Data Science Canvas: Evaluation of a Tool to Manage Data Science Projects

Thomas Neifer
University of Siegen
Bonn-Rhein-Sieg University of Applied Sciences
thomas.neifer@uni-siegen.de

Dennis Lawo
University of Siegen
Bonn-Rhein-Sieg University of Applied Sciences
dennis.lawo@uni-siegen.de

Margarita Esau
University of Siegen
Bonn-Rhein-Sieg University of Applied Sciences
margarita.esau@uni-siegen.de

Abstract

Data emerged as a central success factor for companies to benefit from digitization. However, the skills in successfully creating value from data – especially at the management level – are not always profound. To address this problem, several canvas models have already been designed. Canvas models are usually created to write down an idea in a structured way to promote transparency and traceability. However, some existing data science canvas models mainly address developers and are thus unsuitable for decision-makers and communication within interdisciplinary teams. Based on a literature review, we identified influencing factors that are essential for the success of data science projects. With the information gained, the Data Science Canvas was developed in an expert workshop and finally evaluated by practitioners to find out whether such an instrument could support data-driven value creation.

1. Introduction

In times of ever-increasing data amounts and the ascribed value to data, companies are challenged by creating value from it. Due to volume, velocity, and validity of big data, the toolbox around machine learning and mathematical-statistical procedures becomes ever more important. However, the implementation of such procedures into valuable business models requires expert knowledge [1]. In view of today's fast-moving markets and the prevailing competition, it is urgently necessary for decision-makers to evaluate the opportunities that arise from their data, the existing expertise in their company, and bring both together to create value [2]. To drive digital transformation for their organizations, decision-makers should be aware of how to scale the value of their data assets and the capabilities and opportunities of analytics [3]. However, according to a study by the Data Literacy Project ($n = 7,377$), only 24% of these decision-makers consider themselves to be data competent, although this is perceived as increasing in work performance [4]. Despite its high relevance, the introduction of a data strategy has also not yet reached any maturity in many companies [5]. This requires a basic knowledge regarding data analysis as well as an active data-driven corporate culture to enable data-driven value creation and business models (DDBM) [6].

Therefore there is a practical need to provide decision-makers with tools to support data-driven value creation processes [7]. For other fast-developing and complex business problems, a concept for a quick implementation in the sense of 'Learning by Doing' emerged. For example, Lean Manufacturing [8] and Lean Startup [9] are supported by the Lean Canvas as an orientation basis. Another example is the 'Business Model Canvas' developed by Alexander Osterwalder, which simplifies the development and communication of a business model by helping to visualize and understand its elements [10]. The intention of a canvas model is to summarize a complex problem as clearly and simply as possible and give guidance for solution-finding.

Also for DDBM some canvas approaches, e.g. the Data Canvas [11] or the canvas for data-driven ideation workshops [12] were explored. While those focus rather on the data and algorithmic value creation [13], canvases such as the Data Insight Generator seek to more strongly include the customers value proposition [13, 7]. Still, this canvas only focuses on "an illustration between the [...] two components, key resources and value proposition" [13] rather than providing a holistic approach that considers the cost and revenue structure as well as stakeholders of DDBMs [14, 15].

Having identified the need to support decision-makers regarding DDBMs, which qualifies as a "heretofore unsolved and important business problem" [16] and the lack of a holistic canvas approach, the purpose of this paper is to communicate the design (first design cycle) and evaluation (proof-of-concept) of the Data Science Canvas – a tool to holistically support decision making and stakeholder communication on

URI: https://hdl.handle.net/10125/71277
978-0-9981331-4-0

DDBMs. Our canvas (artifact), thereby, aims at being sufficiently specific and holistic at the same time to be readily used by managers. This means it covers both the technical and managerial aspects of the data problem at hand, to address the specified problem. Thereby, our research follows the Design Science Research [16] approach as outlined by Peffers et al. [17], to first design the canvas as an artifact that aims to solve the problem and derive insights and theoretical implications [18]. As shown in Table 1 the remainder of the paper is structured along the phases outlines by Peffers et al. [17]. Having motivated the problem and defined objectives of a holistic solution in this section, section 2 reports on the design and development of our canvas (artifact). This section includes a literature review of related artifacts as a presentation of the related work [19] and an expert workshop to develop our prototype. In section 3, the canvas is demonstrated and evaluated on real-life data projects as a naturalistic ex-post evaluation [20] to understand usefulness in the context of real projects and decision making. The results are discussed in section 4 and we draw conclusions in section 5.

2. Data Science Canvas Design & Development

2.1. Literature Review: Identifying and Understanding Relevant Elements

2.1.1. Procedure. To highlight important elements and delimitations of the concept to be evaluated, a literature review was conducted according to the four phases of Rowley and Slack [21]. We searched for scientific or practical literature that aims to solve the identified problem by means of designing an artifact. As the terms data science, big data, and machine learning are often used interchangeably [22], we used the following keywords and their combinations: canvas, data science, data analytics, machine learning, big data, business model. Based on the initial literature, a forward and backward reference search was conducted [23, 24]. The main inclusion criteria for the surveyed literature was the provision of an artifact in the form of a canvas and a description of the elements to allow for comparison.

