Transdisciplinary competence development for role models in data-driven value creation

The Citizen Data Scientist in the Centre of Industrial Data Science Teams

Jochen Deuse, René Wöstmann, Lukas Schulte, Thorben Panusch, Josef Kimberger

Increasing digitalisation is fundamentally changing the understanding and possibilities of value creation as well as labour organisation. The systematic collection, storage and analysis of data is becoming a decisive competitive factor and is the basis for intelligent products, processes and production technology. This results in new competence requirements and roles in mechanical and plant engineering and in the manufacturing industry in general. Machine Learning (ML) in particular, as the basis of Artificial Intelligence (AI), poses great challenges for companies, as the demand for experts, so-called Data Scientists, significantly exceeds the offer and furthermore, these experts rarely have the required domain knowledge - the core competences of manufacturing companies. In this context, the new job description of the Citizen Data Scientist (CDS) as a link between the most important disciplines of information technology (IT), domain knowledge and data science enters the focus of attention (Idoine/Brethenoux 2019). The article presents a role model as a basis for team building and systematic development of ML competences in the manufacturing industry and combines the results of various research projects and industrial implementations. For this purpose, required ML competences of the future are derived in section 1 and transferred into a transdisciplinary role model in section 2. Section 3 addresses the exemplary practical application in an industrial use case, while section 4 gives an outlook on the possibilities of target-oriented competence development for the individual roles and actors.

1. Machine Learning Competences of the Future

Increasing digitalisation and the spread of information and communication technologies (ICT) enable the systematic collection and storage of data. The systematic data analysis becomes a decisive competitive advantage through its use for optimisation and decision-making processes (Eickelmann et al. 2015; Wölfl et al. 2019). In particular, the use of ML enables the discovery of unknown and non-trivial structures and relationships and results in new perspectives of providing knowledge (Dragicevic et al. 2020). As an overarching discipline between com-

DOI: https://doi.org/10.30844/wgab_2021_3

puter science and statistics, it empowers data-driven insights, decisions and automated data processing. However, data analysis becomes a decisive success factor for companies when combined with domain-specific knowledge (Deuse et al. 2014). As a result, there are competence requirements with regard to a profound methodological understanding in the field of ML as well as a distinct technical understanding of engineering problems in order to meet the challenges of industrial companies in a demand-oriented manner (Bauer et al. 2018). Therefore, interdisciplinary teams from the three disciplines of computer science, statistics and engineering have to be formed on the company side. However, a lack of resources - especially for small and medium enterprises (SME) - limits the formation of such teams (Rammer et al. 2010). In particular, these companies have a strong domain knowledge of their own processes, but not in implementing ML in production (Bertelsmann Stiftung 2018; Morik et al. 2010). Therefore, the CDS is gaining importance: this role, first introduced by Gartner, describes a domain expert who is capable of using ML in the data-driven decision-making process, but whose job function is outside statistics and computer science (Moore 2017). This role is becoming important as the availability of ML tools is increasing exponentially, but the rate of Data Scientists is not increasing at the same level (Miller/Hughes 2017; Mullarkey et al. 2019). Due to this, the potential of data analysis is not fully exploited, creating an economic disadvantage (Mazarov et al. 2019). In this context competences describe knowledge and skills that are acquired through teaching and learning processes (Fölsch 2010) and can be used by individuals in a targeted manner to successfully solve problems in different situations (Weinert 2001). Thus, in manufacturing industry, competences can only be developed through the concrete and problem-oriented application of knowledge and skills (North et al. 2016).

1.1. Derivation of general ML Competences within the Industry

In order to enable problem-oriented competence development, the relevant competences have to be identified. For this purpose, (Bauer et al. 2018) interviewed companies from various industrial sectors such as automotive, electronics and mechanical engineering regarding the degree of use of ML in industrial practice. Only 5 out of 57 companies (9 %) are successfully using ML in industrial practice, while 27 companies (47 %) are using ML to a small extent or piloting initial applications. The remaining 25 companies (44 %) plan to use ML in the future without using it yet. To identify the obstacles of integrating ML into industrial practice, the companies were surveyed in more detail. The main obstacle of using ML is the lack of methodological competences. Furthermore, the increasing complexity of manufacturing processes leads to an increase of required professional competences. Another major challenge is the transfer of the acquired theoretical knowledge to the practical issues and areas of application. In the survey, the companies indicate a high importance for the competence categories of ML, ICT and domain knowledge, but also social and personal competences. (Schulte et al. 2020).

Based on these findings (Bauer et al. 2018) derive an approach to enable ML in industrial production by developing the tripartite Industrial Data Science (InDaS) model (Fig 1) as the basis for a transdisciplinary course of the same name at TU Dortmund University as part of the research project InDaS

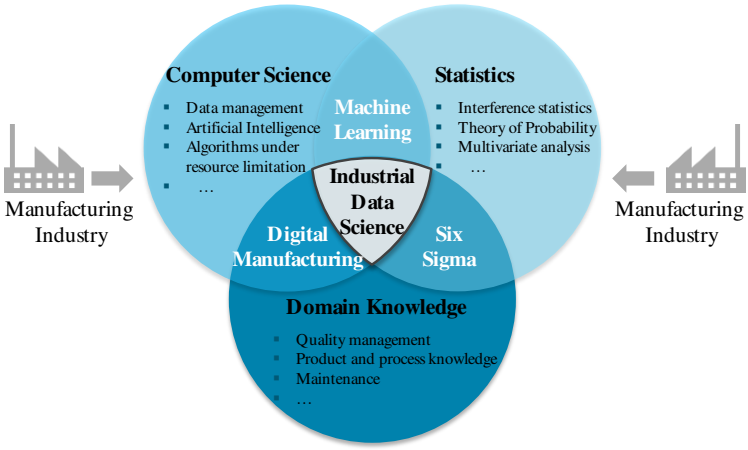

Figure 1: Industrial Data Science as a transdisciplinary approach (Bauer et al. 2018; Schulte et al. 2020)

