Wirtschaftsinformatik 2022 Proceedings

Track 18: Design, Management & Impact of AI-based Systems

Jan 17th, 12:00 AM

Developing an ontology for data science projects to facilitate the design process of a canvas

Lukas-Walter Thiée
Leuphana University Lüneburg, Germany, lukas-walter.thiee@leuphana.de

Developing an ontology for data science projects to facilitate the design process of a canvas

Lukas-Walter Thiée[1]

[1] Leuphana Universität Lüneburg, Institute for Information Systems, Lüneburg, Germany
lukas-walter.thiee@leuphana.de

Abstract. Data science projects can become very complex, due to the complexity of their content, but also due to the nature and composition of their stakeholders. There are several approaches to remedy this, e.g., canvases, which support ideation and common understanding. However, previous approaches are limited to single details or abstract too much, so that it is difficult to carry out entire projects successfully based on them. This paper describes one part of the design process, namely the derivation of the underlying ontology, of a new canvas that integrates both the overall project and detail steps. The ontology is mainly derived from CRISP-DM, literature review and project work.

Keywords: Data science project, Machine learning, Canvas, Ontology

1 Introduction

With the advent of powerful local and cloud hardware, open source software, and even online collaboration tools for big data projects, e.g., Google Colab, data science (DS) should be within reach for all the interested. However, machine learning (ML) and artificial intelligence (AI) are still elusive topics for many companies and individuals. These information systems (IS) topics, nevertheless, are not only important drivers for the optimization of existing products and business processes, but also for the (digital) transformation and foundation of companies. So far-reaching, in fact, that data has been called a new commodity [1]. In order to take part in data science (DS), organizations start their own data-driven projects and due to the complexity of these projects they "need clear and structured guidance at the beginning of the innovation process to formulate and communicate business ideas with data" [2]. Therefore, standard processes have been suggested [3], maturity models have been developed [4], and competency profiles have also been proposed [5]. Yet, there is no generally accepted definition of DS and the discipline's practice. A major obstacle in the projects is the high spread of data literacy among the stakeholders, which hinders the process of value generation [6]. One approach to make the entry and execution of complex projects easier are canvases, such as the Business Model Canvas [7]. Canvases are boundary objects to facilitate teamwork and generate common understanding of complex topics. For the overarching field of DS, several canvas approaches have been proposed in recent years, for example the 'Machine Learning Canvas' [8], the 'Key Activity Canvas' [9], or the 'Data Value

17th International Conference on Wirtschaftsinformatik,
February 2022, Nürnberg, Germany

Map’ [10]. These tools offer support in various phases of a DS project. Most approaches focus on identifying potential before project start. However, it remains unclear how to get from this ideation to concrete tasks in the course of the project. The challenge is that most approaches are structured in such a way that they either address a specific part, e.g., ideation, thus ignoring the overall project, or they look at the overall project very generically, e.g., project justification and budgeting, so that the necessary level of detail is not achieved. Accordingly, an approach that integrates the standard process, i.e., CRISP-DM [3], is missing. Therefore, the question underlying this research project is “*What does a joint working tool, i.e., canvas, need to look like that supports teams during initial and subsequent tasks of ML projects in line with standard data science processes?*” The goal is to design a canvas for DS projects, especially suitable for small organizations. The contribution of this research project is in combining both an integrative view of the overall project and an appropriate level of detail in the individual sections, thus addressing the dichotomy between holistic and compact. This paper presents the derivation of the design requirements and the development of the underlying ontology to facilitate the design process of a canvas.

2 Related Work and Methodology

The approach of using a canvas to facilitate the development of ML or AI solutions has been prominent in IS research in recent years. At its core, most contributions try to support (parts of) the process from data exploration and ideation to a concrete business value. Four categories of such canvases with different thematic foci, namely *ML/DS*, *(AI) Project*, *Data Value*, and *Data Source,* have been identified in prior work [11], as shown in Table 1.

