The Human Side of Service Engineering, Vol. 62, 2022, 312–318
https://doi.org/10.54941/ahfe1002572

Processes in Data Science Projects

Damian Kutzias and Claudia Dukino

Fraunhofer Institute for Industrial Engineering IAO, Nobelstraße 12, 70569 Stuttgart, Germany

ABSTRACT

Data science and artificial intelligence have passed the stage of research in the ivory tower over the last years. Applications are not only found in huge enterprises and corporate groups: Many start-up companies were founded, and also small and medium sized enterprises adapt the new technology and take advantage of the capabilities more and more. For many of them, the use of data-based approaches rapidly become a necessity due to the product and service range of the competition or customer expectations. In particular, companies coming from other business sections than information technology face the challenge to implement new and robust data-based solutions. Classical structures and competencies have to be combined with new ones in data science projects, which usually come with high interdisciplinarity. Some aspects of such projects can be done just as in classical projects whereas others have to be slightly adapted and also some completely new arise. Data science process models can assist enterprises by facing these challenges with a structured approach, however most of them focus on the new or technical aspects of such projects or ignore the business context. This paper focuses on the aspect of business processes from data science projects in practice and shows their relevance in several points of time in and around a project's lifetime. Process-related differences to classical projects are shown and possibilities to take processes into account in an appropriate manner are discussed. Lastly, recommendations are given to cope with processes in the context of data science projects respecting the interplay of processes, humans and technology.

Keywords: Data science, Processes, Methodology, Engineering, Business processes

INTRODUCTION

Data Science as a trend term has several definitions and ambiguities. Addressing the industry and practitioners, we use an open definition aligned with the usage in practice: ≫Data Science is the nontrivial acquisition of knowledge from data≪ (Kutzias et al. 2021). Already in the beginning of the current century, data science was discussed as highly relevant for business and science. Depending on the definition, data science and its aspects were also addressed as data mining and knowledge discovery. It has also been noted early that it may be essential to go beyond algorithmic roots (Kurgan and Musilek 2006). It is a field of diversity and complexity, where different skills from different sub-disciplines come together (Egger and Yu 2022), great expertise is usually required. This expertise is often not covered by the data experts who carry out the project to the extent it would be necessary, as they know little about the field (Spruit et al. 2020). In addition, a lack of actual actions and tools was

noted (Fayyad et al. 2017). Data science process models such as Knowledge Discovery in Databases (KDD) (Fayyad et al. 1996) and the Cross-industry standard process for data mining (CRISP-DM) (Chapman et al. 2000) can assist in the implementation of data science projects, but usually provide structure and not concrete tools for each and every challenge. Widely accepted and known early data science process models such as KDD and the CRISP-DM are not easily applicable anymore due to missing considerations of the technical implementation, deployment and operational activities (Volk et al. 2020). Some of these aspects, especially related to the operation, include new or adapted business processes. Whereas this is nothing special to data science projects, not being covered in structured process models has the risk of addressing them too little or even not at all when relying on these models. Despite to the role of business processes as the basis for enterprises operations, they are rarely modelled explicitly to guide activities and instead they are often implicitly stored and managed (Delgado et al. 2020). There is an obvious risk in neglecting process management and optimisation: bad data-based processes are what they are at core – bad processes. Even if (business) processes are not highly relevant for each and every project, not considering them may often have effects ranging from unused potential to failed projects.

PROCESSES IN DATA SCIENCE PROJECTS

Organisations usually have hundreds or even thousands of business processes and their management is an important task for the organisations (Kriglstein and Rinderle-Ma 2012). Business processes and systems managing them are increasingly complex, including the integration of different versions, techniques and tools (Delgado et al. 2020). This section discusses the role of processes within the context of data science projects and differences to other types of projects. Relevant phases in and around a project's life cycle are highlighted, presented in an overview and accompanied with examples.

Business processes can be of great importance for project implementations from the very beginning: requirements analysis including business process modelling is acknowledged as a critical success factor of information system development for organisations (La Vara et al. 2008). Depending on the projects contents and results, processes can also play an important role afterwards when utilising the results. When business processes are affected by the project's results, the question of change and change management arises. The change of business processes is one important aspect of organisational change and can not only be implicit, but also guided by a structured approach, a change process (Hussain et al. 2018). Whereas neither the change of processes nor the change process itself are new or specific to data science projects, some facets are such as possible barriers for AI (artificial intelligence) implementation owing to negative attitudes among the employees (Lichtenthaler 2020). Such negative attitudes might be just unwillingness to change and adapt but can also be based on missing explainability of black box approaches or media forged pictures of strong AI. More general, a possible reason for resistance of employees is fear of coming changes (Vasiljeva et al. 2021). One step further to ≫just≪ being affected by the projects results which could also

be interpreted as targeted optimisation of a process, data-based analysis and optimisation can be the core of a project. Collecting large amounts of data does not necessarily lead to better processes and services (van der Aalst and Damiani 2015). Process mining is a structured data-based approach addressing this: it aims at extracting information from event logs to capture the business process as it is being executed (van der Aalst and Weijters 2004).