1.2. Detailing transdisciplinary Categories and Competences

The InDaS model (Fig 1) addresses a transdisciplinary approach to solving challenges with ML in an industrial context. It provides a first suitable overview of disciplines and expertise to be integrated and addresses esp. professional competences. However, a deeper level of detail is required. Within the research project ML2KMU, (Reckelkamm/Deuse 2021) work out that in addition to the professional competences also methodological, social and self-competence play a decisive role. In the following, the individual categories are presented in more detail. Based on the InDaS model and the literature research listed by (Reckelkamm/Deuse 2021), domain knowledge and competences in the categories of ML, statistics, data management and ICT are emphasised as professional competences. The domain knowledge addresses e.g. knowledge about manufacturing processes, the ability to identify ML potentials, to assess framework conditions and possibilities for inference in deployment. The ML competences consist of knowledge of ML methods and models, programming languages and software tools, feature engineering and selection, the creation of visualisations, validation and performance metrics as well as operationalisation of models in deployment. The competences in the area of statistics address application oriented skills in descriptive statistics (e.g. metrics, visualisation forms), multivariate statistical methods, theory of probability and time series. The competences in data management include database

technologies, cloud solutions, IT systems and structures as well as mechanisms for privacy and security concerns in industry. The ICT competences address on the one hand the mastery of automation technology (e.g. human machine interfaces (HMI), programmable logical controllers (PLC) as well as field level interfaces and protocols), sensor technology, Industrial Ethernet, shop floor IT systems such as manufacturing execution systems (MES) and machine data acquisition (MDA), but also Edge Computing, which plays an increasingly important role in ML for distributed decentralised data collection and execution of models. Methodological competences include skills and application of Lean Management methods for conventional process optimisation, as well as knowledge of how to perform data-based improvements such as the DMAIC (Define Measure, Analyse, Improve, Control) process model of the Six Sigma philosophy. Knowledge of ML related models such as the CRISP-DM (Cross-Industry Standard Process for Data Mining) of KDD (Knowledge Discovery in Databases) is required to structure projects. In general, methodological skills also include presentation techniques, operational project management as well as special methods such as Design Thinking or agile frameworks like SCRUM, some of which, however, can also be seen as optional depending on the specific tasks and teams. The social competences address the ability to work in teams, communication skills, conflict management, leadership, cooperation and empathy. In addition, self-competence is required, which addresses characteristics such as the willingness to learn, adaptability, curiosity and openness as well as creativity. Figure 2 summarises the categories of competences. It becomes clear that ML is a transdisciplinary cooperation of different professional domains and actors and therefore addresses heterogeneous competences. Many of the existing approaches address only the professional competences, which, however, does not provide a complete picture, especially for application-driven data science, such as the rise of CDS. In the following section, specific role models will be presented in order to be able to assemble and develop teams in a targeted manner.

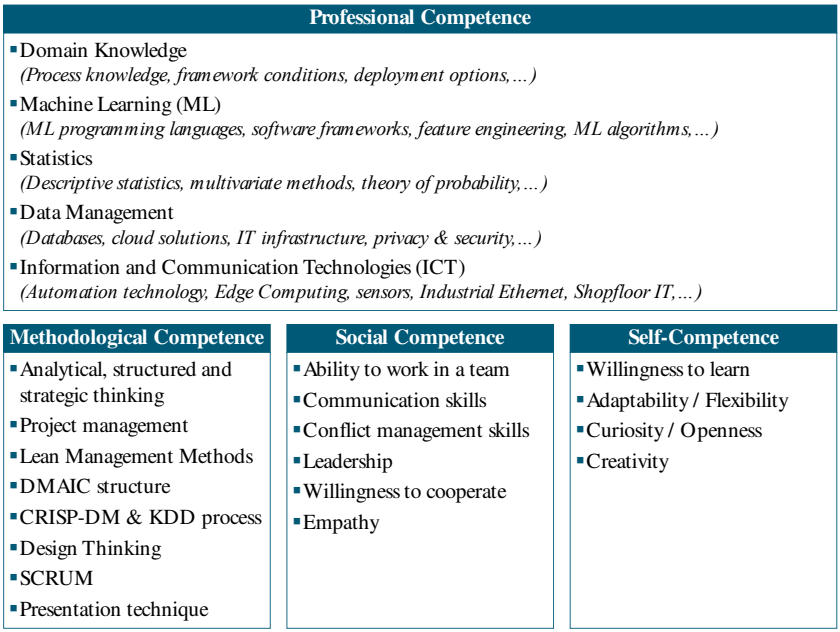

Figure 2: Categories of competences required for ML projects in manufacturing industry

2. The Citizen Data Scientist in the Centre of Industrial Data Science Teams

Although the competences identified are useful for a general overview of the disciplines to be represented in implementation projects or the organisation, for a detailed team composition not only the competences but also the roles and actors need to be specified in more detail. The literature contains both theoretical models for the general composition of ML teams, e.g. (Saltz/Grady, 2017) and practice-oriented guidelines, e.g. (RapidMiner 2020). However, previous approaches are either too specific in detail or too generically and only partially address the domain of the manufacturing industry. In addition, none of the existing work adequately addresses competence development in an integrated role model. Therefore within the research project DaPro (Data-driven process optimisation based on machine learning in the beverage industry) (Wöstmann et al. 2019), transdisciplinary role models were defined based on the preliminary work of the InDaS project and the DPDA (Data Preparation for Data Analytics) project group (Stark et al. 2019) of the prostep ivip association and a practical application context was created for validation. An overview of the role model is shown in Figure 3. It consists of the roles of IT, Domain Expert, Data Scientist and Management, whose particular areas of expertise are integrated by the CDS in a central orchestrating role.

The basis for both the delimitation of the roles and the design of the competence profiles is formed by iterative workshops both in the prostep Group DPDA and within the DaPro project with practitioners from manufacturing industry, mechanical engineering companies, Data Scientists, IT companies and research institutes. The disciplines of the competences were expressed with different competence levels. The levels consist of no competence present or required (0), basic information present (1), which becomes knowledge (2) through application and interconnection. This is extended through willingness and practical experience to competence (3). Furthermore, the highest level (4) addresses the ability to synthesise and evaluate different approaches and disciplines. The gradations are thus based on the knowledge steps according to (North et al. 2016) as well as the stages taxonomy model according to (Bloom/Engelhart 1976). The partially odd values of the competence levels in the following section result from the process of individually interviewed industry and research experts from the InDas, DaPro, AKKORD, DPDA and ML2KMU projects and subsequently combining the ratings in order to obtain a more valid overall view. For aggregation, the mean values of the assessments of the interviewed experts were used. In the following section, the individual roles, their specific tasks and responsibilities as well as the underlying competence profiles are presented.