Building on the SQ3R-Method [25] to review the found papers, we identified nine canvas models, which were then recorded in the form of a concept matrix [26] using a conceptual framework [24].[1] As the identified canvas models used different wordings, we analyzed the tiles according to their meaning and role and clustered and renamed them to form a unified scheme. This procedure was conducted collaboratively by two authors to ensure agreement on the resulting elements. Resulting from this, the initial 20 factors were reduced to 12. As a first result of the literature research, Table 2 shows the basis of comparison as a concept matrix and allows a more detailed analysis of the canvas model elements.

2.1.2. Finding: Identified Elements. The concept matrix in Table 2 shows that there are in principle many elements for a Data Science Canvas concept. However, some of these elements are only considered to a minor extent in literature, as the only occur in single sources. Data source, methods, model quality, added value, and stakeholders are among the most important elements, with four out of five mentions each. Data collection, data quality, and software for data analysis with only one entry each are not often considered. In the following, we more deeply discuss the elements.

Business Case (1): Although the definition of a business case is one of the crucial positions in the data science process, it is integrated into the canvas model as an element from five sources. In some cases, however, an exact definition of a business case is lacking. Thus, this element is often oriented towards the selection of methods and quality rather than the problem definition. Furthermore, used headings and questions supporting the answer are partly misleading and therefore not unambiguous. For example, the AI Canvas uses the heading *”Prediction”* with the supporting question *”What do you need to know in order to make the decision?”* [29]. Moreover, this mainly addresses the ML area and ignores other areas such as clustering or classification.

Data Collection and Source (2, 3): Following the objective delimitation of the business case, data is collected within the framework of its operationalization – in analogy to a statistical investigation – to obtain the required data. If this involves data still to be collected, a survey method and the necessary scale level must be specified. Both the survey method and the scaling are determined by the problem and the methodology [33]. Five approaches have such an element, but they do not sufficiently highlight the data requirements and potentially necessary cleansing needs [31, 11, 32].

To ensure the correctness and accuracy of the model, the source of the data should be analyzed. This is particularly important in the case of external data. Six sources examined represent such an element, whereby only one canvas concept explicitly addresses internal and external data resources [29, 32].

[1] A list of all identified canvas models can be provided on request.

Table 1. Overview of the First Cycle of the DSR Project [17]

Identify Problem and Motivate (Section 1)	Empowering the Management to understandand create value with data and enable collaboration of different stakeholders of data-driven business models is an unresolved problem.
Define Objectives of a Solution (Section 1)	A holistic canvas concept is required to support management in understanding and making decisions of data-driven business models. Such canvas concepts have also been successfully employed in other fast-developing and complex domains.
Design and Development (Section 2)	The holistic Data Science Canvas is designed on the basis of a literature review on other canvas approaches and an expert workshop.
Demonstration and Evaluation (Section 3)	We conducted a real-world trial with 10 participants including a post-interview to prove the usefulness of our concept and understandthe design space for further design cycles.
Communication	The Canvas description and the evaluation results are communicatedin this research paper.

Table 2. Concept Matrix.

Elements Canvas-Models	1	2	3	4	5	6	7	8	9	10	11	12
Big Data Management Canvas [27]			X		X	X		X			X	X
Deep Learning Canvas [28]	X			X		X	X		X		X	X
AI Canvas [29]	X		X			X			X	X		
AI Project Canvas [30]			X		X		X		X	X	X	X
Machine Learning Canvas [31]	X	X	X			X			X		X	X
Data Insight Generator [13]	X	X		X	X	X		X			X	
Data Canvas [11]		X	X									
Data Innovation Board [12]	X	X	X							X	X	X
Data Project Canvas [32]		X					X	X		X	X	X
Legend:	1: Business Case 2: Data Collection 3: Data Source			4: Data Quality 5: Data Integration 6: Methods			7: Skills 8: Software 9: Model Quality			10: Costs, Revenues 11: Value Added 12: Stakeholders		

Data Quality (4): By means of explorative analyses and visualizations, the quality of the data – whether already available or not – should be validated. This may reveal the need for data cleansing, as the presence of erroneous or distorted data could lead to erroneous models and management making the wrong decisions. This element is only rudimentarily and casually represented by two canvas models [28, 13].

Data Integration (5): Once the data quality has been checked, the data (from the different sources) should be merged into one database. This will facilitate the access by the data analyst and improve the quality of information through accessibility, completeness, clarity, accuracy, and consistency [34]. This element occurs in three sources, but is explicit and comprehensible only in two cases [27, 13].

Methods (6): The heart of Data Science is the transformation of data into knowledge [2]. The conversion takes place through algorithms which fall back on mathematical-statistical procedures. The central task of an analyst is to select the right method for the respective problem based on the available data. In principle, there is a wide range of statistical instruments (multivariate methods) for data analysis, but the special nature of the data must be taken into account [33].

Procedures for segmentation, classification, estimation (regression) and association can be differentiated. Different methods are ranging from white box statistical approaches to black box machine learning applications to address these problems [35].

Five canvas concepts have an element for method selection. However, they usually limit themselves to a certain method category (e.g. deep learning or machine learning methods) [31, 28]. Furthermore, no concept offers support in the choice of methods beyond the canvas. This makes it impossible for decision-makers to use the respective canvas without methodological knowledge.