Data preparation is a fundamental stage of data analysis (Zhang et al. 2003). The preparation phase usually requires a lot of time of experts with knowledge about data schemas and structures as well as the domain (Tole and Joshi 2018). However, often the curation and cataloguing of the processes used to integrate and analyse the data are often neglected resulting in avoidable costs (Goble et al. 2008). These technical processes may cease after project finalisation, but their documentation can be of great value for possible further developments and sometimes also for maintenance or operation when exceptional problems or requests arise. Regarding maintenance and operation, another type of processes comes into play: monitoring and maintenance. Machine learning models are not static in quality and may require adaptions due to changes of the processed data (Wu et al. 2020). Such challenges can be addressed with MLOps (Machine Learning Operations), which enables developers to collaborate and increase the pace at which AI models can be developed, deployed, scaled, monitored and retained (Garg et al. 2022). The corresponding tasks of the operations also come with (new) business processes which we call secondary processes. Unfortunately, such secondary processes beginning with deploying models has been identified to be a black art often being ignored since the corresponding tasks are frequently beyond the capabilities of data scientists and the understanding of IT (Information Technology) teams (Fayyad et al. 2017). Hence, these processes deserve special attention especially when organisations implement their first few data science projects and do not have established standards in these areas.

Last but not least the data science project itself defines a process by the structure it follows during implementation. Data science process models such as CRISP-DM, KDD and several others can assist in following a structured approach and reducing risks. Although these models have proven well in assisting data science projects, most of them have known limitations such as missing continuity (e.g., focussing the technical core of a project neglecting downstream activities such as MLOps or change), vendor specialisation or limited concrete tool recommendations (Kutzias et al. 2021). Table 1 contains an overview of all previously discussed process types with a concise description as well as examples.

ADDRESSING PROCESSES IN DATA SCIENCE PROJECTS

Basically, handling processes can range from simply addressing them by informally taking them into account for certain tasks over modelling them in a formal way to making processes a central topic by the utilisation of process mining. Independent of the level of detail and formalisms of handling processes, the first step is the establishment of the awareness of relevant

Table 1. Overview of different types of processes in the context of data science projects with concise description and examples.

Type	Concise Description	Example
Primary Processes	Processes yielding requirements and processes being optimised by the project	A quality assurance process is to be optimised by utilising machine learning based classification on camera images.
Secondary Processes	Assisting processes such as those from monitoring and maintenance during operation	Specialists for quality assurance monitor a machine learning model on a regular basis.
Technical Processes	Processes of detailed technical (data processing) steps during development	Data is processed by several steps to ensure format and completeness before being used by a machine learning model.
Data Science Processes	The processes of implementing data science projects.	A data science project is implemented following project phases with milestones checking for goal adaptions also considering steps back.
Change Processes	Processes of applying change resulting from a projects outcome	A quality manager is retrained to monitor the quality of a machine learning solution and work in coordination with the model.

occurrences in the beginning of an upcoming project. Whereas processes specific to data science projects come with new challenges, tools and methods, it is not required to reinvent the wheel for documenting, analysing and optimising processes of data science projects. Regarding the five different process types discussed in the previous section and listed in Table 1, data science processes come with new challenges addressed by data science process models. CRISP-DM is known to be the de facto standard (Martínez-Plumed et al. 2020), but also has known issues which were not addressed by an update since its publication more than 20 years ago (Mariscal et al. 2010). Newer process models exist which can be utilised depending on the conditions and requirements of the project. We refer to (Kutzias et al. 2021) for a list and discussion of distinguishing characteristics which can assist in determining the best option for a data science process model. Whereas primary processes can be handled by classical tools during requirements analysis and later on during change, process mining is an option to make processes the core of a data science project. Process mining is a research field with a multitude of different software tools for application. An analysis and comparison of such tools would be a research topic itself, but we refer to a comprehensive list of such tools maintained by a university group: (Process and Data Science Group of the RWTH Aachen University).