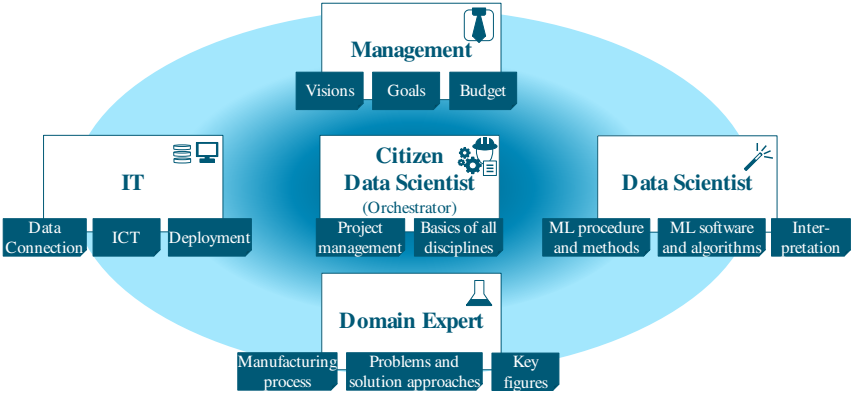

Figure 3: The Citizen Data Scientist in the Centre of Industrial Data Science Teams

2.1. Management

Fundamentally, Management has an important role to contribute. Without the active involvement, initiative or support of the management, there will be no sustainable success for any ML-driven initiative in manufacturing companies. The tasks consist of constituting a common vision and initiating projects. The composition of the project team and the provision of resources (e.g. budget and working time) play an important role. In addition, Management can help to open doors by

encouraging internal cross-departmental communication, but also by engaging external partners. During the project, Management takes on a more passive role, advising the project team as required, assisting in escalations and mediation of conflicts, making decisions at important points and being informed about the progress through reporting and controlling. Even if new agile approaches such as SCRUM are chosen for operational implementation, the Management must enable the organisational conditions for this. In addition, it is both strategically responsible for the integration of ML projects into the corporate strategy and operationally involved in their implementation. The requirements for the role of Management are also reflected in the competence profile (Figure 4). A high level of self- and methodological competence is required to equip the teams with the relevant tools and to inspire enthusiasm for general visions and specific ML-driven initiatives. In addition, social competence helps in assembling the right teams and resolving conflicts. Less necessary in Management are technical competences, which are covered by the other roles. Domain knowledge should be emphasised, as it is a fundamental factor in assessing the practical applicability and relevance of the analysis results.

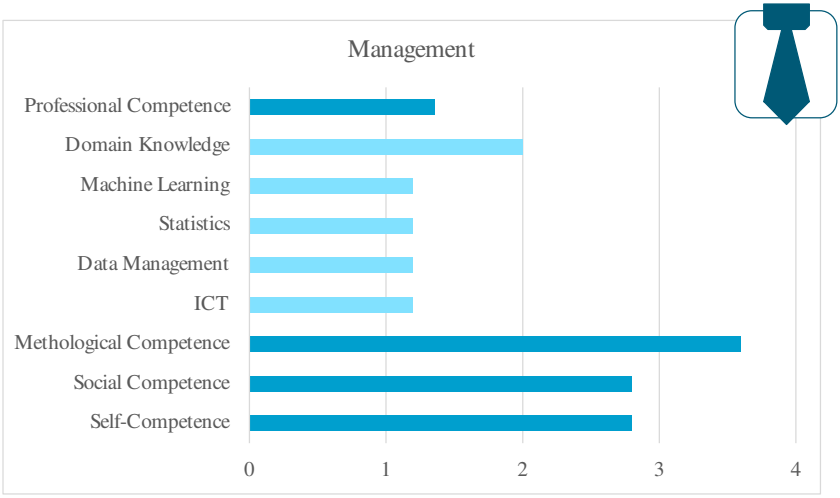

Figure 4: Competence profile of the Management role

2.2. IT

The role of IT is fundamental to the data-driven application of ML. Firstly, it provides insights into the possibilities and restrictions of productive IT systems regarding e.g. accessing data sources. Furthermore, it helps in the consideration of legal concerns such as privacy and security. The IT role should have a comprehensive overview of suitable ICT (e.g. database technologies, IoT protocols, edge devices and platforms). It plays a fundamental role in the implementation of architectures for ML. In smaller companies and one-off projects, this can be rather

simple, including hosting servers and making data access available, installing analytics software and providing deployment options. In more advanced projects with a higher degree of complexity, where, for example, larger amounts of data are processed and greater computational power is required, IT sets up and maintains platform solutions and distributed computing and storage services, or enables the use of commercial Platform-, Infrastructure- or Software-as-a-Service (PaaS, IaaS, SaaS) solutions. Furthermore, services and the operation of an own platform ecosystem are becoming increasingly important for machine manufacturers. The role of IT is therefore to be considered heterogeneous and can be partially differentiated into further sub-roles, e.g. data engineers, solution architects, or platform teams. The most important competence requirements (Figure 5) are specific technical competences like ICT skills and data management. A basic ML knowledge helps to understand the requirements for architectures both for testing and training, but also for deployment, and to be able to translate them into implementations. A basic knowledge of the application domain also contributes to this.

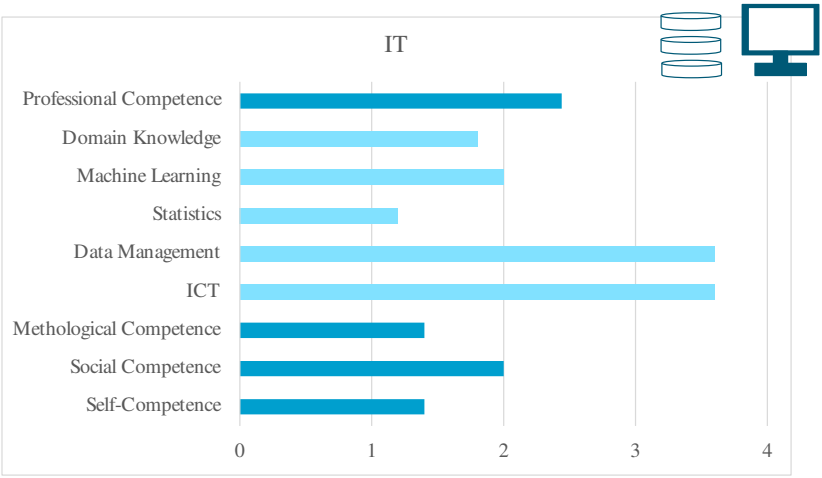

Figure 5: Competence profile of the IT role

2.3. Domain Expert

Domain Experts play a central role in ML implementation projects in the manufacturing industry, as they contribute the most profound knowledge about value adding manufacturing processes. They can represent the starting point of a project and identify needs in real application scenarios. In the interdisciplinary teams, they explain the problems to the other actors involved and play an important role in the brainstorming of solutions, since they are able to assess both the possibilities and the restrictions of the real processes in the most valid way. In this context,

