The Human Side of Service Engineering, Vol. 108, 2023, 38–46
https://doi.org/10.54941/ahfe1003105

Bringing Data Science to Practice: From Protype to Utilisation

Damian Kutzias and Claudia Dukino

Fraunhofer Institute for Industrial Engineering IAO, Stuttgart, BW 70569, Germany

ABSTRACT

Data science and artificial intelligence have passed the stage of innovative trends. The applications in practice increase with every year with enterprises of all industry sectors creating new solutions utilising their data. However, there is much to learn for the enterprises, especially for those new to the implementation of information technology and data-based projects. Data science process models can assist in structuring such projects by giving ideal-typical project structures and assist with the provision of explanations, best practices, and concrete tools. One aspect which is rarely covered by data science process models is the utilisation of the results beyond their technical integration. This includes the risk of failing in operation due to missed requirements regarding affected employees or organisational aspects of the enterprises, especially their business processes. This paper provides an overview of relevant aspects for the integration of new data-based solutions into practice, i. e. the socio-technical system environment of the enterprise. Bridges to different project phases and results are shown to derive measures for integration. In addition, common tools for handling the arising challenges and tasks are listed and briefly discussed.

Keywords: Data science, Artificial intelligence, Process model, Methodology, Project management, Utilisation, Applied research, Change of qualification, Deployment and operation, Change of processes

INTRODUCTION

Data science process models as tools to support project management in data-based projects, especially including artificial intelligence (AI), are a rather new field of research. The Knowledge Discovery in Databases (KDD) Process (Fayyad et al. 1996) was identified as the initial approach and the CRoss-Industry Standard Process for Data Mining (CRISP-DM) (Chapman et al. 2000) has often been regarded as the central approach of evolution (Mariscal et al. 2010). CRISP-DM was often identified as the de facto standard in this area, even in recent years (Martínez-Plumed et al. 2020). Despite the existence of such guidelines, the proven value of the related technologies and the growing maturity of available technologies, it was stated that most projects fail (Volk et al. 2020). The existing guidelines, including explicitly listed KDD Process and CRISP-DM, were identified to be hard to apply and to only cover isolated aspects or at least not full projects from start to end (Volk et al. 2020). Even though CRISP-DM comes from practice and contains a step for utilising the results compared to KDD process, this step is not comprehensive

(Schulz et al. 2020). A data science process model omitting relevant project steps such as the utilisation of the results can thus be seen as a model assisting in prototype development instead of solution development for productivity. Even if the surrounding challenges often come from the business domain and do not always heavily differ from the challenges of classic projects, differences such as AI-specific requirements (e.g. explainability of the model) or human reservations regarding the "magic black box endangering their job" (Kutzias and Dukino 2022) can make the difference between a successful project and a failed one.

The coverage of all relevant project aspects was later identified as one of several desirable characteristics of data science process models and called "Continuity" (Kutzias et al. 2021). This paper focusses on the utilisation of the results as an important late phase of data-based projects, briefly describes the important contents of this phase, points out important dependencies to earlier project phases and finally discusses how to face related steps and challenges from the perspective of a project manager.

UTILISING THE RESULTS

In previous work, contents of data science projects were identified, and a gap analysis was performed with seven data science process models. The results include several relevant contents which are rarely covered by data science process models (Kutzias et al. 2023). The late project phase, the utilisation of the results, is most notably affected by this phenomenon. Usually, only the technical utilisation, namely the deployment, is covered by the analysed process models. The central contents of the phase are briefly described in the following.

"**Deployment and Operation**" consists of the technical tasks for utilising the data-based results from the given project. It encompasses all [technical] steps after model training and evaluation, including packaging the model in a format appropriate for deployment, publishing to a model registry or storage, integrating the model into a broader software system, serving, and monitoring (Kolltveit and Li 2022). The related efforts heavily depend on the projects environment since projects usually do not happen on a green field. The required (sub-)systems and their integration may be already given, be bought or be subject to extensive software development. Some even argue, that a data science project must be embedded in a software engineering process (Hesenius et al. 2019).