Table 1. Canvas artifacts with different foci

Focus	Year	Source	Canvas Artifact
ML/DS	2018	[12]	The ML Canvas (Big Data MBA Version)
ML/DS	2018	[12]	Hypothesis Development Canvas v1.1
ML/DS	2019	[8]	Machine Learning Canvas v0.4
ML/DS	2020	[9]	Key Activity Canvas
ML/DS	2020	[13]	ML Lifecycle Canvas
(AI) Project	2017	[14]	Digitalization Canvas
(AI) Project	2018	[15]	AI Canvas
(AI) Project	2018	[16]	AI Canvas
(AI) Project	2019	[17]	AI Project Canvas
(AI) Project	2020	[18]	AI performance canvas (prototype)
(AI) Project	2020	[19]	Canvas for the use of AI
(AI) Project	2021	[20]	Enterprise AI Canvas
Data Value	2016	[21]	Data Canvas: Data-Need Fit
Data Value	2017	[10]	Data Value Map
Data Value	2019	[6]	Data Innovation Board
Data Value	2020	[2]	Data Product Canvas
Data Source	2019	[22]	Data Collection Map
Data Source	2019	[23]	Data Insight Generator

All of these practice and scientific artifacts are intended as initial tools for generating ideas and/or as a communication platform at the beginning of DS projects. Although many of the approaches include parts of standard processes, such as the 'Data Preparation' phase, there is no explicit alignment between canvas and project progress, i.e., subsequent tasks after project initiation.

The literature review preceding this work posed the research question '*Which canvas models, that address ML or AI implementation, are available, and which contents do they cover?*' The answering of the latter resulted in a catalog of 163 fields with 287 (non-exclusive) questions categorized in a total of 11 categories and 39 subgroups. On the one hand this catalog with its categories is a good starting point for DS projects, on the other hand it also shows that there are still areas that are underrepresented, such as the connection between data preparation and modeling. Thus, the review is not only suitable as a basis for the following design part of an own canvas, but also shows where a new content focus must be set.

The methodological approach of this research follows the design-science paradigm, which at its core seeks to create useful (IS) artifacts through creative problem-solving techniques, thereby enhancing the scientific corpus and practical utility [24]. Since design science is meant to solve an observed (organizational) problem [25], we formulate the problem statement in two parts: (1) *Existing canvases and process models are not aligned, which makes it hard for organizations to use them coherently* and (2) *existing canvases either only focus on parts of the whole project or lack a level of detail, when they take an abstract view on the whole project, which makes it difficult to get to successful solutions*. Wirth and Hipp have already addressed this dilemma between detailed (exhaustiveness) and generic (parsimony) process descriptions [3]. We therefore propose the design requirements for the artifact: The canvas should be exhaustive, in order to address the whole project, and provide the right level of detail in order to be useful, while at the same time the canvas should be kept as simple as possible to provide ease-of-use and common understanding. Optionally, as the canvas might be too complex for a paper based version and workshop, a digital tool could be the preferred solution [26], which would demand the inclusion of user-centered design.

Since various stakeholders, e.g., data scientists, managers, domain experts, and IT specialists, are usually involved in DS projects and the project itself deals with complex topics, two main issues, namely collaboration/communication and structuring of tasks, have to be managed. Avdiji et al. (2018) propose IS "design principles for tools that both support collaboration and are tailored for specific ill-structured problems" [26]. These principles include "(1) framing the ill-structured problem by developing an ontology, (2) representing the ontology into a shared visualization, and (3) instantiating the visualization in a way that supports shared prototyping of the solution" [22]. We build upon these guidelines in the design process. In the following the derivation of the ontology is being described.

3 Developing the Ontology

The first step in the design of the future artifact, is the development of an ontology. In computer science an ontology is a representation of entities and the relationship between these entities in a specific subject area. It can be understood as reference model [27] and helps reducing "conceptual and lexical confusion by providing a unifying framework within an organization" [26], thereby sharpening problem understanding.