their task is also to define requirements for the solutions to be developed. Furthermore, the definition of performance and result indicators plays an important role, which can be used to quantify both problems and improvements. Domain Experts know the IT systems from a user's perspective and can define the conditions for deployment options. They also help in evaluating and validating model outputs in terms of transferability and applicability to real processes. With regard to the competence profile (Figure 6), conditions of self, social and methodological competence are comparable to those in IT and indeed useful qualities for the daily work. In ML implementation projects, however, these do not necessarily have to be fully developed, as the project management tasks are performed by the orchestrating role. Therefore, professional competence is of highest importance. On the one hand, a basic ICT knowledge cannot cause any disadvantage in order to be able to comprehend the IT systems, underlying data as well as possible deployment scenarios. However, the key competence of Domain Experts is domain knowledge that includes process knowledge as well as general ML and deployment potentials.

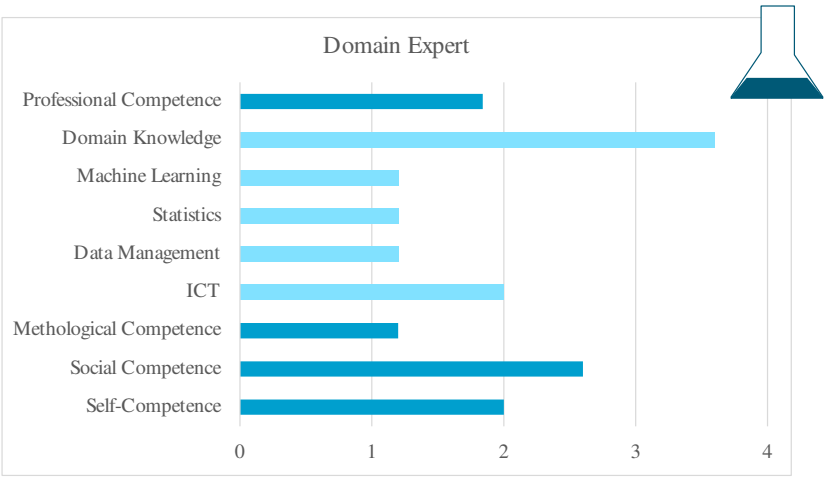

Figure 6: Competence profile of the Domain Expert role

2.4. Data Scientist

To enable ML-based analysis, the role of the Data Scientist aims to integrate ML and AI related know-how. Fundamentally, the role brings structuring options and process models for ML projects. Initially, an essential task is to enable the assessment of data availability and quality from an ML-oriented perspective. The goal in this context is to translate visions and ideas into realistic expectations. Furthermore, it helps in the selection of software for the ML environments. The content-related Data Science tasks such as feature engineering, explorative analyses as well

as training and validation of models are carried out by this role. Permanent coordination with the Domain Experts is important in order to ensure the feasibility and practicality of the analyses, to increase the quality of solutions and to create acceptance for later deployments. Furthermore, the Data Scientist plays a major role in the design of the deployment environment, which addresses the operationalisation of models. In particular, the transfer of analysis processes into scoring processes must be carried out by the data scientist in close coordination with IT and Domain Experts. The competence profile (Figure 7) addresses basic methodological competences to enable a structured implementation of ML projects. Since much of the implementation depends on the Data Scientist, self and social competences are also required. The core tasks, however, consist of technical competences in ML like programming languages, learning strategies and software tools. Furthermore, a Data Scientist must have high competences in the areas of ICT, data management and statistics in order to build up an adequate data connection management and to be able to represent the ability of model interpretation.

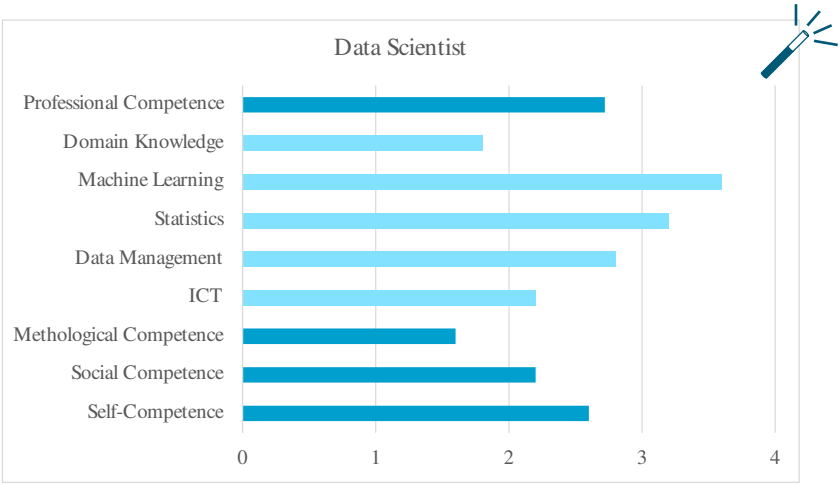

Figure 7: Competence profile of the Data Scientist role

2.5. Citizen Data Scientist

As a coordinating role between the specific roles of the Domain Expert, IT and Data Scientist as well as the Management, an integrating position of the operational project management is required. This orchestrating role of the CDS is responsible for typical project management tasks such as organising deadlines, scheduling and stakeholder management. It leads the analysis of the current situation and the collection of requirements in the early project phases of business and data understanding. The CDS plays an important role by translating real problems

into ML problems. During the technical work, it moderates between the various actors, contributes to the decision-making process and reports on the progress of the project to the Management. In the course of an ML implementation project, it has to assess various ML relevant contents, such as IT systems, data quality, software and model selection, evaluation criteria, performance metrics, etc. All of this leads to a new job description of CDS, since numerous Data Science related skills must also be developed to an action-oriented degree in this role. The competence profile of the CDS role (Figure 8) reflects the requirements of the heterogeneous tasks. Thus, the central orchestrating role has to meet the most diverse competence requirements. As the central contact for the project, he or she must have the highest level of self-social and methodological competence. This includes team building and motivational skills as well as organisational talent, analytical, structured and strategic thinking, flexibility, openness and willingness to adapt. In addition, the Citizen Data Scientist needs high hands-on competences in ML and esp. domain knowledge and thereby creates acceptance of models and deployments.