Skills (7): The analyses requires skills from the fields of statistics, computer science and mathematics. Skill sets to be defined are also suitable for a job description and advertisement for the respective activities corresponding to the business case. Against the background of insufficiently though-out and formulated job advertisements to a data scientist, this seems to make sense [36]. In addition to methodological knowledge (methods, tools and libraries, programming, databases), a basic understanding of business and economics as well as domain knowledge from the respective company should also be addressed to generate deeper insights [37].

Furthermore, the various roles in connection with data preparation and analysis should also be differentiated: The Data Scientist is mainly concerned with complex methods of data analysis, while the

Data Engineer is responsible for software solutions for Big Data and, e.g., creates data pipelines. This differentiation is still missing in many companies, although even success could be endangered due to a lack of understanding of the distribution of tasks and roles [38]. Only three of the identified sources make it possible to deal with the required skills of the (required) personnel.

Software (8): Based on the method, the existing skills/personnel, and resources, appropriate software must be selected. The provided program libraries also play a role in this, to fall back on predefined and useful functions. The 19th KDnuggets Software Survey from 2018 identifies Python, RapidMiner, and R as the most popular software of the 2,052 software users asked [39], but also Excel is present in the top 5. From a managerial perspective, software is an importantissue, as it directly relates to the cost of procuring software packages or services for training and runningthe models.

There are only three sources that have an element for mapping the software. However, there is no clear distinction between method and software selection nor are libraries, as a fine-granular decision, addressed [27].

Model Quality (9): Once the model has been developed, its quality must be assessed. Depending on the chosen method, different indicators are used. For classification problems, a confusion matrix is often used, which compares actual with predicted classification based on the existing classes [40]. In the case of regression-based methods, criteria such as the mean square error or the (adjusted) coefficient of determinationare used [33].

The literature research showed that four canvas models allow an evaluation of the model. One concept event differentiates evaluations before and during the use of the model. If the models are working in real-time, such a distinction proves to be useful. However, there is no further explanation or clarification of the indicators.

Costs and Revenues (10): In the business context, costs and revenues are key variables, especially for justifying (data analysis) projects to top-level management. In analogy to the Business Model Canvas, the costs and their structure on the one hand, and the possible revenue streams of the project on the other hand are listed here. Four of the sources have elements that include costs and revenues. One analyses only the costs with and without the model, while the other differs concretely costs and incomes [29, 30].

Value Added (11): This element also originates from the Business Model Canvas and is intended to show how the model can be used to support the internal and/or external customers [10]. Without a prior specification of the added value, any project is questionable. Seven sources have such an element.

Stakeholders (12): Finally, the results of the analysis and the resulting added value must be presented to the stakeholders involved in the project. If the model is implemented in the company, it is also important to support them in dealing with the model. For target-oriented control, the ability of Data Storytelling is also required here. Data Storytelling stands for the preparation and presentation of data in the form of a scenario within a story, tailored to the target group. This makes it easier for the target group to understand the meaning behind the data and motivates action [41, 42].

The analyzed canvas concepts underline the relevance of the consideration of stakeholders in a canvas model. Six concepts have such an element. Often, however, the requirements of the respective target group for the presentation and communication of the results are not mapped. In analogy to the various notation methods and their specific depth in process modeling, it makes sense to use a different representation, history, and complexity for the top-level management than for the specialist departments and their users [43].

2.1.3. Findings: Relation of Elements. To understand the relation of our elements and add a procedural view, we mapped them to a data science process (cf. Figure 1). The relation is thereby very close to the iterative processes as identified by Cielen et al.[44] but also CRISP-DM [45].

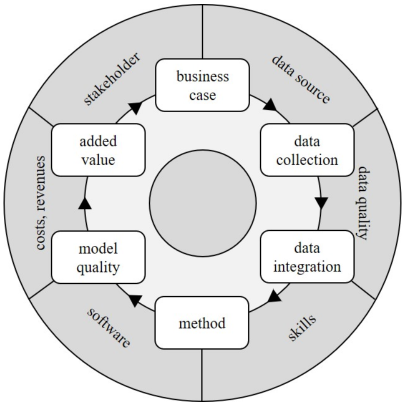


Figure 1. Data Science Process.

As a starting point and core of a new product or service, the business case represents a strategic decision to focus data analysis on a specific business objective. At the same time, this also implies an orientation to the requirements and needs of the customer (stakeholder) and commits to continuous measurementand optimization [46].

The data required to map the business case may not yet be available or its quality (data quality) may not be acceptable. Therefore, the data must (additionally) be collected (data collection), possibly cleaned up, and integrated into an existing system (data integration). To ensure sustainable quality, the data source should also be evaluated [34, 47].

If the required quality and quantity of the data are given, a model is to be developed according to the problem definition of the business case, which supplies the demanded result through a method of data analysis. Often a software is used, which can turn out differently depending upon the question, availability, and existing competencies. Furthermore, the analytical skills of the analyst are required, which should be specified beforehand. This can also be helpful for the exact definition of a position [48].

Finally, the model quality measures the adaptability to the respective data for the evaluation of the model and describes its ability to apply these data to the business case. Depending on the method chosen, different quality criteria have to be applied [33].

A cost-benefit-analysis and added values should be highlighted for project prioritization, as they improve the possibilities of management and controlling. Due to result-oriented companies, many stakeholders are particularly interested in the resulting added value. The challenge is to present the right information to the right target group and thus avoid complexity. This is where Data Storytelling plays a significant role [42].