The field of DS encompasses a wide range of disciplines and skills, e.g. computer science, programming, statistics, or data management. In order to make this complex term more tangible, attempts have already been made to develop an ontology for DS, e.g., the Data Science Ontology [28]. This ontology indexes various concepts from the DS discipline as well as annotations of commonly used software libraries. However, the relationship between the elements is not evident from the index. The context and usefulness of the integrated software libraries can only be understood with appropriate prior knowledge. "Its long-term objective is to improve the efficiency and transparency of collaborative, data-driven science."[1] However, this publicly available ontology does not lay claim to completeness, rather it is a living and editable online document. We must therefore use other references.

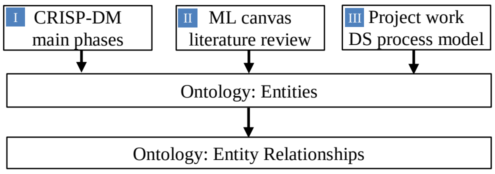
I CRISP-DM main phases
II ML canvas literature review
III Project work DS process model
Ontology: Entities
Ontology: Entity Relationships

Figure 1. Process and references for the development of the ontology

For this reason, we integrate three main sources to initialize the entities in our ontology (Figure 1). We use selected entities from these sources and then build the relationships between these entities. First of all, we utilize the CRISP-DM process stages (I) [3], as they are fundamental to DS. These stages include the entities 'Business Understanding', 'Data Understanding', 'Data Preparation', 'Modeling', 'Evaluation', and 'Deployment' Then we integrate two sources from prior work, namely the results of the aforementioned literature review regarding ML canvases (II) [11], and an item list of topics and questions, which is a part of project work regarding the development of a DS process model (III) [29]. Our approach ensures, that on the one hand the ontology is built on a tested scientific artifact, as CRISP-DM can be seen as a fundamental basis for DS projects, and on the other hand, both a content focus on DS and ML as well as actuality are taken into account. Additionally, we consider enhancements of the standard process, such as CRISP-ML(Q), which integrates quality assurance in ML projects [30].

Exemplarily, we describe the main path of the ontology (Figure 2), which was taken from CRISP-DM (blue): 'Business Understanding' enables 'Data Understanding', 'Data Understanding' in turn supports 'Business Understanding' and is simultaneously the basis for 'Data Preparation'. 'Modeling' requires 'Data Preparation' and is assessed

[1] https://www.datascienceontology.org/about

by the ‘Evaluation’, which in turn influences the ‘Deployment’, and recursively updates the ‘Business Understanding’. The ontology is structured in such a way that the main entities are composed of sub-elements, e.g. (shaded), ‘Business Understanding’ is composed of ‘Business Key Performance Indicators’ (KPI), which in turn are derived from ‘Customer‘, ‘Financial‘, ‘Product‘, ‘Organizational‘, and ‘Technological Understanding’. The interlinking of the sub-elements results in a web structure, which reflects the iterative nature of DS projects, and CRISP-DM respectively. Another central aspect of DS projects is captured in the ontology, namely the paths between ‘Data Quality’ and ‘Model Training’ (red). Extent and kind of the entire data preparation is substantially dependent on the selection and programming of an appropriate ML/AI algorithm, vice versa. The estimator selection in turn affects the model training and tuning. Therefore, depending on project maturity, different entities have to be incorporated. The ontology in Figure 2 represents an interim result and contribution of our research. It provides a holistic overview of a DS project.