Depending on the expected efforts for deployment and operation, especially the integration, monitoring during operation and changes after the project, the degree of support and automatism may become an important aspect. To this end, machine learning operations (MLOps) describe approaches for this support. MLOps is still a vague term and its consequences for researchers and professionals are ambiguous (Kreuzberger et al. 2022). In simple terms, it means DevOps for machine learning, enabling developers to collaborate and increase the pace at which AI models can be developed, deployed, scaled, monitored, and retrained (Garg et al. 2021).

"**Change of Qualification**" consists of all kinds of changes related to the qualification of affected employees ranging from training over completely new roles to ceasing ones. While implicitly re-defining parts of the required job profiles of the enterprise, the change of qualification as part of a project is about reaching the required qualification in sufficient quantity for utilising the projects results. It is not necessarily restricted to the core of the results: New tasks and also roles (and therefore qualification requirements) can for example result from new or adapted secondary processes such as monitoring and maintenance from deployment and operations (Kutzias and Dukino 2022).

Particularly with AI in practice, employee learning plays an essential role for organisational development (Sen et al. 2022). A common consequence of AI in practice is the need for skilled workers with AI knowledge and competencies (Rott et al. 2022). This especially impacts the change of qualification when humans must collaborate with AI-solutions not hiding in the background. Another common consequence is the requirement of increased social skills when AI is solving a larger part of the analytical tasks (Huang et al. 2019).

"**Change of Processes**" consists of all kinds of process adaptions, ceasing of processes and establishing of new ones. In addition, the adaption of work activities as the smallest units of processes is included. Humans, as the proverbial creatures of habit, tend to repeat the same behaviours in recurring contexts (Wood and Rünger 2016). The resulting challenge is to effectively make employees change the way they work. In addition to this fundamental resistance to change, different challenges may arise such as resistance due to fear of coming changes (Vasiljeva et al. 2021), which may be particularly high for AI as the magic black box technology, potentially endangering their jobs. It is subject to the change of processes to identify such challenges in a timely manner and to overcome them before implementing the concrete changes.

Inputs From (Earlier) Project Phases

The previously described contents of the utilisation of the results contain several changes which are subject to planning in a timely manner. Thus, a project leader should think about these changes and their handling in the very beginning of a project. By such an early planning, several earlier steps within different phases of the project can be tailored to give useful input for the utilisation of the results. An overview of earlier phases with their contents can be found in (Kutzias et al. 2023). This section describes several important steps with input for the utilisation of the results. The details may and will vary for each project. The overview can be seen in Figure 1.

- The **Goal and Requirements** can directly yield important input for the utilisation phase. Requirements such as the existence of CI/CD (Continuous Integration and Continuous Deployment) for the resulting solution directly result in input for the deployment and operation. Processes themselves are often subject to change and their modelling is acknowledged as a critical success factor for information system development (La Vara et al.

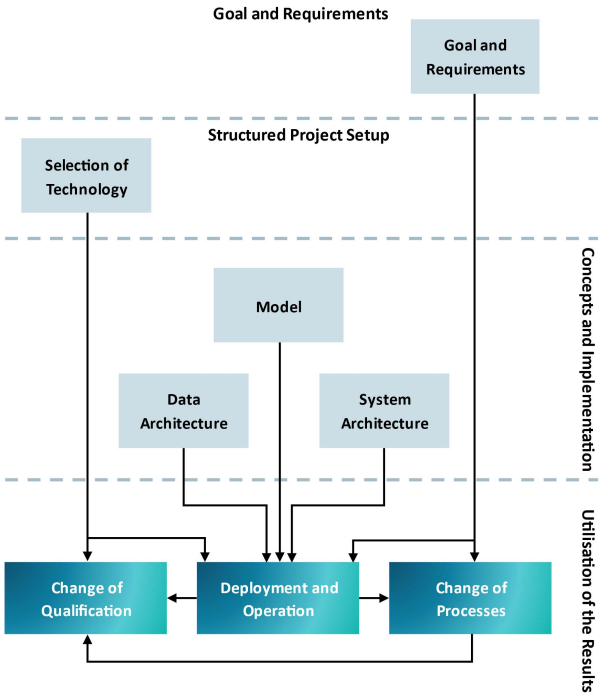

Figure 1: Visualisation of the central steps for utilising the results with relevant steps from earlier project phases giving important input.

