2020). Data-Science-Prozessmodell (DASC-PM). In: WISU, Basiswissen Wirtschaftsinformatik, 49 Jg., Heft 4/2020, 387 ff.

Schulz, M., Neuhaus, U., Kühnel, S., Rohde, H., Hoseini, S., & Theuerkauf, R. (eds.) (2023): DASC-PM v1.1 Fallstudien, NORDAKADEMIE gAG Hochschule der Wirtschaft, Hamburg 2023.

Theuerkauf, R., Daurer, S., Hoseini, S., Kaufmann, J., Kühnel, S., Schwade, F., Alekozai, E. M., Neuhaus, U., Rohde, H., & Schulz, M. (2022). Vorschlag eines morphologischen Kastens zur Charakterisierung von Data-Science-Projekten. In: Informatik Spektrum 45, https://doi.org/10.1007/s00287-022-01508-6.

Index of Authors

The list of authors includes everyone who actively participated in revising and editing Version 2.0 and consented to being listed.

They would like to thank everyone who participated in Version 1.0 and 1.1 for their work on the previous publications.

Prof. Dr. Michael Schulz, NORDAKADEMIE Hochschule der Wirtschaft

Prof. Dr. Jens Kaufmann, Hochschule Niederrhein

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

Dipl.-Inform. Uwe Neuhaus, Europa-Universität Flensburg

Heiko Rohde (M.Sc.), valantic Business Analytics GmbH

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

Prof. Dr. Carsten Lanquillon, Hochschule Heilbronn

Prof. Dr. Stephan Daurer, DHBW Ravensburg

Jonas Dieckmann (M.Sc.), Philips

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

Felix Welter (M.Sc.), CGI

Prof. Dr. Uwe Haneke, Hochschule Karlsruhe

Prof. Dr. Christian Beecks, FernUniversität in Hagen

Dr. Martin Böhmer, Martin-Luther-Universität Halle-Wittenberg

Prof. Dr. Bahne Christiansen, NORDAKADEMIE Hochschule der Wirtschaft

Prof. Dr. André Drews, Technische Hochschule Lübeck

Prof. Dr. Arne Ewald, NORDAKADEMIE Hochschule der Wirtschaft

Dr. Viktor Harkov, HEON GmbH

Franziska Herrmann (B.Sc.), NORDAKADEMIE Hochschule der Wirtschaft

Tim Hilbig (M.Sc.), Euler Hermes

Prof. Dr. Dirk Johannßen, Fachhochschule Westküste

Dipl.-Ing. oec. Lutz Kretschmann, Hapag-Lloyd AG

Prof. Dr.-Ing. Bernhard Meussen, NORDAKADEMIE Hochschule der Wirtschaft

Dr. Christian Pasold, Datron Consulting GmbH

Stefan Rösl (MBA), Ostbayerische Technische Hochschule Amberg-Weiden

Dr. Klaus Schmerler, Martin-Luther-Universität Halle-Wittenberg

Theo Schnelle de Lourenco (M.Sc.), Aclue GmbH

Prof. Dr. Martin Schultz, Hochschule für Angewandte Wissenschaften Hamburg

Prof. Dr. Michael Seifert, Technische Hochschule Rosenheim

Marcus Soll (M.Sc.), NORDAKADEMIE Hochschule der Wirtschaft

Dr. Robert Stahlbock, Universität Hamburg

Niklas Ullmann (M.Sc.), Synvert GmbH

dasc°pm