Michael Schulz ▪ Uwe Neuhaus ▪ Stephan Kühnel ▪ Heiko Rohde ▪ Sayed Hoseini ▪ René Theuerkauf (Eds.)

DASC-PM v1.1 Case Studies

dasc°pm v1.1

Elmshorn 2023

info@dasc-pm.org

Publisher:

NORDAKADEMIE gAG Hochschule der Wirtschaft
Köllner Chaussee 11
25337 Elmshorn

Supported by the
NORDAKADEMIE foundation

The current version of the DASC-PM was published as
Schulz, Michael; Neuhaus, Uwe; Kaufmann, Jens; Kühnel, Stephan; Alekozai, Emal M.; Rohde, Heiko; Hoseini, Sayed; Theuerkauf, René; Badura, Daniel; Kerzel, Ulrich; Lanquillon, Carsten; Daurer, Stephan; Günther, Maik; Huber, Lukas; Thiée, Lukas-Walter; zur Heiden, Philipp; Passlick, Jens; Dieckmann, Jonas; Schwade, Florian; Seyffarth, Tobias; Badewitz, Wolfgang; Rissler, Raphael; Sackmann, Stefan; Gölzer, Philipp; Welter, Felix; Röth, Jochen; Seidelmann, Julian; Haneke, Uwe: "**DASC-PM v1.1 - A Process Model for Data Science Projects**", NORDAKADEMIE gAG Hochschule der Wirtschaft, Hamburg 2022, ISBN: 978-3-9824465-1-6, http://dx.doi.org/10.25673/91094.

Table of Contents

Foreword

The development of the DASC-PM (Data Science Process Model) is based on the knowledge of a large working group consisting of experts in Data Science. We believe that the procedure described in the DASC-PM can be used beneficially in data-driven projects and offers support in structuring complex projects. In version 1.1 of the DASC-PM, which was published in German in March 2022 and in English in June 2022, we significantly revised the original documentation based on feedback from readers. We hope that the new document is easier to understand and more structured than the previous one, thus reducing the barriers to using the process model for the first time.

With this publication, we realize another suggestion of the readers: a collection of case studies in which the DASC-PM is applied. On the one hand, the documentation of the DASC-PM claims to show the procedure of a Data Science project as detailed as possible. On the other hand, however, it is also intended to ensure universality, which is reflected in a thoroughly broad spectrum of applications. Showing examples of the direct application and usability of DASC-PM in different application domains represents another aspect that is intended to reduce the aforementioned application barriers.

This collection of case studies is intended to provide readers with guidance on the use of DASC-PM. This use is presented from different perspectives and at very different levels of abstraction. Thereby, projects from practice and science are considered. We thank all authors of the case studies for their commitment to the working group and are very pleased to see the exciting contributions that have emerged.

However, the present collection of case studies should not be seen as complete. We are always interested in your feedback on the use of DASC-PM in practical or scientific projects and would also be happy to add your case study to our collection.

If you are interested in joining our working group or would like us to keep you informed about current developments regarding the DASC-PM, feel free to get in touch at the contact address below.

Elmshorn, Flensburg, Halle (Saale), Hamburg, and Krefeld in February 2023

The DASC-PM Core Team

Contact: info@dasc-pm.org

1 Retaining Knowledge with Knowledge Graphs

Carsten Lanquillon, Sigurd Schacht

1.1 Introduction

When employees depart from small and medium-sized companies, the loss of their experience and knowledge can threaten a company's existence. Yet, there is also tremendous value in daily work when employees have easy access to existing experience and knowledge in a company, as it promotes exchange and collaboration. Central building blocks for comprehensive knowledge management are securing and transferring existing experience and knowledge (Probst, 2013).

Successfully implementing knowledge management typically requires tedious manual work for recording knowledge that often only exists implicitly in documents and the minds of people or was expressed in dialogues. Structured interviews and debriefings can document and pass on relevant experience and knowledge at the end of a project when an employee departs (Müller and Kaiser, 2006). Typically, the manual form of recording knowledge is a bottleneck that inhibits securing and sharing experience and knowledge, putting the successful implementation of comprehensive knowledge management at risk.

Therefore, this data science project aimed at developing and applying AI-based methods for the greater automation of the knowledge recording process. Machine learning and natural language processing (NLP) techniques were used to extract knowledge fragments from project artifacts such as reports, documentation, or communication between involved persons and charted in a knowledge graph. Since these methods map the relevant experience and knowledge often only incompletely and with a certain degree of uncertainty, they were refined and complemented by applying classical methods of knowledge recording, as illustrated in Figure 1.1.

A suitable preparation and display of the existing knowledge graphs in the form of knowledge maps, for example, can help knowledge management experts prepare for their departure interviews and debriefings, so the recording process is significantly accelerated, and the quality of the recorded knowledge fragments is improved. Furthermore, knowledge graphs are crucial for further knowledge management applications such as chatbots or other AI-based assistance systems.

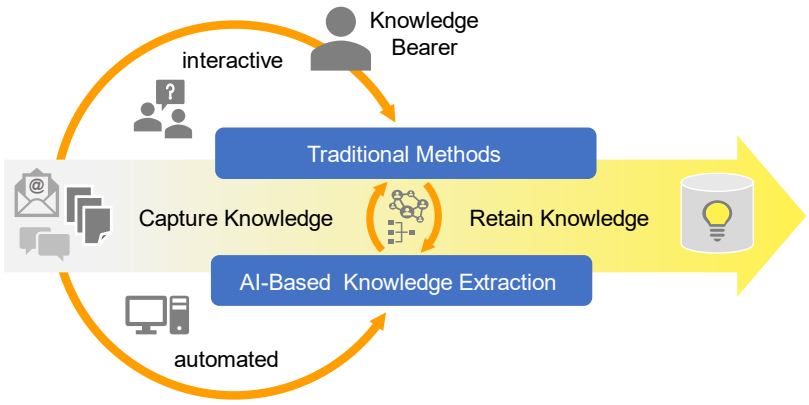

Figure 1.1: Various sources should supply knowledge for recording and storage automatically (AI-based) and interactively, allowing for use during new tasks or sharing. Source: Authors' own illustration.