2.2. Expert Workshop

2.2.1. Procedure. Based on the results of the literature research, it becomes clear that most canvas concepts are more technically oriented, neglecting the cost and revenue structure or even the skills needed in the organization. In many cases, the derivation of the problem via a business case is missing. No single canvas provides additional support for the choice of methods and software, and even the completion of supporting questions is not always available.

Building on these results, we developed a model that can be used for any data analysis problem. Our canvas model should address non-data-affine managers and executives to make data science projects more plannable, controllable, and manageable. This allows companies to gain an overview of their resources and the utility of their data. The model should further help to evaluate the use of such a canvas in practice.

For the development of the first version of our Data Science Canvas, we used the design thinking process by Brown and Katz [49]. In the beginning, the problem space and the interest groups were defined. With the help of a workshop on data analysis with eight researchers and scientific experts from Germany from the fields of informatics, statistics, mathematics, economics, and business administration, it was examined how they proceed with the analysis of data and what problems they encounter. Researchers act at the transition point between science and practice by communicating research data and project results and, thus, can argue from both perspectives.

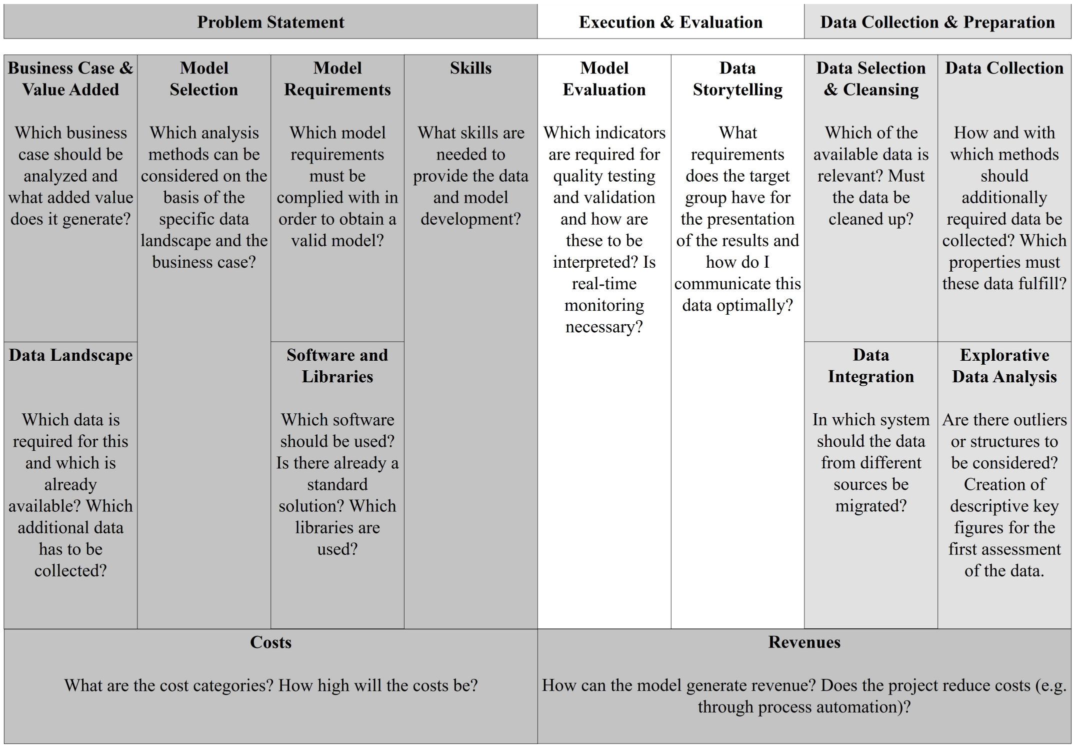
The insights gained were processed and served to define the viewpoints of the interest groups (managers, specialists, users). Further they were used in the idea generation phase to discuss the instantiation of a canvas through group dynamics and with discussion-enhancing ”How can we...?” questions. The elements were therefore placed on a whiteboard via sticky notes and their meaning and arrangement were critically discussed by the group.

The interim result of the workshop was an artifact, which was finally tested and evaluated by each participant based on a use case in the form of a project to optimize a sustainable energy supply of a production plant. One of the main findings was that selecting the appropriate method still encounters problems. In comparison to the identified results of the literature review, explorative data analysis and model requirements were here mentioned as additional elements. The insights gained in this process were incorporated into the actual version of our Data Science Canvas (see Figure 2).[2]

2.2.2. The Data Science Canvas. Consisting of three parts, our Data Science Canvas addresses the different aspects of data analysis: The dark grey elements are dedicated to the definition of the (business) problem (**Problem Statement**). The light grey fields deal with the collection and preparation of data for the model (**Data Collection and Preparation**). Finally, the white elements describe the implementation and evaluation of the previously derived method (**Execution and Evaluation**).

Users can access the Data Science Canvas basically from two different directions: On the one hand, it allows them to start by defining the business case via the data collection and preparation to derive and evaluate a model. On the other hand, it is possible to start from the data collection and preparation perspective (e.g. in the context of a research project) to derive an application case which leads to a model.