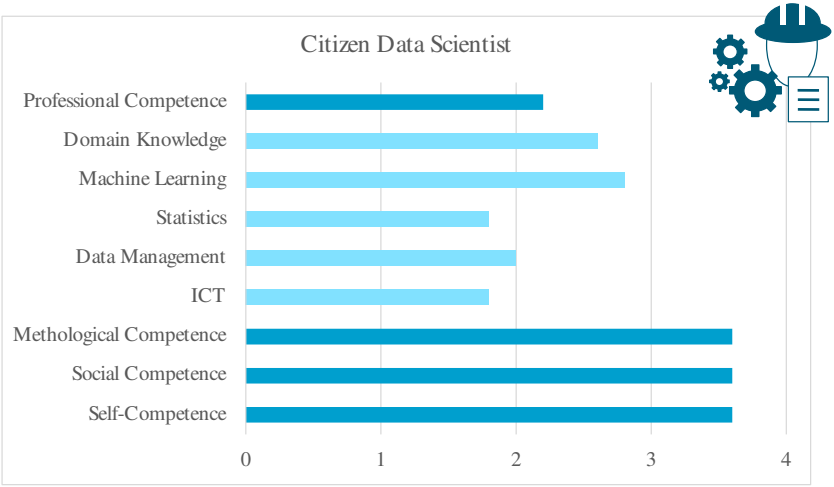

Figure 8: Competence profile of the Citizen Data Scientist role

3. Case Study in the Beverage Industry

The role model shows a conceptual structuring of actors in ML projects in manufacturing industry, which is to be specified individually for each initiative. For demonstration and validation, the following section shows an implementation example. For this purpose, interdisciplinary teams consisting of Domain Experts, IT, Data Scientists and Management were formed for each use case scenario in the DaPro project, which were orchestrated by a Citizen Data Scientist within a case

study on the use of ML to optimise malt yield at the Bitburger Braugruppe (Wöstmann et al. 2020).

3.1. General Matching of the Role Model and Project Organisation

In general, ML and AI methods needs to be part of the digitalization strategy of a company to have any chance for success. The mission for implementing data science has to be communicated from top down by the Management role to get the support from all stakeholders in a brewery. Failing in this basic requirement will delay a serious implementation and will result, if at all, in isolated solutions. Nevertheless, isolated solutions might also be the vital spark of interest for a more general approach.

An ideal proof of concept was to implement ML in a highly automated industry such as a large scale brewery as Bitburger Braugruppe. Having only minor experience in explorative data analytics such as methods like Six Sigma, implementation goes along with defining role models inside an already implemented project management system, based on the PRINCE2 project management method (AXELOS 2018). Working on a ML-based use case has many parallels to a standardized project cycle, adding specific steps of the CRISP-DM model to the project definition and the execution stage. In a regular managed project, there are roles defined such as a Steering Committee including members of supplier and customer representatives which monitor the definition and the result of a project. Experts on demand give general support with interfaces to all stakeholders and departments. A Project Manager to set up a project plan and sets milestones for the unique project steps. The Manager reports to the steering committee. And finally, an analysis and implementation team, which is responsible to complete the defined tasks. Depending on the complexity of the given project topic, the project itself can have multiple subprojects with different focuses. In this case there will be a responsible Project Manager for every subproject, but one Coordinating Manager above for orchestration. Optionally there is the possibility to include a coach for different tasks such as key lessons on new methods, consultation on best practices and in the worst case for de-escalation. In ML-based use cases the relevant departments or stakeholders, included in a project have been clearly defined (Figure 3). In the beginning it is necessary to map the roles of these stakeholders to positions and departments in breweries. Having also developed a brewery specific reference architecture (Figure 9) within the research project DaPro for developing ML-based use cases it is necessary to map the interface partners, which are actively involved in the use case, to key elements of the ML architecture. (Wöstmann et al. 2020).

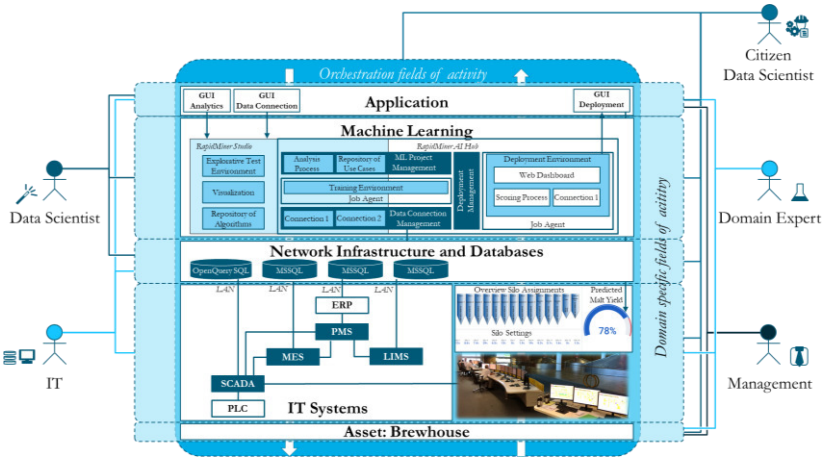

Figure 9: Key Fields of Activity of involved Roles in a Use Case-specific ML Architecture

The key project handler in data science projects is the CDS. In Breweries this position is responsible to set up the project management and has to be acquainted the basics of the production processes, quality assurance, database and IT architectures as well as ML methods and advanced statistics. Having all this know how combined, he is the most valuable team member of working on data science use case. The transdisciplinarity of this position requires an open minded engineer, either as a Domain Expert or from the Data Scientist area to deeply dive into the respective other discipline. The CDS has to identify possible use cases, get the necessary stakeholders (Domain Experts, IT) into the project team, and also reports the milestone results to both management and the customer representatives.

The role "Experts on Demand" of PRINCE2 is analogue to the Domain Experts in this proposed role model. The term "domain" is always referred to the focus of the respective department of the company where a use case is being implemented. Domain experts give the CDS the required information regarding the processes and the scope and boundaries of the use case itself. Respective to the DaPro project team members of the Domain Expert team should combine all aspects of production interests. E.g. a production manager should always consider quality aspects, thus a quality expert should be on the team as well. If the desired prediction results in changing downstream processes an expert of downstream departments might need to be consulted. The interface between production and the supply chain management is bottling and logistics.