1.2 Procedure

Key area: "Scientificity"

This case study is based on findings and experiences gained during research projects in AI-based knowledge management. All phases of the case study in this project were accompanied by comprehensive analyses of the literature documenting the latest developments in the respective subareas to select suitable approaches and methods. In the areas below, there is no further explicit reference to the cross-sectional character of this key area. Since this case study focuses on the description of the approach, there will be no detailed presentation of the findings from the literature analysis or the analysis methods adopted.

Phase: "Project description"

The project dealt withknowledge management within companies. It is a cross-sectional discipline that can be deployed in any technical domain. Background knowledge in functional or technical domains is crucial for determining, recording, and evaluating relevant knowledge fragments. Still, it does not affect the preparation and representation of the procedure in this case study. For this reason, no technical domain is discussed in this case study. It is assumed that project-based activities and suitable text-based artifacts are available. These artifacts are also referred to as documents in short.

Use case

The project's overarching goal was to develop a knowledge graph with all the relevant relations (knowledge fragments) in a defined functional or technical domain of a company as a component of securing knowledge. In knowledge management, the construction of a knowledge graph shows the data preparation that takes unstructured documents (project artifacts) and produces a structured database (knowledge graph) from them. This, in turn, equates to the analytical data source regarding its use in knowledge management applications and other AI-based assistance systems.

While the solution of simple canonical data science tasks such as classification or numeric forecasting can often be mapped and described directly and holistically in a DASC-PM cycle, many complex tasks such as securing knowledge with knowledge graphs based on documents require a break-down of the problem into subtasks.

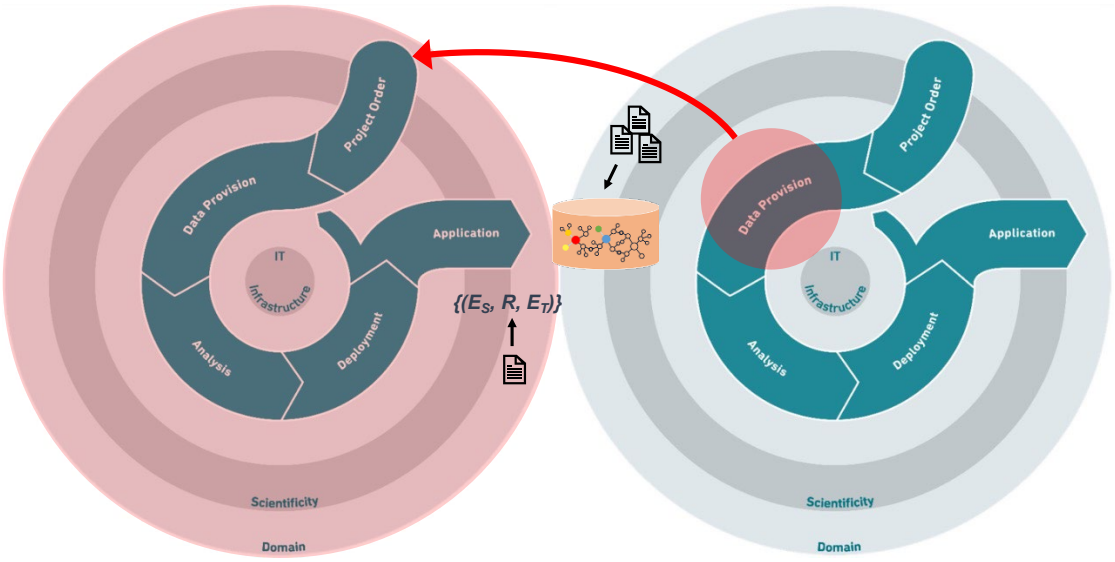

Figure 1.2: Applying the DASC-PM by constructing a knowledge graph from texts to prepare data to secure knowledge on a macro-level (right) and comprehensively extract knowledge from documents on the micro-level (left). Source: Authors' own illustration, based on Schulz et al. (2022).

Such nested solution patterns are typical for complex problems in real application domains. Concerning implementation, it is essential to use a process model that allows for uniform application both on the macro-level of the entire project and on the micro-level of individual steps and subtasks. This case study describes how the DASC-PM is applied with its key areas and phases on all granularity levels.

We regard knowledge fragments here to be so-called SPO triples or, in short, triples (E_S, R, E_T) consisting of subject E_S, predicate R and object E_T. The predicate defines a relation between the start entity E_S and the target entity E_T, which is represented in a graph as a directed arrow between the entities as nodes. The set of all triples constitutes the knowledge graph. As described above, it is a complex task that can be implemented using suitable machine learning methods and natural language processing (NLP). Figure 1.2 shows the nested application of the DASC-PM intended for this.

Project outline

The scholarly literature describes two extremes for constructing knowledge graphs (Zhao et al., 2018): In the top-down approach, an ontology is defined in advance as a set of schemas with all the required types of entities and relations. Then it is filled with suitable entities and relations between these entities, extracted from the available documents. In the bottom-up approach, the entities and relations are initially extracted as isolated knowledge fragments that are usually not assigned a type. A typification and summary for meaningful structures are constructed afterwards.