2008). Modelled processes can be used for planning of changes as well as their transparent communication.

- The **Selection of Technology** directly affects the required future qualifications. Users must be capable of using the technology, may it be end users, developers, or maintainers. This selection can have negative impacts on the required resources when technological debts arise (Magnusson and Bygstad 2014) and should therefore be aligned with the strategy of the enterprise.
- The **System Architecture** describes the components of the final system as well as its interfaces. It can be hard to define the architecture in the beginning, therefore some changes may occur during the process (Kienzler 2019). In theory, implementation of the system architecture could be parallelised with the model development but doing so would include the risk of implementation and failing in the final evaluation which decides about exiting, refining, or bringing the solution into operation.
- The **Data Architecture** defines all relevant aspects of storing and accessing the data. Some parts of it are systems and therefore also part of the system architecture. In addition, more details such as database schemes and types are included. An architecture, which automatically integrates data and enables the usage of variable analysis tools, without the consideration about the specific data formats of the sources, can greatly enhance the impact of data analysis (Trunzer et al. 2017). It heavily influences the deployment and operation both for the initial deployment as well

as operational tasks such as handling new data or later improvement processes.
- The **Model** and its fundamental characteristics such as being an online model affect the necessary system environment by requiring interfaces for data provision. Depending on format requirements of the data and the integration of possibly relevant robustness aspects directly in the model, it may be necessary to implement pre- and postprocessing steps in addition to the model. Such implementations are subject to the deployment and operations.
- **Deployment and Operation** can itself require new secondary processes such as (semi-)automatic monitoring. Related processes are part of the change of processes in the same way as other business processes. Such new processes and the related technologies as well as the deployment and operations (e. g. MLOps technology) may require certain qualifications which are subject to the change of qualification within the project.
- The **Change of Processes** also affects the change of qualification by changing the way the employees work. This reflects the previously mentioned shift to higher social skill requirements as well as AI-related skills for human-AI-cooperation. Since such changes are rarely the direct goal or requirement of data-based projects (they may be in strategic projects, though), the input is presented indirectly by the change of processes.

Each of these steps includes one or several common inputs for the utilisation of the results. Some of them arise in the beginning of a project and can therefore directly be communicated and distributed while others can be planned as likely to happen, but arise later, e. g. during the implementation phase. A summary of these inputs including their phase of likely occurrence and examples can be found in Table 1.

Facing the Arising Tasks and Challenges

Deployment and Operation: Implementing DevOps concepts such as CI/CD for MLOps has challenging issues and exclusive tooling for the implementation is usually provided by cloud providers (Garg et al. 2021). Depending on experiences (not only in data-based projects, but also software development), the degree of automatism can be chosen which heavily impacts the complexity of MLOps systems. Despite the novelty of the topic, some guidelines and samples exist. John et al. describe the different degrees of automatism as their MLOps maturity model with four steps: 1) automated data collection, 2) automated model deployment, 3) semi-automated model monitoring, and 4) fully-automated model deployment (John et al. 2021). The authors also describe their MLOps framework as an abstract system architecture. Another such MLOps architecture is presented by Kreuzberger et al. and supplemented by an interview-based study which presents concrete tools for the different components of the architecture used by their interviewees in practice (Kreuzberger et al. 2022).

Change of Processes: There is a multitude of different definitions of "change" in the literature (Hrytsenko et al. 2021). Several definitions talk of change as a pure strategic aspect, consisting of serious changes on the

Table 1. Summary of the most common inputs for the utilisation of the results. The phases are numbered as follows: 1) goal and requirements, 2) structured project setup, 3) concepts and implementation, and 4) utilisation of the results.