For the other (types of) processes, classical approaches for documenting, analysing and optimising them can be utilised. The difficulty for them lies in knowing the subtle differences which especially occur for the secondary

processes of the operations as described in Section Processes in Data Science Projects. Two prominent options are the Unified Modeling Language (UML) which is considered a general purpose language but can also be extended depending on the requirements (Lindemann et al. 2002) and the Business Process Model and Notation (BPMN) designed for business users handling processes (White 2004). Lastly, change processes may have subtle but relevant different challenges such as AI-based negative attitudes as noted within the last section. However, they can be handled by known methods such as Lewins model (unfreeze, change, refreeze) or an adaption such as discussed in (Hussain et al. 2018).

For the selection of tools and methods for process handling, we recognised a gap between theory and practice. Theory and especially research tends to describe the theoretical optimum, which often fails in practice due to knowledge or resource restrictions. Such lack of relevance of research for practitioners is known as the ≫ivory divide≪ (Fuetsch and Suess-Reyes 2017). On the other hand, not handling processes or only handling them implicitly can result in unused potentials, staggered downstream costs or project fails. It is therefore a very important (and most probably non-trivial) task of organisations to decide for an appropriate level of detail and formalism of handling processes within their data science projects.

CONCLUSION

Within this paper, processes were discussed within the context of data science projects. Relevant occurrences of different types of processes in and around a project's life cycle were identified and discussed. The process types are 1) primary processes for requirements analysis and process optimisation, 2) secondary processes assisting data-based solutions, 3) technical processes containing technical (data processing) steps, 4) data science processes describing the project process of data science projects itself and 5) change processes describing the way of bringing project results to practice within the business. Three major differences were shown: 1) data science projects can differ from other types of projects, 2) primary processes handled by process mining are a research field themselves and 3) data-based solutions usually require more operational secondary processes. The other processes usually only have subtle differences. Lastly, possibilities of handling processes in data science projects were discussed, whereat data science process models for data science project processes and data mining tools make major differences to process handling in other project types. Furthermore, many classical methods and tools such as UML and BPMN remain useful tools for process handling such as in classical projects.

REFERENCES

Chapman, Pete/Clinton, Julian/Kerber, Randy/Khabaza, Thomas/Reinartz, Thomas/Shearer, Colin/Wirth, Rüdiger (2000). CRISP-DM 1.0. Step-by-step data mining guide.

Delgado, Andrea/Marotta, Adriana/González, Laura/Tansini, Libertad/Calegari, Daniel (2020). Towards a Data Science Framework Integrating Process and Data Mining for Organizational Improvement. In: Proceedings of the 15th International Conference on Software Technologies, Lieusaint - Paris, France. SCITEPRESS - Science and Technology Publications, 492–500.

Egger, Roman/Yu, Joanne (2022). Data Science and Interdisciplinarity. In: Roman Egger (Ed.). Applied Data Science in Tourism. Cham, Springer International Publishing, 35–49.

Fayyad, Usama M./Candel, Arno/La Ariño de Rubia, Eduardo/Pafka, Szilárd/Chong, Anthony/Lee, Jeong-Yoon (2017). Benchmarks and Process Management in Data Science. In: Stan Matwin/Shipeng Yu/Faisal Farooq (Eds.). Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax NS Canada. New York, NY, USA, ACM, 31–32.

Fayyad, Usama/Piatetsky-Shapiro, Gregory/Smyth, Padhraic (1996). The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 39 (11), 27–34. https://doi.org/10.1145/240455.240464.

Fuetsch, Elena/Suess-Reyes, Julia (2017). Research on innovation in family businesses: are we building an ivory tower? Journal of Family Business Management 7 (1), 44–92. https://doi.org/10.1108/JFBM-02-2016-0003.

Garg, Satvik/Pundir, Pradyumn/Rathee, Geetanjali/Gupta, P. K./Garg, Somya/Ahlawat, Saransh (2022). On Continuous Integration / Continuous Delivery for Automated Deployment of Machine Learning Models using MLOps. Available online at http://arxiv.org/pdf/2202.03541v1.

Goble, Carole/Stevens, Robert/Hull, Duncan/Wolstencroft, Katy/Lopez, Rodrigo (2008). Data curation + process curation=data integration + science. Briefings in bioinformatics 9 (6), 506–517. https://doi.org/10.1093/bib/bbn034.