To clarify the feasibility, a test of the two extremes showed that each exhibits grave disadvantages. Modeling a complete ontology ex-ante and structuring entities ex-post are very time-consuming. In addition, defining the schemas in advance is often not feasible since the entity and relation types needed in an application are usually not entirely known at the beginning of a project.

Clear parallels to data warehousing are evident here. That is because data warehousing is also holistic modeling in advance according to the pure top-down approach based on Inmon, but often cannot be achieved in practice. In contrast, according to Kimball, the bottom-up approach

can result in problems integrating the data. In practice, a mixed form has been established that is in line with the well-known principle of “think big, start small, grow step by step” (SCN, 2013).

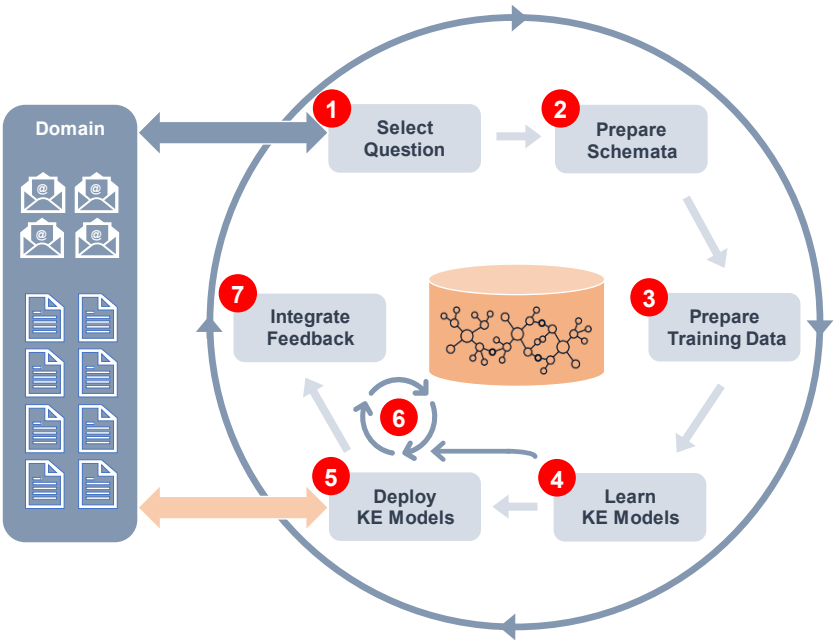

Figure 1.3: The iterative construction of a knowledge graph based on questions provided and selected by users in the domain requires the preparation of individual models for knowledge extraction (KE models). Knowledge is modeled in well-defined steps that are seamlessly inserted into the DASC-PM. Source: Authors’ own illustration.

This principle is also based on the procedure selected here for constructing knowledge graphs. The vision is a completely typed knowledge graph with corresponding schemas. However, these are not fully modeled and provided ex-ante as in the pure top-down approach, but iteratively. As depicted in Figure 1.3, the knowledge graph is successively expanded by selected questions to include new knowledge fragments. Each iteration in the steps for knowledge modeling triggers the processing of a separate data science task that is then also processed on the micro-level, also based on the DASC-PM. The definition of a relevant question (step 1), which recurs in each iteration, equates to an internal project description to prepare a suitable model for the knowledge extraction as an analysis artifact and includes the key area of domain with its embedding in the application domain. The technical knowledge obtained from the application is used to identify the entity and relation types necessary and relevant for the answer and to model them as schemas (step 2). This step and the procurement of training data (step 3) as an analytical data source can be assigned to the data procurement phase. In step 4, the models are prepared for knowledge extraction as the main task in the analysis phase. The models' deployment and application (steps 5 and 6) for extracting knowledge fragments from all the relevant documents are explicitly complemented here by the possibility of integrating feedback from users (step 7). The feedback, which the system may actively request, is also included in the knowledge graph and used for evaluating and improving the models deployed. The advantage of this procedure is that the knowledge fragments are already available in typed form, and the effort is limited to the ex-post typification and structuring. Furthermore, the initial results from the knowledge modeling are already used in applications early and thus validated.

Phase: “Data preparation”

After the definition of the problem, the preparation of the analytical data source plays a central role in the success of a data science project. Since the basis for the learning process in this case study consists of mostly unstructured textual data, details of the steps to be carried out and suitable methods substantially differ from projects with structured data. The main aspects of the DASC-PM phases retain their validity and are applicable for structuring the project both on the macro-level and the micro-level.

Original data sources

The basis for the data in this case study is a set of text-based project artifacts such as reports and documentation and electronically available correspondence between people involved in the projects. These text-based artifacts are briefly referred to as documents and the entire set as a collection.

Data management

For the sake of simplicity, it is assumed that all artifacts that should and may be used for the knowledge extraction are in directories specially designated for this and are available in a format allowing for the text to be directly extracted. The data management must be adapted and expanded for operational use. Not only is automatic access to relevant project artifacts necessary, but it also ensures the protection of personal or confidential documents.

Exploratory data analysis

Since the available data are unstructured, the exploratory data analysis for understanding the data differs significantly from projects with structured data. Nonetheless, it is essential to view the documents in their original form. It is necessary to check whether the documents are available in the same language or multilingual, which requires special treatment. Likewise, the linguistic quality of the documents may play a critical role. Quickly formulated messages often contain many abbreviations and more linguistic errors than carefully written texts. In these cases, the upstream substitution of abbreviations or the automatic correction of mistakes can improve the quality of the following processing steps. In this case study, however, we assume that single-language documents are available and that the quality is acceptable.

Data preparation: from the original data source to the analytical data source

On the macro-level of the project, extracting relevant knowledge fragments from the documents provided resembles data preparation. The first knowledge graph is the analytical data source on which the subsequent analysis method builds. In the deployment and application phases, the theme of regular knowledge extraction is addressed to consider new documents.