[2] Print version: https://github.com/tomalytics/datasciencecanvas



<table>
<tr><th colspan="4">Problem Statement</th><th colspan="2">Execution &amp; Evaluation</th><th colspan="2">Data Collection &amp; Preparation</th></tr>
<tr><td><b>Business Case &amp; Value Added</b>
Which business case should be analyzed and what added value does it generate?</td><td rowspan="2"><b>Model Selection</b>
Which analysis methods can be considered on the basis of the specific data landscape and the business case?</td><td><b>Model Requirements</b>
Which model requirements must be complied with in order to obtain a valid model?</td><td rowspan="2"><b>Skills</b>
What skills are needed to provide the data and model development?</td><td rowspan="2"><b>Model Evaluation</b>
Which indicators are required for quality testing and validation and how are these to be interpreted? Is real-time monitoring necessary?</td><td rowspan="2"><b>Data Storytelling</b>
What requirements does the target group have for the presentation of the results and how do I communicate this data optimally?</td><td><b>Data Selection &amp; Cleansing</b>
Which of the available data is relevant? Must the data be cleaned up?</td><td><b>Data Collection</b>
How and with which methods should additionally required data be collected? Which properties must these data fulfill?</td></tr>
<tr><td><b>Data Landscape</b>
Which data is required for this and which is already available? Which additional data has to be collected?</td><td><b>Software and Libraries</b>
Which software should be used? Is there already a standard solution? Which libraries are used?</td><td><b>Data Integration</b>
In which system should the data from different sources be migrated?</td><td><b>Explorative Data Analysis</b>
Are there outliers or structures to be considered? Creation of descriptive key figures for the first assessment of the data.</td></tr>
<tr><td colspan="4"><b>Costs</b>
What are the cost categories? How high will the costs be?</td><td colspan="4"><b>Revenues</b>
How can the model generate revenue? Does the project reduce costs (e.g. through process automation)?</td></tr>
</table>

Figure 2. Data Science Canvas.

Problem statement: Based on the business case and the expected added value, the already existing data landscape of the company is to be analyzed. The actual data situation shows whether additional (external) data is required under certain circumstances. The information about the existing data and the application case is used to select the specific model. If an analysis method was chosen, the requirements of the model (e.g. regarding scaling) should be noted. The software used is also influenced by the choice of model. Thus, several software tools are usually suitable, depending on the existing resources in the company as well as the skills to be defined subsequently. These are equally determined by the method used. For meaningful resource planning and project prioritization, the costs and revenues are also defined. In the case of values that are difficult to estimate, an enumeration of possible cost and revenue categories can also be helpful.

Data Collection and Preparation: If not all the data required for the model are available, additional data must be collected. This can be achieved using various methods and sources (e.g. surveys, open data, etc.). The requirements of the model, e.g. for the scale level, must be considered. Following their collection, the quality of the data must be checked. Data inconsistencies and outliers can be quickly identified using descriptive statistics and visualizations. Furthermore, meaningful aggregations of the data to key figures (e.g. mean values, scatter, and correlations measures) help to gain a deeper understanding of the circumstances to be modelled. If it is necessary to clean up the data (e.g. due to outliers or missing values), the required measures are documented (e.g. interpolation of missing values) and the data relevant to the model is selected. The split into training and test data can also be carried out at this point. If there are (no longer) any quality deficits, it should be clarified into which system the data possibly originating from different sources are to be integrated, so that the model can access them flawlessly. In principle, the tasks of data collection and preparation should be performed by a data engineer.

Execution and Evaluation: Finally, during the execution and evaluation phase, the model is evaluated using specific quality indicators that depend on the method. For some models it is also helpful if continuous monitoring is carried out in real time (e.g. in the case of predictive maintenance solutions). The remaining element can be used to present the stakeholders requirements for the presentation of the data, its further communication, and the use of the model. The element should support a target group oriented communication through data storytelling to create understanding, motivation, and change.

Table 3. Overview of Participants.

ID	Age, Gender	Branch	Job
F01	38, f	Media	Business Development
F02	42, m	Engineering	Consultant
F03	41, f	Automotive	Project Manager
F04	35, m	Chemical	Developer
F05	37, m	IT	Data Scientist
F06	48, m	Manufacturing	Procurement Manager
F07	40, m	Trade	CDO, Head of IT
F08	43, m	Project Sponsoring	Head of Software Development
F09	29, f	University	Researcher
F10	27, f	University	Researcher

3. Demonstration & Evaluation

3.1. Data Collection and Analysis

To conduct an ex-post naturalistic evaluation [20], we introduced the canvas to 10 managers and data science practitioners to test and evaluate it in their organizations (cf. Table 3). We then interviewed all of them along a semi-structured interview guideline, to understand and evaluate the canvas in terms of its suitability and usefulness for empowering managerial-decisions on DDBMs and effective stakeholder communications. In the case of the canvas artifact, a numbering of the elements was explicitly left out in the sense of a procedure to provide scope for own layouts and processes.

We have analyzed the data based on the thematic analysis approach, as described by Brown et al. [50]. We used MAXQDA to code the transcribed interviews and in an open coding procedure, the codes were developed and discussed in the team of researchers to derive the relevant themes. We ended this process after all researchers agreed on the low probability of finding additional themes. The results of the practical evaluation interviews can be divided into three categories: Use cases, application problems, and design implications. They are presented in the following chapter.