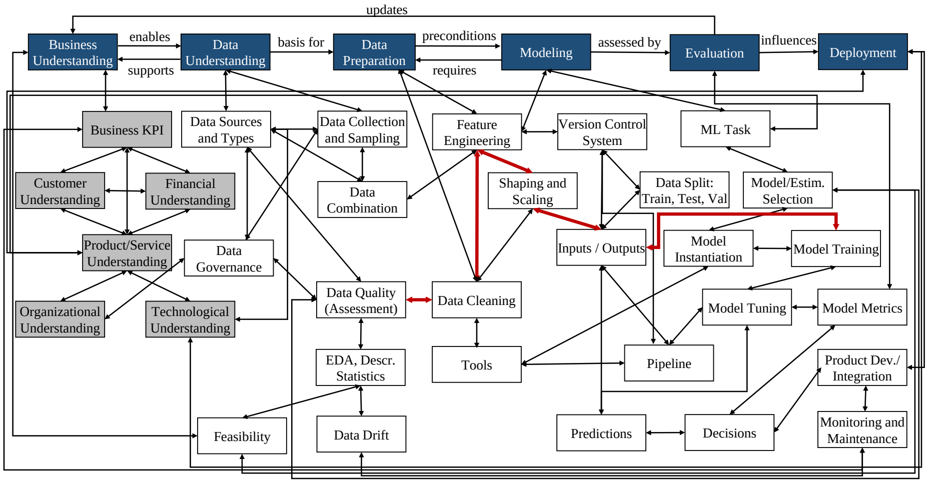

Figure 2. Simplified view of the ontology of a data science project

4 Conclusion and Future Research

In this article we call for the design of a new canvas for DS projects that is both holistic and compact, as previous approaches either address only partial aspects or are too generic. The artifact should support (small) organizations not only in generating ideas, but also in supporting the overall project. The design of the canvas is to be done in three steps, defining the ontology, designing a shared visualization and initializing the canvas. In this paper, the derivation of the ontology is presented as an intermediate result and contribution of our research. Future research consequently involves finalizing the design process, integrating principles of user-centered design [31], and evaluating the artifact. The evaluation is to be done essentially through qualitative methods, e.g., workshops and case studies, as is common in design science projects, and refers mainly to plausibility, usability, and perceived usefulness of the artifact.

References

1. The Economist: The world’s most valuable resource is no longer oil, but data. https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data. Accessed 30.08.2021 (2021)
2. Fruhwirth, M., Breitfuss, G., Pammer-Schindler, V.: The Data Product Canvas - A Visual Collaborative Tool for Designing Data-Driven Business Models. BLED 2020 Proceedings (2020)
3. Wirth, R., Hipp, J.: CRISP-DM: Towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining (2000)
4. Alsheiabni, S., Cheung, Y., Messom, C.: Towards An Artificial Intelligence Maturity Model: From Science Fiction To Business Facts. PACIS 2019 Proceedings (2019)
5. Provost, F., Fawcett, T.: Data Science and its Relationship to Big Data and Data-Driven Decision Making. Big data, 1(1):51–59 (2013)
6. Kronsbein, T., Mueller, R.: Data Thinking: A Canvas for Data-Driven Ideation Workshops. Hawaii International Conference on System Sciences 2019 (HICSS-52) (2019)
7. Osterwalder, A., Pigneur, Y.: Business model generation. A handbook for visionaries, game changers, and challengers. Wiley&Sons, New York (2010)
8. Dorard, L.: The Machine Learning Canvas. A handbook for innovators and visionary managers striving to design tomorrow's Machine Learning systems (2019)
9. Hunke, F., Seebacher, S., Thomsen, H.: Please Tell Me What to Do – Towards a Guided Orchestration of Key Activities in Data-Rich Service Systems. In: Hofmann, S., Müller, O., Rossi, M. (Hrsg.), *Designing for Digital Transformation. Co-Creating Services with Citizens and Industry. 15th International Conference on Design Science Research in Information Systems and Technology, DESRIST 2020, Kristiansand, Norway, December 2–4, 2020, Proceedings.* Springer International Publishing; Imprint: Springer, Cham (2020)
10. Sammon, D., Nagle, T.: The Data Value Map: A framework for developing shared understanding on data initiatives. ECIS 2017: 25th European Conference on Information Systems:1439–1452 (2017)
11. Thiée, L.-W.: A systematic literature review of machine learning canvases. Workshop: Künstliche Intelligenz für kleine und mittlere Unternehmen (KI-KMU 2021). INFORMATIK 2021. Lecture Notes in Informatics (LNI) - Proceedings. Series of the Gesellschaft für Informatik (GI) (in publication) (2021)
12. Schmarzo, B.: Data Science “Paint by the Numbers” with the Hypothesis Development Canvas. https://www.linkedin.com/pulse/data-science-paint-numbers-hypothesis-development-canvas-schmarzo/. Accessed 27.04.2021 (2018)
13. Zhou, Z., Sun, L., Zhang, Y., Liu, X., Gong, Q.: ML Lifecycle Canvas: Designing Machine Learning-Empowered UX with Material Lifecycle Thinking. Human–Computer Interaction, 35(5-6):362–386 (2020)
14. Heberle, A., Löwe, W., Gustafsson, A., Vorrei, Ö.: Digitalization Canvas - Towards Identifying Digitalization Use Cases and Projects. Journal of Universal Computer Science, 23(11):1070–1097 (2017)