Another kind of expertise is the participation of the IT department. In a classical project management approach, the IT is part of the project team or an "Expert on Demand". During a ML project the IT is a vital partner, as the ML architecture has to be hosted, maintained and monitored. When implementing the first ML use

case the first step is to manage the connectivity to all necessary data sources and a first introduction to the database schemas and its tables and its content. As a preferred conceptual focus, the IT role in breweries has both a dedicated technical IT focused on production and quality and a general IT for network administration. The IT supports the CDS in setting up a data science platform and provides the resources for hosting computationally intensive analytics processes. The IT has also a major responsibility for deploying process into production.

Depending on the progress of ML implementation breweries may have more than one Data Scientist employed and integrated in the brewery's organization. Data Scientists in breweries however are a luxurious asset, since their focus is more generalized, and they do not have much experience in domain processes. Data Scientist however have the general background of data analysis, its explorative approach, deep understanding of algorithms and modeling and also the ability to understand and interpret the model results. This asset may therefore be also part of the tasks of a CDS if a general Data Scientist is not at hand in the respective brewery's organization.

3.2. Exemplary industrial application scenario of malt yield prediction

Having described the individual specifications of the roles within a brewery, a use case to predict the malt yield in the brewhouse is used to validate this model. The malt yield is an important key figure to measure how much extract, has been extracted from the raw material malt depending on the produced quantity of wort and the total amount of malt used in the brewing process. Predicting of this key figure gives brewmasters enhanced possibilities to react in the automation of the processes and a choice in selecting the right ratio of raw materials. The earlier a brewmaster has prediction assistance the better will be the possibility to react.

A brewmaster in this use case has two roles, both being the Domain Expert and on the other hand the contracting entity with the initial idea and definition of the problem. Along the project cycle the Domain Expert has to give advice and evaluate anomalies found in the data for further interpretation. The use case has three major data sources for evaluating processes and raw materials. First being the supply and storage of raw materials, secondly the production processes itself with input and output parameters and thirdly the quality control of both product and raw materials. For this the Data Scientist or CDS is relying on data from SCADA (Sequence Control and Data Acquisition), LIMS (Laboratory Information Management System) and PMS (Production Management System) systems (Figure 9). Brushing against points of contact in quality assurance, the quality department can also give valuable input. In an initial phase of data understanding the Domain Experts (quality and production) help to understand which data is being collected and what the data means for employees outside the domain. A significant issue is the timestamp when a data point is being measured and appended to a data point list.

Shortly explained the time offset of a datapoint to a process is substantial to whether include this feature in a model or not.

It is important to note that most Domain Experts work with specialized reporting systems and seldom have information where the date is being stored and how the raw data looks like. At this point it is necessary to include the IT in the use case. The best-case scenario is to have an IT expert at hand who has a good inside how the backend of reporting systems look like to get fast access to raw data and support for basic modelling in the data preparation phase of the cycle. The data understanding and preparation phase is a constant circuit to help the Data Scientist or the CDS setting up a data pipeline and analyse which data becomes crucial for the prediction model.

The goal for this use case was to get the prediction feedback to the start of the process, meaning that when a production batch is started the result needs to be clear beforehand. The result of the data understanding phase was to shift the analysing focus on a complex silo management to develop a tracking system which malt deliveries ends up in a batch, paired with malt specifications of the incoming inspection. Bitburger uses 14 different silos, all reserved to a different supplier, generating a complex composition and mapping of deliveries. Having solved the silo black box on an abstract level, a large part of the complexity of this use case was removed, enabling training a model on historical data and providing first predictions. For predictions several algorithms were used, with Random Forest and Gradient Boosted Trees performing best in accuracy and error giving the brewmaster results within the measurement inaccuracy of the automation sensors and laboratory analysis.

ML models always have their own cross-validation included to measure prediction performance. However, these "raw" first predictions also need to be validated in a production environment. Here the Domain Experts play a significant role to check on the results itself. In collaboration with Domain Experts and Data Scientist a validation concept needs to be developed, setting up a prototype deployment in production and feeding the model with live production data. The IT department may have to be included on how to implement a deployment for the domain departments. The validation concept has to be evaluated on a regular basis to improve handling and the prediction results if necessary.

The management of a department or the brewery director needs to be informed of the progress and the results recurrently. Management always needs to assess the costs and risks of a use case and weigh up the benefit for the department or brewery since asset and personnel resources are bound during development and deployment. Moreover, the Management is responsible to mediate between departments and special interests. In this special use case interests of production planning and procurement are necessary to be informed in a future deployment strategy.

The CDS is the orchestrator to integrate all stakeholders in the use case. Depending on the progress the result may lead to a paradigm shift to reappraise procurement or planning processes in general, but this is a distant scope of the use case.

This use case is a significant part of proofing how a properly set up role model can help for introducing data science to a domain like breweries. Having implemented this use case several steps had to be solved along the project cycle. Beside finding and defining the use case itself it was also necessary to set up a ML-architecture (Figure 9), find all necessary data sources, built up a data pipeline for data understanding and preparation which was presented in more detail in (Wöstmann et al. 2020). Enhancing a reliable standardized project management method to ML-based use cases has helped to find the common thread for both data science and ordinary project management. Having developed several tools inside the project to enable non-data scientist to do basic modelling themselves the learning curve since introduction has been steep. Furthermore, a group of students was integrated in order to enable the development of practical competences by combining academic lectures with real InDaS challenges.

4. Outlook: Development of Citizen Data Science Competences

After considering the competences required for ML in the manufacturing industry and differentiating them into roles, the question arises how exactly the development of competences can take place, especially in a job-related perspective. Especially for SME there is a lack of practice-oriented ML competence development concepts, e.g. assistance in choosing training offers. Therefore, it is necessary to develop methodological support that allows the derivation of individual competence profiles and corresponding individual competence development activities. Based on the competence categories presented in section 1 and the role model presented in sections 2 and 3, a platform for ML competence development is to be created in the ML2KMU research project (Reckelkamm/Deuse 2021). The core component is a matching mechanism on basis of the role model for the target oriented derivation of competence development actions. For this purpose, it is required to evaluate existing competences of the respective employees for the designated roles. This evaluation could be conducted, for example, employing an assessment test, a self-assessment, or an external assessment. The assessment results could then be used to determine the competences that are not sufficiently fulfilled for the corresponding role in terms of a target-performance comparison. These competences, which do not fulfil the minimum requirements, can then be addressed individually with target-oriented and practical training programs. The recommendation and provision should be carried out automatically by the platform.