On the micro-level of the project, the knowledge extraction is realized by learning and applying suitable methods and models. As most methods cannot handle text as input directly, the text representation constitutes a critical aspect here. Since individual subtasks in knowledge extraction often have different data preparation requirements, it is impossible to define a generally applicable representation method. Figure 1.4 shows some classical steps in the data preparation for a text.

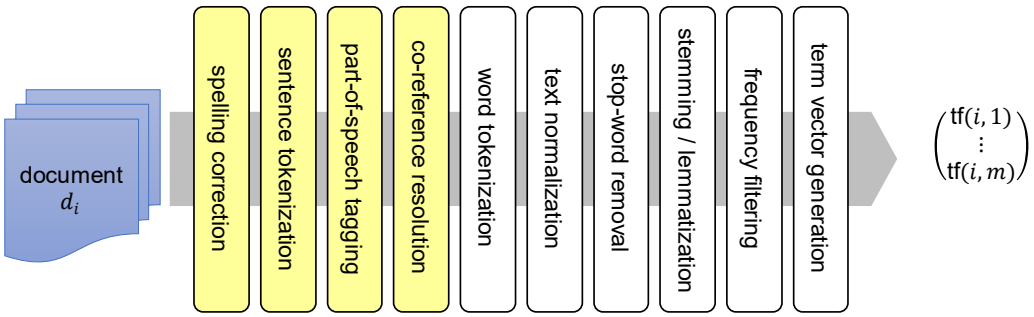

Figure 1.4: Classical steps in preparing data for transfer from text to a numeric vector representation easily accessible for algorithmic processing. Less typical steps are displayed in yellow, and frequently used steps are in white. Source: Authors' own illustration.

The classical preparation of text illustrated above often involves many steps until a document is converted to a numeric vector representation. It is necessary to consider that this simple form of preparation regards documents as a unit, and the order of the words is ignored. However, some tasks in knowledge extraction require an analysis on the word level, with consideration given to the context and the order of the words. Individual sub-steps remain relevant, but their application must be selected very carefully. In part, the results of the sub-steps no longer lead to a transformation of the original data but an enrichment of the representation, for example, by including tags for certain types of words. Applied to individual words, this vector representation results in so-called one-hot encoding where the mere presence is a one at the position, with the index matching the position of the corresponding word in a defined word list (vocabulary) and all other values being 0. Since this representation does not consider the semantics of the words, so different words are all equally dissimilar, modern NLP approaches primarily use global word embeddings, which assign a statistical numeric representation to each word according to a linguistic model, or encoder representations that take account of the specific context of a word in the determination of suitable representation. These modern approaches automate a significant part of the data preparation. Components to produce appropriate text representations are then part of comprehensive model architectures and are either taken from the pre-trained models without changes or explicitly adjusted in a fine-tuning step to the unique features of a domain so that elementary preparation steps are not necessary for this phase.

The procedure for enriching the analytical data source to include possible labels for target variables when applying supervised learning procedures is discussed during the viewing of the individual NLP tasks in the key area of the analysis method.

Phase: "Analysis"

The extraction of knowledge fragments (triples) comprises multiple NLP tasks, as illustrated in Figure 1.5 as a pipeline. While the first step (PREP) with classical data preparation is assigned to the data procurement phase, the following steps consist of specialized models learned from data either in isolation or in combination (analysis phase). The knowledge fragments extracted from a document are stored in a knowledge graph. The extraction process is applied to all prepared documents and may be done multiple times in the event of improvements to the individual model after receiving feedback (use phase).

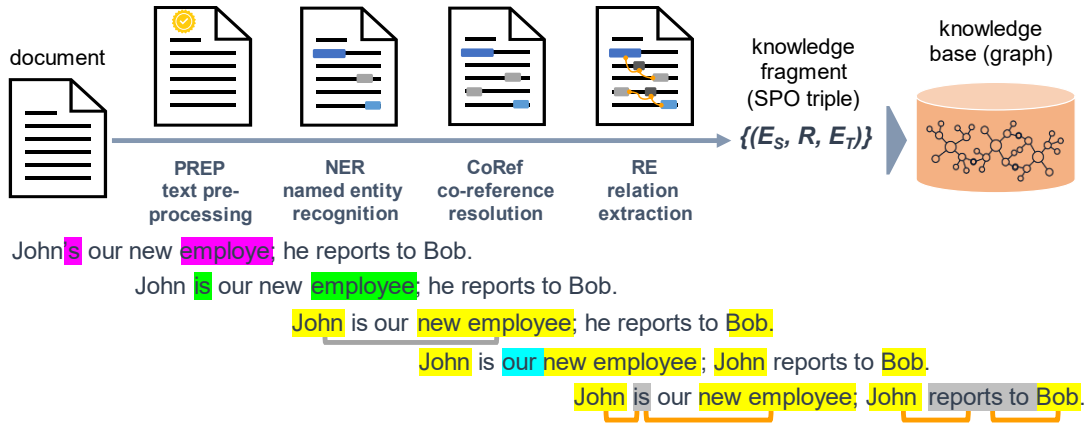

Figure 1.5: NLP tasks in a processing pipeline for extracting knowledge fragments from a document. Source: Authors' own illustration.

To better understand and simplify the process-supporting controls and evaluation, we will start by examining the steps in isolation. Note that for each NLP task to be addressed and solved, a separate DASC-PM cycle is required (micro-level). It is essential to study the latest developments in research in-depth to consider them in the project adequately. Below, the objectives of the individual phases are only briefly outlined with some current solution variants.