3.2. Findings

3.2.1. Use Cases of the Data Science Canvas. In general, almost all participants accepted the canvas positively and addressed various use cases. The canvas was often perceived as a structuring element, but also as a checklist to describe progress or the current situation. This is also discussed in the context of the respective target group. For example, a manager can use it to inform about a new project and as a checklist for **progress review**. Within the project team, the canvas can be used to structure the project during the planning phase and to point out problems that need to be solved.

"I find many of the points in the canvas very useful, and I will surely use them for my further work, especially with regard to the evaluation of the data situation and the data still needed, I find this extremely helpful and would pass it on to colleagues [...] and tell them, please: Work on these points." –[F08]

"Well, I think that's a good basis for argumentation to structure one's project. That you can think about how to handle it. [...] Or I could think of it as a checklist to check off during the project to see if it's there." –[F03]

It was also mentioned that the canvas can be used as an **instrument of trust** to make it clear to the manager that the previous preparation has been dealt with in a structured way.

"Such a canvas can also create trust with a supervisor by showing that you have dealt with it in a structured way, with your project. With my colleagues I would consider it more exploratory, for my boss it would be more like a check-up." –[F09]

"Another extreme value is that you can show a booth by showing that you can see: "I've just got that far and this and that needs to be done." –[F10]

There is also an application for cross-departmental projects to solve silo thinking through the structured and **interdisciplinary communication** of the topic or to identify the resulting problems.

"It also allows you to discuss with different actors across departments and to address overlapping problems and recognize them in the run-up to a project." –[F08]

A participant from the area of research also notes that she likes the fact that a problem can be approached from different directions. So on the one hand, if the data is available and something is to be investigated, and on the other hand, from a research perspective: I still want to collect the data and I can deal with the model requirements beforehand.

"I think it's good that it can also be filled out from a research perspective, that the problem can also be approached through data collection." –[F09]

Another use case is that the Canvas can help to define a specific **job description**.

"You can use this to have a better understanding of requirements for the definition of a position, what you are lacking in skills." –[F03]

Two participants also see data storytelling as an important factor to communicate results and progression to stakeholders.

"Data storytelling is then the customer's viewpoint, with which you can communicate the revenue again. It's important that it's communicated correctly." –[F05]

"I also like data storytelling, how it is prepared for the target group. It also helps to think in terms of goals, that you don't forget anything." –[F09]

3.2.2. Limitations of Applicability. During the interviews, however, problems regarding the use of such a canvas also became apparent. These mainly address interface problems with other departments, the scope of a canvas at the beginning of a project, and a general understandingof the elements.

Some participants pointed out that such a canvas is difficult to fill in at some points, especially at the **beginning of the planning phase**. The majority, however, emphasizes the added value addressed in the use cases. The Data Science Canvas provides a good overview of the implementation and weaknesses of a project, as well as of the resources still lacking.

"Well, I think that especially in the planning phase, where not much is known, that it was a bit difficult to fill in completely there." –[F06]

In this context, the problem of the **interdisciplinarity** of some projects is addressed. The canvas is also perceived positively here, to establish structure and overview. Especially in non-data-driven companies, there might be little understanding of cooperation in this area.

"A project manager working in the problem-space is not necessarily interested in data quality, at least not yet. But maybe in the future and the canvas could serve as an opener if everyone understands that data-driven business is important. But unfortunately, there are still too few of these people today. –[F08]

"Perhaps you should be able to specify which department the data comes from. Especially if you don't have a common database or silo thinking it might be important to have at least one contact person. I often had problems with that in practice." –[F05]

A further problem of use was seen in the **choice of methods**. Many in the project may have no experience with it and need supportin assessing the use of a specific model.

"The choice of a method is difficult for me. I don't know anything about these things. I can deal with the data, but I find it difficult to understand how I evaluate it. Perhaps recommendations could be made." –[F04]

It was generally suggested that there should be **more detailed guidance** on such a canvas model in addition to the supporting questions, also with a view to a common understanding.

"I always think it makes sense for the canvas models to come with a little manual or something, so that everyone understands the same thing." –[F01]

3.2.3. Further Improvements. The design of the Data Science Canvas was widely regarded as very extensive. On the one hand, this helped to ensure a structured and comprehensive approach to the problem, on the other hand, it required a more intensive study of the instrument. This also reflected the demand for **guidance for a common understanding**.

"[I think] the process is a bit difficult, it should at least be numbered in terms of phases." –[F06]

"So after I filled out the canvas, I noticed that I would arrange it differently, I would sort of design the structure from left to right, so first problem statement, then data collection and then execution. so it would have a clearer flow." –[F05]

It was also noted that a **differentiation into existing and still needed skills** would be useful. This could help to fill in the costs to be able to plan more precisely and also to create job profiles.

"One could perhaps still divide skills into skills that we already have and skills that we still need. This is certainly also relevant in practice. You can then apply that directly to the cost of what you would have to buy." –[F10]

It was also mentioned that a time component should be queried to better estimate the costs.