There are various formats for training programs, that mostly are offered online as e-learning courses, or partially as on-site events (Zschech et al. 2018). Online learning platforms esp. address massive open online courses (MOOC) such as Udemy,

Udacity, edX, Coursera, and Pluralsight. Also more ML-specific platforms such as DataCamp and Big Data University exist. In addition, large tech companies, including Amazon Web Services, Microsoft, Google, SAP or Cloudera as well as Data Science Platforms like RapidMiner, now offer their own training series customized to their own software. There are also larger academic institutions that offer corresponding ML courses and programs, such as the TU Dortmund University (e.g. the InDaS lecture), the TDWI Academy, the BITKOM Academy, the Data Science Academy in Karlsruhe, the Ludwig Maximilian University in Munich, the Westphalian Wilhelms University in Münster or the Zurich University of Applied Sciences (Zschech et al. 2018). All these courses, however, have in common that they are theoretical in nature and attempt to teach ML in a general way. Building up competences, however, requires experience and a context of practice as well as an industrial context for manufacturing related demands.

For this reason, the Institute of Production Systems, together with the RIF Institute in Dortmund, is developing a joint demonstrator for CDS (Figure 10), to enable students developing applicable oriented competences in the context of the InDaS course and to facilitate a transfer to industry in the context of the ML2KMU project in a strategy workshop concept. The demonstrator consists of a Cyber-Physical micro-brewery with industrial control technology, in which data-driven optimization of recipes is implemented by using ML. Thus, for example, regression analyses can be applied based on resource and process data as influencing parameters. Furthermore, it is also intended to gain a general understanding of complex patterns and relationships of cause and effect in brewing. Using the approachable example of brewing processes, the demonstrator acts as a test environment for the Industrial Internet of Things, ICT and is an ideal way to provide practical experience in the field of ML, whether for students or employees in the industry. Thereby, data flows from the PLC, Edge Devices and sensor technology via pre-processing steps to the Data Science Platform of RapidMiner are to be made transparent and tangible. For companies, this will be provided through the strategy workshops mentioned above, which offers another opportunity of developing ML competences. The employees can step into the individual roles themselves and in this way carry out a practical Data Science project exemplarily. This is to demonstrate and emphasize the relevance of the corresponding ML competences and roles. Besides developing fundamental ML competences, this should be an opportunity to derive strategies for future ML approaches in the own products and processes. These practical impressions should serve as a catalyst for future business models and a deeper use of ML methods at the participating companies.

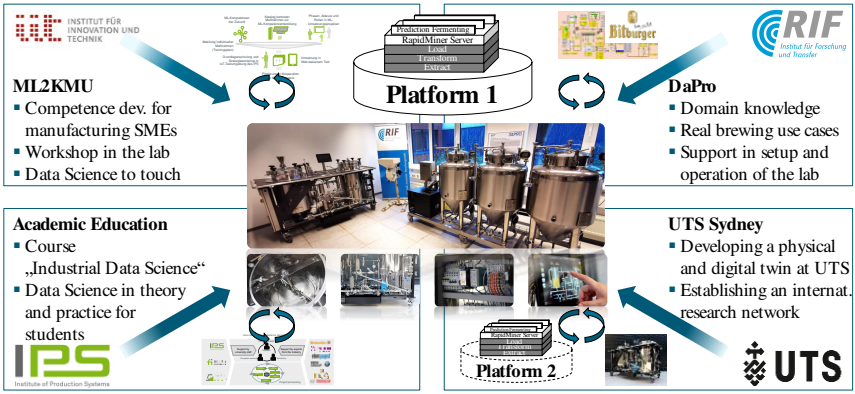

Figure 10: Interdisciplinary Citizen Data Science Demonstrator in Cyber-Physical Brewing Lab

The transdisciplinary work will continue to be expanded in the future through an international exchange with the University of Technology Sydney (UTS). In this context, a physic and digital twin of the brewing demonstrator will be set up in collaboration with UTS. Furthermore, the fundamentally important discussion must be conducted further in detail about which roles are to be mapped internally and externally in the area of conflict between (short-term) economic efficiency and flexibility as well as long-term digital sovereignty. In the vast majority of cases, however, the lack of corresponding competences is not due to an unwillingness on the part of management to support and develop employees accordingly, but to a lack of opportunities to do so in concrete terms. On the one hand, the constant competitive pressure prevents employees from having time to undergo lengthy training courses alongside their day-to-day work. On the other hand, the willingness of employees for further development and to actively demand this must also be created. This results in a growing tendency of outsourcing these activities to external service providers and thus to hand over the future asset of data (Reckelkamm/Deuse 2021). However, the consequent omission of internal development of these competences leads to endangering the company's own digital sovereignty. In other words, the development of these competences is a prerequisite for maintaining the own digital sovereignty and thereby continuing to retain authority over the own data. This is of great significance because digital sovereignty is the key factor for future competitiveness. External dependencies are reduced and self-determination in the digital space is strengthened in terms of autonomous and independent ability to act (BITKOM 2015).

Acknowledgements

This research and development project is/was funded by the German Federal Ministry for Economic Affairs and Energy (BMWi) in the program "Smarte Datenwirtschaft" (funding code 01MT19004D) and supervised by the DLR Projektträger.

Supported by:

Federal Ministry for Economic Affairs and Energy

on the basis of a decision by the German Bundestag

References

AXELOS (2018). Erfolgreiche Projekte managen mit PRINCE2 (6th ed.). Norwich: TSO (The Stationery Office).

Bauer, N., Stankiewicz, L., Jastrow, M., Horn, D., Teubner, J., Kersting, K., . . . Weihs, C. (2018). Industrial Data Science: Developing a Qualification Concept for Machine Learning in Industrial Production. In *European Conference on Data Analysis (ECDA),* Paderborn.

Bertelsmann Stiftung (2018). *Zukunft der Arbeit in deutschen KMU*. https://doi.org/10.11586/2019059

BITKOM (2015). *Digitale Souveränität: Positionsbestimmung und erste Handlungsempfehlungen für Deutschland und Europa*. Berlin. Retrieved from https://www.bitkom.org/sites/default/files/file/import/BITKOM-Position-Digitale-Souveraenitaet.pdf

Bloom, B. S., & Engelhart, M. D. (Eds.) (1976). *Beltz-Studienbuch: Vol. 35. Taxonomie von Lernzielen im kognitiven Bereich* (5. Ed.). Weinheim: Beltz.