NLP tasks for the extraction of knowledge fragments

The objective of named entity recognition (NER) is identifying entities with specific, predefined, semantic types such as persons, places, or organizations based on fixed designations given to them. Numerous pre-trained models exist for these generic types and often generate good results in general texts. Domain-specific applications often require not only an adjustment of the occurring entities of these types but also an expansion to additional domain-specific types, such as identifying data science terms in research articles. The NER task is currently often successfully solved with deep learning approaches such as bidirectional long-short-term memory (LSTM) architectures. Dependencies between words, for example, with entities that consist of multiple words, are considered very well with a final conditional random field layer (Goyal et al., 2018; Panchendrarajan & Amaresan, 2018). When entity linking is added, entities are combined, referring to the same object in the real world. Methods for coreference resolution (CoRef) replace reference words in documents with the entities to which they refer in the text. Only then can knowledge fragments be sensibly extracted from isolated parts of the sentence independently of the larger context. A well-known class of methods for this can be found in the Mention Pair models (Ng, 2017). The Relation Extraction (RE) should identify relevant relations between two entities where at least one of the relevant entities is involved. Recurrent deep learning approaches are also increasingly used for this task (Wang et al., 2021).

Procurement of training data for supervised learning methods

Supervised deep learning approaches currently achieve outstanding performances for most NLP tasks. Using pre-trained models with subsequent fine-tuning for the respective domains reduces the set of labeled data needed to create the model. A two-step approach simplifies the procurement of the training data: Based on rules, the first training data are generated with lists of known entities and relations as seed values. These data are used to fine-tune the deep learning approaches in the second step. The learning process is also supported by active learning based on new knowledge fragments that can be systematically queried via feedback components (Shen et al., 2018).

Organization of learning processes

The results generated in each step are combined into a processing knowledge extraction pipeline in the use phase. Note that with this approach, errors in any of the steps along the pipeline will accumulate and, hence, are likely to aggravate over time. This can be avoided by the combined learning of a joint knowledge extraction model (Singh et al., 2013; Geng et al., 2021). However, providing sufficient training data for this approach is usually more complex and time-consuming. Furthermore, the joint model is more difficult to reuse for extracting knowledge fragments that share specific entity or relation types. Therefore, joint modeling is not used here.

Phases: Deployment and Application

On the micro-level, pipelines with models for the different NLP tasks are developed and applied for each question. In combination, they enable the extraction of dedicated knowledge fragments. They are applied in the deployment phase such that the knowledge graph is initially filled based on all available documents.

The quality of the extraction is monitored during use via the feedback components provided. Any new knowledge fragments, which the system may also actively query, will be used to expand the analytical data sources. The knowledge graph is expanded accordingly as soon as new documents are available or models of the extraction pipeline have been improved based on user feedback. This is comparable to the application and maintenance of ETL pipelines in data warehousing.

Regarding the overall project (macro-level), using all extraction pipelines to construct the knowledge graph as an analytical data source is an essential component of the data procurement phase. The analytical data source is used for further knowledge management applications, such as creating knowledge maps to support interviews and debriefings or developing AI-based assistance systems. On the macro-level, these applications are based on analysis artifacts generated in the analysis phase, and their deployment and use must be planned explicitly.

1.3 Conclusion

This case study shows that the DASC-PM as a process model offers an ideal framework for projects with more complex problems. Derived from the use case, the AI-based securing of knowledge in this case study, the project was broken down into smaller tasks in the project description phase. This supplied a clear framework for setting up the macro-level in the project. The solution and implementation of each subtask also took place via the DASC-PM as a process model to handle the individual activities such as the problem recording, data provision, and the execution of the analysis in a structured way. In this case study, the results were combined in the deployment and application phases. This showed that the DASC-PM is flexible enough to develop complex problems with multi-layered solution components. Finally, the key area of scientificity should be emphasized since hardly any AI project can be pursued without knowing the latest developments in research by analyzing the literature and selecting suitable methods on all levels of implementation. The DASC-PM also offers the necessary framework for this.

Literature

Geng, Z., Zhang, Y. & Han, Y. (2021). Joint entity and relation extraction model based on rich se-semantics. Neurocomputing, vol. 429, pp. 132–140.

Goyal, A., Gupta, V. & Kumar, M. (2018). Recent named entity recognition and classification techniques: A systematic review. Computer Science Review, Vol. 29, No. 1, pp. 21–43.

Müller, M. & Kaiser, R. (2006). Wissensbewahrung bei der Stadt Erlangen - Dokumentation und Kommunikation der Erfahrungen ausscheidender Wissensträger. In the conference volume Know- Tech 2006, pp. 425–432.

Ng, V. (2017). Machine Learning for Entity Coreference Resolution: A Retrospective Look at Two Decades of Research. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), pp. 4877–4884

Panchendrarajan, R. & Amaresan, A. (2018). Bidirectional LSTM-CRF for Named Entity Recognition. In Proceedings of PACLIC 2018, pp. 531–540.

Probst, G., Raub, S. & Romhardt, K. (2013). Wissen managen: Wie Unternehmen ihre wertvollste Ressource optimal nutzen. 7th edition, Springer Gabler.

Schulz, M., Neuhaus, U., Kaufmann, J., Kühnel, S., Alekozai, E., Rohde, H., Hoseini, S., Theuerkauf, R., Badura, D., Kerzel, U., Lanquillon, C., Daurer, S., Günther, M., Huber, L., Thiee, L., zur Heiden, P., Passlick, J., Dieckmann, J., Schwade, F., Seyffarth, T., Gölzer, P., Welter, F., Röth, J., Seidelmann, J.

& Haneke, U. (2022). DASC-PM v1.1 - Ein Vorgehensmodell für Data-Science-Projekte. NORDAKA- DEMIE gAG.