"You ask here already, must the data be cleaned up. Maybe. If I were to ask, how much longer or is it too much work?" –[F09]

4. Discussion & Implications

4.1. Practical Implications

The predominant contribution of this research for practice is the provision of a holistic Data Science Canvas that covers both technological and managerial aspects. Regarding the use of a Data Science Canvas model in practice, it becomes clear that it fits for several functions. On the one hand, it can be used as a structuring tool to discuss a project deeply by analyzing the current organizational situation and, thus, it is possible to identify problems that have to be solved beforehand. If the elements were filled out, the interviewees see also a use case regarding a communicationinstrumentto provide trust in the project team's work as well as to grant a commonunderstanding of the goals and challenges of the project.

On the other hand, it can also be used as a checklist for monitoringthe project's progress by getting visualized what is done and what is left, to be able to intervene if it is necessary. Furthermore, there is a possibility to start breaking up silo-mentality by addressing and discussing topics that are interesting for both stakeholders Nevertheless, there are still problems left that need to be considered in further developments of the Data Science Canvas to make it more relevant to practical applicability.

4.2. Theoretical Implications

From a theoretical perspective, our research contributes to literature through the following insights: First, in line with the practical contribution, our research provides an overview of yet existing data science canvas approaches and their respective elements. Furthermore, we combine those elements into a holistic canvas and discuss their relations.

An essential insight that contributes to design theory of data science canvases, is that such an artifact can be seen as both a structuring tool and as a checklist, depending on the phase of the project. This information can serve regarding a necessary adaption of the canvas structure by maybe offering different versions - one in the manner of a canvas and one as a checklist. If a canvas model is provided as software, this could simplify the filling out, because only the canvas needs to be prepared and the checklist is generated automatically.

For a common understanding, guidance can be helpful, which should be additionally available. To make the selection of a method more comprehensible for those who do not have an affinity for data, supporting small cards with the essential requirements and possibilities of the respective method could be offered. This could also improve interdisciplinary cooperation by giving everyone an equal understanding of the method. This also essentially supports the purpose as a structuring element of a project and shows where a participant still has to work on.

To improve the design, the canvas should further be numbered according to the elements or described accordingly in a manual. In our case, the numbering therefore should be presented differently depending on the initial situation in a manual, so that several applications are possible.

It is also an important finding that supporting questions time requirements should be addressed and queried to ensure a better presentation of the cost and benefit view and to support the decision for or against such a project. Furthermore, skills should be differentiated into those that are already available and those that are still needed in terms of personnel. This will also allow better planning of personnel resources. This can also help with job definitions, e.g. by making it clear if a data scientist or data engineer is required.

5. Conclusion

According to the Data Literacy Project study, data literacy is often poorly represented among decision-makers [4]. Regarding the central research question, a concept like the Data Science Canvas can therefore contribute to the success of a data science project through the clearly arranged presentation of relevant influencing factors and help to democratize Data Science and thus support digitization by making it accessible for non-data-affine project managers. According to the practical evaluation it serves as a communication base between the different project stakeholders and can be seen as a structuring element or checklist before, during, and after a project.

But the evaluation interviews also showed that there are some problems regarding the understanding, scope, and problems resulting from interdisciplinarity in projects. Therefore a Canvas concept should consider this information and solve the derived problems.

References

[1] Bitkom, “Digital-Design-Manifest,” 2018.

[2] T. Neifer, A. Schmidt, P. Bossauer, and A. Gadatsch, “Data Science Management: Planung, Steuerung und Kontrolle von Data Science,” *Rethinking Finance*, no. 3, 2019.

[3] K. Panetta, “Analyst Answers: The Biggest Challenges for Data & Analytics Leaders Today.”

[4] Qlik, “How to Drive Data Literacy With the Enterprise.”

[5] M. Comuzzi and A. Patel, “How organisations leverage big data: A maturity model,” *Industrial management & Data systems*, 2016.

[6] P. M. Hartmann, M. Zaki, N. Feldmann, and A. Neely, “Big data for big business? a taxonomy of data-driven business models used by start-up firms,” *Cambridge Service Alliance*, 2014.

[7] B. Kühne and T. Böhmann, “Requirements for representing data-driven business models-towards extending the business model canvas,” 2018.

[8] J. P. Womack, D. T. Jones, and D. Roos, *The machine that changed the world: The story of lean production–Toyota’s secret weapon in the global car wars that is now revolutionizing world industry*. Simon and Schuster, 2007.

[9] E. Ries, *The lean startup: How today’s entrepreneurs use continuous innovation to create radically successful businesses*. Currency, 2011.

[10] A. Osterwalder, Y. Pigneur, G. Bernarda, and A. Smith, *Value proposition design: How to create products and services customers want*. John Wiley & Sons, 2014.

[11] K. Mathis and F. Köbler, “Data-need fit–towards data-driven business model innovation,” in *Service Design Geographies. Proceedings of the ServDes. 2016 Conference*, no. 125, pp. 458–467, Linköping University Electronic Press, 2016.

[12] T. Kronsbein and R. Mueller, "Data thinking: A canvas for data-driven ideation workshops," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[13] B. Kühne and T. Böhmann, "Formative evaluation of data-driven business models–the data insight generator," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[14] R. Schüritz, S. Seebacher, and R. Dorner, "Capturing value from data: Revenue models for data-driven services," in *Proceedings of the 50th Hawaii International Conference on System Sciences*, 2017.

[15] A. Zolnowski, J. Anke, and J. Gudat, "Towards a cost-benefit-analysis of data-driven business models," 2017.