Deuse, J., Erohin, O., & Lieber, D. (2014). Wissensentdeckung in vernetzten, industriellen Datenbeständen. In H. Lödding (Ed.), *Schriftenreihe der Hochschulgruppe für Arbeits- und Betriebsorganisation e.V. (HAB), Industrie 4.0: Wie intelligente Vernetzung und kognitive Systeme unsere Arbeit verändern* (pp. 373–395). Berlin: Gito.

Dragicevic, N., Ullrich, A., Tsui, E., & Gronau, N. (2020). A conceptual model of knowledge dynamics in the industry 4.0 smart grid scenario. *Knowledge Management Research & Practice*, *18*(2), 199–213. https://doi.org/10.1080/14778238.2019.1633893

Eickelmann, M., Wiegand, M., Konrad, B., & Deuse, J. (2015). Die Bedeutung von Data Mining im Kontext Industrie 4.0. *Zeitschrift Für Wirtschaftlichen Fabrikbetrieb (ZWF)*, *110*(11), 738–743.

Fölsch, T. (2010). *Kompetenzentwicklung und Demografie*. Zugl.: Kassel, Univ., Diss., 2010. *Schriftenreihe Personal- und Organisationsentwicklung: Vol. 9*. Kassel: Kassel Univ. Press.

Idoine, C., & Brethenoux, E. (2019). *Maximize the Value of Your Data Science Efforts by Empowering Citizen Data Scientists*.

Mazarov, J., Wolf, P., Schallow, J., Nöhring, F., Deuse, J., & Richter, R. (2019). Industrial Data Science in Wertschöpfungsnetzwerken: Konzept einer Service-Plattform zur Datenintegration und -analyse, Kompetenzentwicklung und Initiierung neuer Geschäftsmodelle. *Zeitschrift Für Wirtschaftlichen Fabrikbetrieb (ZWF)*, *114*(12), 874–877.

Miller, S., & Hughes, D. (2017). *The Quant Crunch: How the Demand for Data Science Skills is Disrupting the Job Market*. Boston, MA.

Gartner, Inc. (2017, January 16). *Gartner Says More Than 40 Percent of Data Science Tasks Will Be Automated by 2020: Analysts to Explore Trends in Data Science at Gartner Data & Analytics Summits 2017* [Press release]. Sydney, Australia.

Morik, K., Deuse, J., Stolpe, M., Bohnen, F., & Reichelt, U. (2010). Einsatz von Data-Mining-Verfahren im Walzwerk. *Stahl Und Eisen*, *130*(10), 80–82.

Mullarkey, M. T., Hevner, A. R., Grandon Gill, T., & Dutta, K. (2019). Citizen Data Scientist: A Design Science Research Method for the Conduct of Data Science Projects. In B. Tulu, S. Djamasbi, & G. Leroy (Eds.), *Lecture Notes in Computer Science. Extending the Boundaries of Design Science Theory and Practice* (Vol. 11491, pp. 191–205). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-19504-5_13

North, K., Brandner, A., & Steininger, T. (2016). Die Wissenstreppe: Information – Wissen – Kompetenz. In K. North, A. Brandner, & M. Steininger (Eds.), *essentials. Wissensmanagement für Qualitätsmanager* (pp. 5–8). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-11250-9_2

Rammer, C., Köhler, C., Murmann, M., Pesau, A., Schwiebacher, F., Kinkel, S., . . . Som, O. (2010). *Innovationen ohne Forschung und Entwicklung: Eine Untersuchung zu Unternehmen, die ohne eigene FuE-Tätigkeit neue Produkte und Prozesse einführen* (Studien zum deutschen Innovationssystem No. 15-2011). Mannheim und Karlsruhe.

RapidMiner (2020). *Building the perfect AI team*. Retrieved from https://rapidminer.com/resource/building-ai-team/

Reckelkamm, T., & Deuse, J. (2021). Kompetenzentwicklung für Maschinelles Lernen zur Konstituierung der digitalen Souveränität. In E. A. Hartmann (Ed.), *Digitalisierung souverän gestalten: Innovative Impulse im Maschinenbau* (pp. 31–43). Berlin: Springer Vieweg. https://doi.org/10.1007/978-3-662-62377-0_3

Saltz, J. S., & Grady, N. W. (2017, December). The ambiguity of data science team roles and the need for a data science workforce framework. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 2355–2361). IEEE. https://doi.org/10.1109/BigData.2017.8258190

Schulte, L., Schmitt, J., Stankiewicz, L., & Deuse, J. (2020). Industrial Data Science: Interdisciplinary Competence for Machine Learning in Industrial Production. In T. Schüppstuhl, K. Tracht, & D. Henrich (Eds.), *Annals of Scientific Society for Assembly, Handling and Industrial Robotics* (pp. 161–171). Berlin: Springer Vieweg.

Stark, R., Deuse, J., Damerau, T., Reckelkamm, T., & Lindow, K. (2019). Data preparation for data analytics (DPDA): Arbeitsgruppe des prostep ivip e.V. *News. Wissenschaftliche Gesellschaft Für Produktentwicklung*. (2), 4–6.

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Ed.), *Leistungsmessungen in Schulen* (pp. 17–31). Weinheim: Beltz.

Wölfl, S., Leischnig, A., Ivens, B., & Hein, D. (2019). From Big Data to Smart Data – Problemfelder der systematischen Nutzung von Daten in Unternehmen. In W. Becker, B. Eierle, A. Fliaster, B. Ivens, A. Leischnig, A. Pflaum, & E. Sucky (Eds.), *Geschäftsmodelle in der digitalen Welt* (pp. 213–231). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-22129-4_11

Wöstmann, R., Reckelkamm, T., Deuse, J., Kimberger, J., Temme, F., Schlunder, P., & Klinkenberg, R. (2019). Datengetriebene Prozessoptimierung in der Getränkeindustrie. *Fabriksoftware*, *24*(03), 21–24.

Wöstmann, R., Schlunder, P., Temme, F., Klinkenberg, R., Kimberger, J., Spichtinger, A., . . . Deuse, J. (2020). Conception of a Reference Architecture for Machine Learning in the Process Industry. In *2020 IEEE International Conference on Big Data*. Symposium conducted at the meeting of Institute of Electrical and Electronics Engineers (IEEE), Atlanta (virtual).

Zschech, P., Fleißner, V., Baumgärtel, N., & Hilbert, A. (2018). Data Science Skills and Enabling Enterprise Systems. *HMD Praxis Der Wirtschaftsinformatik*, *55*(1), 163–181. https://doi.org/10.1365/s40702-017-0376-4