Input	Phase	Example
Requirements	1	There must be a continuous monitoring of the results performed by the affected divisions.
Modelled processes (current vs. target)	1	A process for handling support tickets is changed by automatising manual matching to responsible divisions.
Technology Competence	2	Python programmers and analysts are required to continuously optimise the developed models in productivity.
Technology Restrictions	2	Monitoring must be possible in the enterprises' standard visualisation tool.
System Architecture	3	A device management and related interfaces must be implemented for maintenance in productivity.
Data Architecture	3	A new data mart must be implemented for a focussed access in addition to the integration of new sensor data into a data warehouse.
Model Requirements	3	Data must arrive in interpolated streamlines and needs to be post processed by applying a noise filter for higher model robustness.
New Secondary Processes	4	A model needs to be retrained with new market data in recurring intervals, requiring a continuous improvement process.
Change of Processes and Work	4	A new chatbot handles a huge share of standard questions, resulting in different skill requirements for a service team.

enterprise level or at least inter-divisional. Other definitions take smaller changes for processes and activities into account. In this work, the latter understanding of change is used. Nevertheless, both levels of change are highly related and can often be handled by the same or similar measures, since managing change is about managing people as the core activators of workplace performance (Moran and Brightman 2000). Motivation, including the elimination of the aforementioned possible resistances, is a prerequisite for successful change. In addition, to be able to mentally change the way of working, time and focus are required to bring the changes to effect. Classical change management approaches such as Kurt Lewin's change management model (consisting of three phases: 1) unfreezing, 2) change, and 3) refreezing) or adaptions such as discussed in (Hussain et al. 2018) can assist the project management in successfully managing change. An overview of different change management methods can be found in (Smith et al. 2022).

Change of Qualification: The change of qualification contains several classical aspects of project management not specific to data-based projects. An important factor is the planning and starting in a timely manner. For example, there may be major differences in the required lead time for training when a) doing it with the project team, b) doing it with a different division,

or c) outsource the training to another enterprise. Some of the new qualifications can directly be derived from the selected technologies. When, for example, Python is chosen as a programming language with a defined set of packages and libraries, these selections directly result in qualification requirements. More complex questions arise for new, data-related areas, especially MLOps. Kreuzberger et al. give a detailed overview of roles and competences required for this area complementing their architecture (Kreuzberger et al. 2022). The previously mentioned AI knowledge is another new competence field, bolstering end user performance in areas where humans and machines work in cooperation. Such human-machine symbiosis is advanced by the human understanding of the machine (Grigsby 2018), resulting in required AI knowledge.

CONCLUSION

This work analyses the often-omitted utilisation of results in data-based projects as a late phase of such projects making the difference between prototype development and solution development for operation. The key contents of that phase are outlined and discussed, which are "deployment and operation", "change of processes", and "change of qualification". For effective handling, different inputs from earlier project phases should be aligned with these late steps by the project management. These inputs and their origin are presented and discussed as important factors for successful project completion. Finally, different approaches and methodologies are briefly presented, outlining how to face the arising tasks and challenges. Many of the tasks and challenges are not specific to data-based projects, but several differences exist, and it is an important task for project managers to know and handle them for project success. An early planning, especially for clean preparations of inputs for the utilisation phase of a project, is identified being crucial for success.

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.

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.

Garg, Satvik/Pundir, Pradyumn/Rathee, Geetanjali/Gupta, P. K./Garg, Somya/Ahlawat, Saransh (2021). On Continuous Integration / Continuous Delivery for Automated Deployment of Machine Learning Models using MLOps. In: 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Laguna Hills, CA, USA. IEEE, 25–28.

Grigsby, Scott S. (2018). Artificial Intelligence for Advanced Human-Machine Symbiosis. In: Dylan D. Schmorrow/Cali M. Fidopiastis (Eds.). Augmented Cognition: Intelligent Technologies. Cham, Springer International Publishing, 255–266.

Hesenius, Marc/Schwenzfeier, Nils/Meyer, Ole/Koop, Wilhelm/Gruhn, Volker (2019). Towards a Software Engineering Process for Developing Data-Driven Applications. In: 2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE), 2019 IEEE/ACM 7th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering (RAISE), Montreal, QC, Canada. IEEE, 35–41.

Hrytsenko, P. V./Kovalenko, Y. V./Voronenko, V. I.Smakouz, A. M./Stepanenko, Y. S. (2021). Analysis of the Definition of "Change" as an Economic Category. Mechanism of Economic Regulation.

Huang, Ming-Hui/Rust, Roland/Maksimovic, Vojislav (2019). The Feeling Economy: Managing in the Next Generation of Artificial Intelligence (AI). California Management Review 61 (4), 43–65.