SCN Education B.C. (2013). Data Warehousing: The Ultimate Guide to Building Corporate Business Intelligence. Vieweg+Teubner Verlag.

Shen, Y., Yun, H., Lipton, Z. C., Kronrod Y. & Anandkumar, A. (2018). Deep Active Learning for Named Entity Recognition. In Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 252–256.

Singh, S., Riedel, S., Martin, B., Zheng, J. & McCallum, A. (2013). Joint inference of entities, relations, and coreference. In Proceedings of the 2013 Workshop on Automating Knowledge Base Construction (AKBC 2013). Co-located with CIKM 2013, pp. 1–6.

Wang, H., Lu, G., Yin, J. & Qin, K. (2021). Relation Extraction: A Brief Survey on Deep Neural Network Based Methods. In Proceedings of the 4th International Conference on Software Engineering and Information Management (ICSIM 2021), pp. 220–228.

Zhao Z., Sung-Kook H. & So I.-M. (2018). Architecture of Knowledge Graph Construction Techniques. International Journal of Pure and Applied Mathematics 118(19): pp. 1869–1883.

2 Collaborative Recommendation Service for Highly Correlated Dialogue Inputs

Julian Seidelmann

2.1 Motivation and project description

In 2020 Germany imported over one trillion euros of goods. Control over imported goods is only gained after they have been declared in a customs procedure and released by a customs authority. A customs declaration is necessary for the declaration of the goods. This declaration is an official document in which imported or also exported goods are listed and the information on them is stated.

Today, customs declarations are usually issued electronically via software services and ATLAS, the information system of the German customs authority. The recording of customs declarations is both a time-consuming and complex process and requires in-depth expertise in customs and regulations under customs law. In addition to the complexity and the amount of time expended, the manual recording of data is always subject to the risk of human error. The digitalisation of declarations and forms can be viewed as an opportunity to support the entry of information by users.

In a project spanning 75 person-days, it is necessary to examine whether data science techniques can support users when they enter information and thus reduce the workload of users in terms of both technical complexity and the expenditure of time. This proof of concept is intended as a basis and suitability test for follow-up projects and a possible integration in the existing system. The risk of this project is to be classified as low. The DASC-PM process model is used to carry out the project. The structuring provided by the process model is especially helpful for the project participants with limited experience in data science projects.

A high degree of digitalisation in the customs process can entail not only more efficient processes, but also new paths for better compliance, more visibility in the supply chain and ultimately a competitive advantage. One approach for supporting users in the use of dialogues is to recommend possible entries to be made in dialogues. One type of recommendation service consists of collaborative recommendation services. The implementation manuals of German customs authorities define various requirements and validations that entail a high correlation of data. The various company-specific processes also allow for high data correlations to be drawn. Only these different correlations can be used to take the large quantity of customs declarations and infer the entries of one single customs declaration from them.

2.2 Data procurement

Customs declarations as evidence for tax purposes are subject to a statutory retention period of ten years. The data for this research project was collected over many years.

It was decided that a selection of ten companies would be examined and the period would be limited to the last five months. Those ten companies were selected to capture a range of variations in the number of customs declarations. The number of datasets per month varies from 200 to 5,000 customs declarations. The original source of the data is a relational database in which all the required data are structured and internally available. A special feature of this project is that only one data source is used. It is not necessary to enrich these data with external sources. The correlation of the data from the customs declarations is sufficient to train this recommendation service. An exploratory analysis of the data and a visualisation as a heatmap (see Figure 2.1) can confirm the already assumed high correlation, but also reveal a differently distributed correlation in the dialogue fields and between the companies.

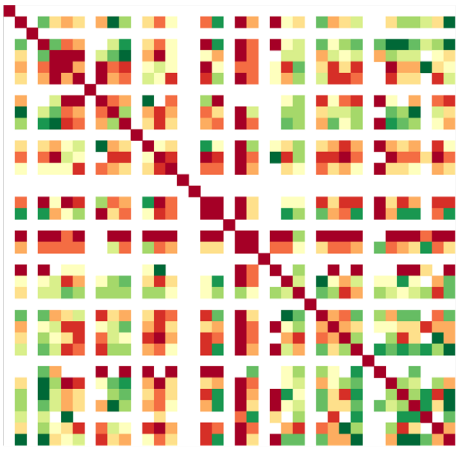

Figure 2.1: Example of heatmap of the correlation coefficient Cramérs V for all dialogue fields in a company.

To reduce the high data dimensionality resulting from the free text fields, entries with a low occurrence were viewed as outliers and thus summarized as “other”. No other adjustments were made, so the information content and the degree of detail in the data were not reduced even more. Entries from free text fields are now viewed as categorical.

Since only binary and categorical data are available, it is not necessary to standardize the data. It is also not necessary to extensively validate the data, as only positive customs declarations confirmed by customs were selected and the quality of the data is already secured. Missing data do not have to be filled in. This specific domain characteristic reduces the usually time-consuming data preparation in data science projects enormously.

2.3 Analysis

Identification of suitable analysis methods

Collaborative recommendation services are usually based on distances, for example, k-nearest neighbour or probability distributions such as those in Bayesian networks. Machine learning also offers approaches for collaborative recommendation services. Autoencoders, a type of neural network, are especially good for extracting structures from a set of data and using this learned information as a basis for making recommendations.

In models based on probabilities, the learned representations are always connected with *a posteriori* distributions. These distributions can become very complicated and also unsolvable in models with multiple layers. The sought numeric values are ultimately derived from the probability distribution. An autoencoder as an alternative calculates the deterministic numeric values directly and efficiently, and was identified as a suitable analysis method.