Hussain, Syed Talib/Lei, Shen/Akram, Tayyaba/Haider, Muhammad Jamal/Hussain, Syed Hadi/Ali, Muhammad (2018). Kurt Lewin's change model: A critical review of the role of leadership and employee involvement in organizational change. Journal of Innovation & Knowledge 3 (3), 123–127. https://doi.org/10.1016/j.jik.2016.07.002.

Kriglstein, Simone/Rinderle-Ma, Stefanie (2012). Change Visualizations in Business Processes. Requirements Analysis. Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications 1, 584–593. https://doi.org/10.5220/0003815505840593.

Kurgan, Lukasz/Musilek, Petr (2006). A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review 21 (1), 1–24. https://doi.org/10.1017/S0269888906000737.

Kutzias, Damian/Dukino, Claudia/Kett, Holger (2021). Towards a Continuous Process Model for Data Science Projects. In: Christine Leitner/Walter Ganz/Debra Satterfield et al. (Eds.). Advances in the Human Side of Service Engineering. Cham, Springer International Publishing, 204–210.

La Vara, Jose Luis de/Sánchez, Juan/Pastor, Óscar (2008). Business Process Modelling and Purpose Analysis for Requirements Analysis of Information Systems. Advanced Information Systems Engineering, 213–227.

Lichtenthaler, Ulrich (2020). Extremes of acceptance: employee attitudes toward artificial intelligence. Journal of Business Strategy 41 (5), 39–45. https://doi.org/10.1108/JBS-12-2018-0204.

Lindemann, Christoph/Thümmler, Axel/Klemm, Alexander/Lohmann, Marco/Waldhorst, Oliver P. (2002). Performance Analysis of Time-enhanced UML Diagrams Based on Stochastic Processes. Proceedings of the 3rd international workshop on Software and performance, 25–34. https://doi.org/10.1145/584369.584375.

Mariscal, Gonzalo/Marbán, Óscar/Fernández, Covadonga (2010). A survey of data mining and knowledge discovery process models and methodologies. The Knowledge Engineering Review 25 (2), 137–166. https://doi.org/10.1017/S0269888910000032.

Martínez-Plumed, Fernando/Contreras-Ochando, Lidia/Ferri, Cesar/Hernandez Orallo, Jose/Kull, Meelis/Lachiche, Nicolas/Ramirez Quintana, Maria Jose/Flach, Peter A. (2020). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 1. https://doi.org/10.1109/TKDE.2019.2962680.

Process and Data Science Group of the RWTH Aachen University. Process Mining Software List. Available online at http://processmining.org/software.html.

Spruit, Marco/Dedding, Thomas/Vijlbrief, Daniel (2020). Self-service Data Science for Healthcare Professionals: A Data Preparation Approach. In: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies, Valletta, Malta. SCITEPRESS - Science and Technology Publications, 724–734.

Tole, Dipali/Joshi, Nikhil (2018). Simplifying Data Preparation for Analysis using an Ontology for Machine Data. In: Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 10th International Conference on Knowledge Engineering and Ontology Development, Seville, Spain. SCITEPRESS - Science and Technology Publications, 167–174.

van der Aalst, W.M.P./Weijters, A.J.M.M. (2004). Process mining: a research agenda. Computers in Industry 53 (3), 231–244. https://doi.org/10.1016/j.compind.2003.10.001.

van der Aalst, Wil/Damiani, Ernesto (2015). Processes Meet Big Data: Connecting Data Science with Process Science. IEEE Transactions on Services Computing 8 (6), 810–819. https://doi.org/10.1109/TSC.2015.2493732.

Vasiljeva, Tatjana/Kreituss, Ilmars/Lulle, Ilze (2021). Artificial Intelligence: The Attitude of the Public and Representatives of Various Industries. Journal of Risk and Financial Management 14 (8), 339. https://doi.org/10.3390/jrfm14080339.

Volk, Matthias/Staegemann, Daniel/Bosse, Sascha/Häusler, Robert/Turowski, Klaus (2020). Approaching the (Big) Data Science Engineering Process. In: Proceedings of the 5th International Conference on Internet of Things, Big Data and Security, Prague, Czech Republic. SCITEPRESS - Science and Technology Publications, 428–435.

White, Stephen A. (2004). Introduction to BPMN.

Wu, Yinjun/Dobriban, Edgar/Davidson, Susan B. (2020). DeltaGrad: Rapid retraining of machine learning models. Available online at http://arxiv.org/pdf/2006.14755v2.

Zhang, Shichao/Zhang, Chengqi/Yang, Qiang (2003). Data Preparation for Data Mining. Applied Artificial Intelligence 17, 375–381. https://doi.org/10.1080/08839510390219264.