15. Agrawal, A., Goldfarb, A., Gans, J.: A Simple Tool to Start Making Decisions with the Help of AI. https://hbr.org/2018/04/a-simple-tool-to-start-making-decisions-with-the-help-of-ai. Accessed 16.04.2021 (2018)
16. Dewalt, K., Rands, R.: Become an AI Company in 90 Days: The No-Bullshit Guide for Understanding AI, Identifying Opportunities, and Launching Your First Product. Prolego (2018)
17. Zawadzki, J.: Introducing the AI Project Canvas - Towards Data Science. https://towardsdatascience.com/introducing-the-ai-project-canvas-e88e29eb7024. Accessed 15.04.2021 (2019)
18. Engel, M., Lang, F.: A Pilot Study on Designing a Data & AI Performance Canvas. AMCIS 2020 Proceedings (2020)
19. Kreutzer, R. T., Sirrenberg, M.: AI Challenge - How Artificial Intelligence Can Be Anchored in a Company. In: Kreutzer, R. T., Sirrenberg, M. (Hrsg.), *Understanding Artificial Intelligence. Fundamentals, Use Cases and Methods for a Corporate AI Journey*. Springer International Publishing; Imprint: Springer, Cham (2020)
20. Kerzel, U.: Enterprise AI Canvas Integrating Artificial Intelligence into Business. Applied Artificial Intelligence, 35(1):1–12 (2021)
21. Mathis, K., Köbler, F.: Data-Need Fit – Towards Data-Driven Business Model Innovation. ServDes. 2016, Fifth Service Design and Innovation conference (2016)
22. Kayser, L., Mueller, R., Kronsbein, T.: Data Collection Map: A Canvas for Shared Data Awareness in Data-Driven Innovation Projects. Proceedings of the 2019 Pre-ICIS SIGDSA Symposium (2019)
23. Kühne, B., Böhmann, T.: Data-Driven Business Models - Building the Bridge between Data and Value. Twenty-Seventh European Conference on Information Systems (ECIS2019), Stockholm-Uppsala, Sweden. (2019)
24. Hevner, A., March, S. T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quarterly, 28(1):75 (2004)
25. Peffers, K., Tuunanen, T., Rothenberger, M. A., Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems, 24(3):45–77 (2007)
26. Avdiji, H., Elikan, D., Missonier, S., Pigneur, Y.: Designing Tools for Collectively Solving Ill-Structured Problems (2018)
27. Osterwalder, A.: The business model ontology a proposition in a design science approach. undefined (2004)
28. Chuprina, S., Alexandrov, V., Alexandrov, N.: Using Ontology Engineering Methods to Improve Computer Science and Data Science Skills. Procedia Computer Science, 80:1780–1790 (2016)
29. 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.: DASC-PM v1.0. Ein Vorgehensmodell für Data-Science-Projekte. valantic Business Analytics GmbH; Nordakademie gAG Hochschule der Wirtschaft; Universitäts- und Landesbibliothek Sachsen-Anhalt, Hamburg, Elmshorn, Halle (Saale) (2021)

30. Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., Mueller, K.-R.: Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology (2020)
31. Hussain, Z., Slany, W., Holzinger, A.: Investigating Agile User-Centered Design in Practice: A Grounded Theory Perspective. In: Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Holzinger, A., Miesenberger, K. (Hrsg.), *HCI and Usability for e-Inclusion.* Springer Berlin Heidelberg, Berlin, Heidelberg (2009)