Autoencoders are a special form of unsupervised learning and attempt to reconstruct the given inputs as precisely as possible in the output layer. An autoencoder consists of two parts: the encoder and the decoder. The encoder projects the given data onto a latent representation of these data, and, in turn, the decoder uses this representation to reconstruct the actual data. So that the autoencoder does not only learn the identity function, it can be restricted by various techniques to learn the relevant dependencies.

The most typical restriction is to reduce the dimension of the latent representation. The autoencoder is forced to map only relevant data in this layer. Another important form is the denoising autoencoder. This variant adds noise to the data and tries to restore these partially destroyed data.

Application of analysis methods

The successive filling-out of dialogues can be viewed as an analogy for the removal of noise from data. The initial autoencoder model is based on the work of Wut et al. (2016). This work also uses implicit feedback and relies on binary data.

To have consistent comparability between test runs, the data were divided into three sets: training, test and validation. The training and test dataset was used to train the autoencoder model, and this trained model was deployed to calculate the accuracy and recall metrics from the validation dataset. In addition, the metrics were calculated several times and averaged to eliminate statistical inaccuracy. This process ensures comparability and reproducibility at all times.

The accuracy and recall metrics allow for the continuous evaluation of the initial model developed by Wu et al. (2016), which passes through the evaluation phase multiple times. The model is thus checked and expanded repeatedly.

This initial rudimentary autoencoder model implemented in Python has an accuracy of 77% and a recall of 69%, delivering good results and exceeding the metrics that are achieved with the typically used datasets such as MovieLens.

Table 2-1: The recommendation quality with different activation functions and optimisation algorithms.

	Sigmoid			ReLu		
	Epochs	*Accuracy*	*Recall*	*Epochs*	*Accuracy*	*Recall*
SGD	4999	0.777	0.688	4999	0.787	0.692
AdaGrad	3000	0.862	0.834	700	0.871	0.833
Adam	360	0.875	0.846	170	0.878	0.844

Replacing the SGD optimisation algorithm with AdaGrad can improve these metrics to 86% accuracy and 83% recall. If the Adam Optimizer is used instead of AdaGrad, comparable values can be achieved in a tenth of the epochs and thus much more efficiently.

Wut et al. (2016) use the Sigmoid function as an activation function both for the encoder and decoder. If you replace the Sigmoid function of the encoder with the known ReLu function, an accuracy of 87% and a recall of 84% can be calculated after 170 epochs.

No influence on the accuracy or the recall is seen as a result of replacing the sigmoid function of the decoder. Other influential parameters are the number of neurons in the hidden layer or the used batch size. However, a careful examination of these two parameters did not show any influence on the system's recommendation quality.

The previous observations have shown that some validation datasets constitute outliers despite the good results. The assumption is that artificially increasing the data that has noise by copying them will reduce the variance. It is possible to see that four additional datasets can raise the recall to 87% and accuracy to 91%. The degree of noise as an analogy for the number of fields already filled by the user is also to be called a parameter. When the percentage of noise is distributed equally, it also leads to consistent quality in the proposed recommendations for the different numbers of inputs made.

Comparing the results of all ten companies reveals, that the correlation between the dialogue inputs along with the amount of data available are to be emphasised as the critical characteristic in regards to the selected metrics. The sparseness of the data, which is often used as a describing characteristic, has no relevance in this context.

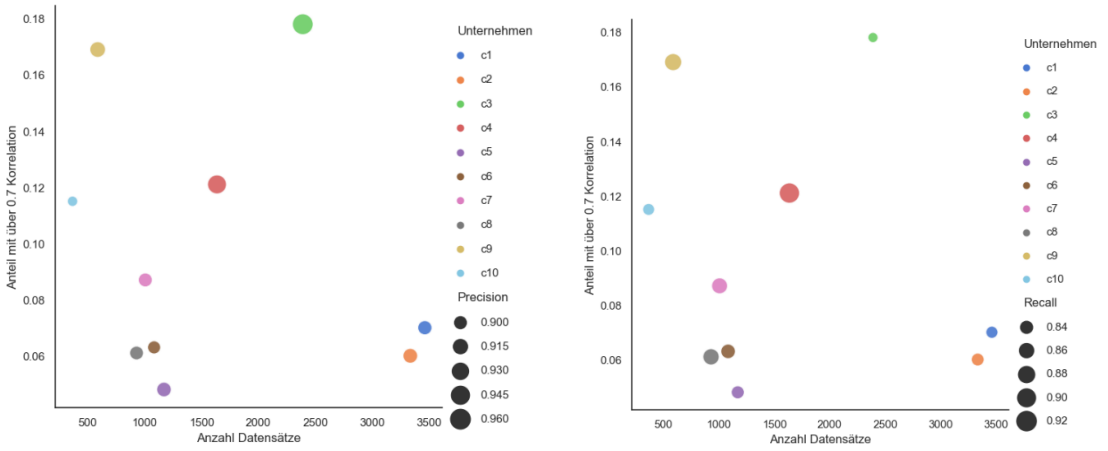

Figure 2.2: Dependence of accuracy/recall on the amount of data and the number of dialogue inputs with a Cramérs V of over 0.7.

Besides the mean correlation across all inputs, the distribution of the correlation is also important for the quality of the recommendations. The targeted examination of fewer inputs made has shown that there is a minimum set of dialogue fields that determines the remaining content of the customs declaration with an accuracy of 96% and a recall of 93%. The heatmap of the correlation can be added here as an explanatory element.

The targeted input of the most heavily correlated fields can reduce the impact of the cold start problem and accelerate the input process. An examination over multiple months showed a slight decrease in the metrics, but not to a degree that would indicate a possible overfitting. The measurable correlation between dialogue fields can also serve as an indicator for the explanatory power of the recommendations and increase the transparency of the recommendation system.