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.

John, Meenu Mary/Olsson, Helena Holmström/Bosch, Jan (2021). Towards MLOps: A Framework and Maturity Model. In: 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), Palermo, Italy. IEEE, 1–8.

Kienzler, Romeo (2019). The lightweight IBM Cloud Garage Method for data science. A process model to map individual technology components to the reference architecture. Available online at https://developer.ibm.com/articles/the-lightweight-ibm-cloud-garage-method-for-data-science/.

Kolltveit, Ask Berstad/Li, Jingyue (2022). Operationalizing Machine Learning Models — A Systematic Literature Review. Proceedings of the 1st Workshop on Software Engineering for Responsible AI, 1–8. https://doi.org/10.1145/3526073.3527584.

Kreuzberger, Dominik/Kühl, Niklas/Hirschl, Sebastian (2022). Machine Learning Operations (MLOps): Overview, Definition, and Architecture.

Kutzias, Damian/Dukino, Claudia (2022). Processes in Data Science Projects. In: The Human Side of Service Engineering, 13th International Conference on Applied Human Factors and Ergonomics (AHFE 2022). AHFE International.

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.

Kutzias, Damian/Dukino, Claudia/Kötter, Falko/Kett, Holger (2023). Comparative Analysis of Process Models for Data Science Projects. Proceedings of the 15th International Conference on Agents and Artificial Intelligence.

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.

Magnusson, Johan/Bygstad, Bendik (2014). Technology Debt: Toward a New Theory of Technology Heritage. Proceedings of the 22nd European Conference on Information Systems.

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, 3048–3061. https://doi.org/10.1109/TKDE.2019.2962680.

Moran, John W./Brightman, Baird K. (2000). Leading organizational change. Journal of Workplace Learning, 66–74.

Rott, Karin Julia/Lao, Lena/Petridou, Efthymia/Schmidt-Hertha, Bernhard (2022). Needs and requirements for an additional AI qualification during dual vocational training: Results from studies of apprentices and teachers. Computers and Education: Artificial Intelligence 3. https://doi.org/10.1016/j.caeai.2022.100102.

Schulz, Michael/Neuhaus, Uwe/Kaufmann, Jens/Badura, Daniel/Kerzel, Ulrich/Welter, Felix/Prothmann, Maik/Kühnel, Stephan/Passlick, Jens/Rissler, Raphael/Badewitz, Wolfgang/Dann, David/Gröschel, Alexander/Kloker, Simon/Alekozai, Emal M./Felderer, Michael/Lanquillon, Carsten/Brauner, Dorothee/Gölzer, Philipp/Binder, Harald/Rhode, Heiko/Gehrke, Nick (2020). DASC-PM v1.0 - Ein Vorgehensmodell für Data-Science-Projekte.

Sen, Wang/Xiaomei, Zhu/Lin, Deng (2022). Impact of Job Demands on Employee Learning: The Moderating Role of Human–Machine Cooperation Relationship. Computational Intelligence and Neuroscience 2022. https://doi.org/10.1155/2022/7406716.

Smith, Tianqi G./Norasi, Hamid/Herbst, Kelly M./Kendrick, Michael L./Curry, Timothy B./Grantcharov, Teodor P./Palter, Vanessa N./Hallbeck, M. Susan/Cleary, Sean P. (2022). Creating a Practical Transformational Change Management Model for Novel Artificial Intelligence-Enabled Technology Implementation in the Operating Room. Mayo Clinic proceedings. Innovations, quality & outcomes 6 (6), 584–596. https://doi.org/10.1016/j.mayocpiqo.2022.09.004.

Trunzer, Emanuel/Kirchen, Iris/Folmer, Jens/Koltun, Gennadiy/Vogel-Heuser, Birgit (2017). A Flexible Architecture for Data Mining from Heterogeneous Data Sources in Automated Production Systems. In: 2017 IEEE International Conference on Industrial Technology (ICIT). 22–25 March 2017. Piscataway, NJ, IEEE.

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.

Wood, Wendy/Rünger, Dennis (2016). Psychology of Habit. Annual Review of Psychology 67, 289–314. https://doi.org/10.1146/annurev-psych-122414-033417.