[16] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in informationsystems research," *MIS quarterly*, pp. 75–105, 2004.

[17] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.

[18] R. L. Baskerville, M. Kaul, and V. C. Storey, "Genres of inquiry in design-science research: Justification and evaluation of knowledge production," *Mis Quarterly*, vol. 39, no. 3, pp. 541–564, 2015.

[19] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS quarterly*, pp. 337–355, 2013.

[20] J. Venable, J. Pries-Heje, and R. Baskerville, "A comprehensive framework for evaluation in design science research," in *International Conference on Design Science Research in Information Systems*, pp. 423–438, Springer, 2012.

[21] J. Rowley and F. Slack, "Conducting a literature review," *Management research news*, 2004.

[22] F. Provost and T. Fawcett, "Data science and its relationship to big data and data-driven decision making," *Big data*, vol. 1, no. 1, pp. 51–59, 2013.

[23] Y. Levy and T. J. Ellis, "A systems approach to conduct an effective literature review in support of information systems research.," *Informing Science*, vol. 9, 2006.

[24] J. Webster and R. T. Watson, "Analyzing the past to prepare for the future: Writing a literature review," *MIS quarterly*, pp. xiii–xxiii, 2002.

[25] D. Ridley, *The literature review: A step-by-step guide for students*. Sage, 2012.

[26] P. Salipante, W. Notz, and J. Bigelow, "A matrix approach to literature reviews," *Research in organizational behavior*, vol. 4, pp. 321–348, 1982.

[27] M. Kaufmann, "Big data management canvas: a reference model for value creation from data," *Big Data and Cognitive Computing*, vol. 3, no. 1, p. 19, 2019.

[28] C. E. Perez, *The deep learning AI playbook: Strategy for disruptive artificial intelligence*. CreateSpace, 2017.

[29] A. Agrawal, J. Gans, and A. Goldfarb, "A simple tool to start making decisions with the help of ai," *Harvard Business Review*, 2018.

[30] J. Zawadzki, "Introducing the AI Project Canvas," Jan. 2019.

[31] L. Dorard, "Machine Learning Canvas."

[32] D. Kolkman and R. Sneep, "Challenges to data science projects with smes: An analysis and decision support tool," *Available at SSRN 3343092*, 2019.

[33] C. A. Mertler and R. V. Reinhart, *Advanced and multivariate statistical methods: Practical application and interpretation*. Taylor & Francis, 2016.

[34] K. Hildebrand, M. Gebauer, H. Hinrichsen, and M. Mielke, eds., *Daten- und Informationsqualitt: auf dem Weg zur Information Excellence*. Wiesbaden: Springer Vieweg, 4., berarbeitete und erweiterte auflage ed., 2018. OCLC: 1044698251.

[35] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[36] A. Savage, "Dear Companies: Your Data Science Job Descriptions are Awful," Apr. 2019.

[37] P. Zschech, V. Fleißner, N. Baumgärtel, and A. Hilbert, "Data science skills and enabling enterprise systems," *HMD Praxis der Wirtschaftsinformatik*, vol. 55, no. 1, pp. 163–181, 2018.

[38] J. Anderson, "Data engineers vs. data scientists," Apr. 2018.

[39] G. Piatetsky, "Python eats away at r: Top software for analytics, data science, machine learning in 2018: Trends and analysis," 2018.

[40] A. Géron, *Praxiseinstieg Machine Learning mit Scikit-Learn und TensorFlow: Konzepte, Tools und Techniken für intelligente Systeme*. O'Reilly , 2018.

[41] B. Dykes, "Data storytelling: The essential data science skill everyone needs," *Forbes Magazine*, 2016.

[42] T. Neifer, D. Lawo, P. Bossauer, M. Esau, and A.-M. Jerofejev, "Data storytelling als kritischer erfolgsfaktor von data science," *HMD Praxis der Wirtschaftsinformatik*, pp. 1–14, 2020.

[43] D. Breuker, D. Pfeiffer, and J. Becker, "Reducing the variation in intra-and interorganizanional business process modeling-an empirical evaluation.," in *Wirtschaftsinformatik (1)*, pp. 203–212, 2009.

[44] D. Cielen, A. Meysman, and M. Ali, *Introducing data science: big data, machine learning, and more, using Python tools*. Manning Publications Co., 2016.

[45] R. Wirth and J. Hipp, "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*, pp. 29–39, Springer-Verlag London, UK, 2000.

[46] A. Taschner, *Business Cases*. Wiesbaden: Springer Fachmedien Wiesbaden, 2017.

[47] O. Brazhnik and J. F. Jones, "Anatomy of data integration," *Journal of biomedical informatics*, vol. 40, no. 3, pp. 252–269, 2007.

[48] A. De Mauro, M. Greco, M. Grimaldi, and G. Nobili, "Beyond data scientists: a review of big data skills and job families," *Proceedings of IFKAD*, pp. 1844–1857, 2016.

[49] T. Brown and B. Katz, "Change by design," *Journal of product innovation management*, vol. 28, no. 3, pp. 381–383, 2011.

[50] V. Braun, V. Clarke, and P. Weate, "Using thematic analysis in sport and exercise research," *Routledge handbook of qualitative research in sport and exercise*, pp. 191–205, 2016.