2.4 Utilisation

In the research project, the iterative and structured expansion of an already existing research basis facilitated the development of a recommendation service capable of generating recommendations with a high degree of accuracy for highly correlated dialogue fields such as those for a customs declaration. This project was therefore a success and will be transitioned to utilisation in future projects and be integrated into the system landscape.

Literature

Wu, Y., DuBois, C., Zheng, A. X. & Ester, M. (2016). Collaborative Denoising Auto-Encoders for Top-N Recommender Systems. In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WSDM '16). Association for Computing Machinery, New York, NY, USA, 153–162

3 Development of a Machine Learning Model for Materials Planning in the Supply Chain

Jonas Dieckmann, Daniel Badura

3.1 Domain and project description

SCHRAMME AG is a leading provider of dressings, band aids and bandages. The management thinks that there are qualitative optimisation potential and savings opportunities in materials planning and the resulting production processes. Management assigns an internal project manager the task of developing a model based on machine learning to plan the materials and requirements in the supply chain. Due to negative experiences in previous data science projects, it is proposed that this project should initially be developed by using a process model. The DASC-PM is chosen to ensure a structured and scientific process for project management. To gain an overview of the project assignment, the project manager initially works out various use cases that are then checked for suitability and feasibility. The suitable use cases then serve as the basis for figuring out the specific problems and the design of the project. This design is then checked again for suitability and feasibility.

Starting point and use case development

The company manually plans and then produces over 2,500 different products at the present time. In the last few quarters, they increasingly had inventory shortages for some product series, while for individual products inventories exceeded storage capacities. While the controlling department complains about rising storage costs due to the imprecise planning, the demand planners lament the insufficient amount of time for the planning. For some time, the head of the supply chain has criticised the fact that the planning is done solely manually, and the opportunities of digitalisation appear not to be taken advantage of.

One goal of the project is the development of a machine learning model where a large part of the product requirements should be planned automatically in the future, based on various influential factors. The demand planners should increasingly address the planning of important product groups and the advertising. The system should take account of seasonality, trends and market developments, and achieve planning accuracy of 75%. This means that the forecasts for quantities of each individual product should deviate from actual requirements by no more than 25%. Order histories, inventory and sales figures for customers and internal advertising plans should be used as potential data sources. Along with the inclusion of the Supply Chain department, close collaboration with Sales and IT is also expected.

The planning team in the Supply Chain department now consists of a global market demand planning team that deals with the long-term planning (6-18 months) based on market developments, product life cycles and strategic focus. In individual markets, there are local customer demand planning teams that implement the short-term materials and advertising planning (0 - 6 months) for

retail through the corresponding sales channels. The data science model to be developed should support the monthly planning cycles and quantify the need for short-term and long-term materials. The projection is then loaded into the internal planning software and should be analysed and, if need be, supplemented or corrected. The final planning quantity will ultimately be used by the factories for production planning. To take account of customer- and product-specific expertise, seasonality and experiences from the past, individual team members of the planning team should be included in the project, allocating up to 20% of their working hours to it.

An important partial aspect during the use case selection is the suitability test. The project manager tries to examine whether the project can fundamentally be classified as feasible and whether the requirements can be carried out with the available resources. Expert interviews have shown that the problem in general is very well suited for the deployment of data science and corresponding projects have already been undertaken externally and also published. The data science team confirmed that there are a sufficient number of potentially suitable methods for this project and the required data sources are available.

Finally, the project manager conducts an analysis of feasibility. It is necessary to coordinate with the IT department to check the available infrastructure and the expertise of the involved employees. The available cloud infrastructure from Microsoft and the experience of the data science team with Databricks software makes the project appear fundamentally achievable. The project risk is classified as moderate in general since the planers assume a major role as controllers in the implementation phase and the results are checked.

Project design

On the basis of the problem and specific aspects of the domains, the project manager, the head of the supply chain and a data scientist are now responsible for formally designing the project. The project objective is assumed to be an improvement in planning accuracy and a reduction in the manual processes, and is tied to the aim of developing an appropriate model for the project. According to an initial estimate, the cost framework totals EUR 650,000. A period of six months is proposed as the timeframe for the development, with an additional six months planned for process integration.

Since full planning and a description of the course of projects in the data science context are usually not possible in contrast to many other projects, the project manager solely prepares a project outline for this process with the basic cornerstones that were already indicated in the previous sections. The budget includes financial resources for 1 full-time project manager, 2 full-time data scientists and 0.5 full-time data engineers. As already mentioned, the demand planners should allocate roughly 20% of the working hours to share their expertise and experience.

The project as a whole should be handled with an agile working method and on the basis of the DASC-PM phases according to the Scrum methodology. The work is done iteratively in the areas of data procurement, analysis, utilisation and use, with the preceding and following phase moving into focus in each phase. The back-steps are especially important if gaps or problems are found in key areas and can only be solved by returning to the previous phase. The project outline is prepared visually and placed in a very visible area of the SCHRAMME AG office for all participants. Then the entire project description is checked for suitability and feasibility once again until the process moves on to the next phase.

3.2 Data procurement

Data preparation

SCHRAMME AG has a number of data sources that can be included in the automatic planning. Besides the historical sales data from the ERP system, order histories and customer data from the CRM system are options, along with inventories and marketing measures. Azure Data Factory is used to prepare a cloud-based pipeline that loads, transforms and integrates the data from various source systems. The primary basis for the automatic forecasts should be the order histories: The remaining data is used either as background information for the planning teams or to carry out cluster analyses in advance if need be. In the initial phase of the project, the individual data sources still exhibit big differences regarding quality and structure. That is why adjustments are made together with the IT and technical departments to prepare the forecasts later on a solid basis.

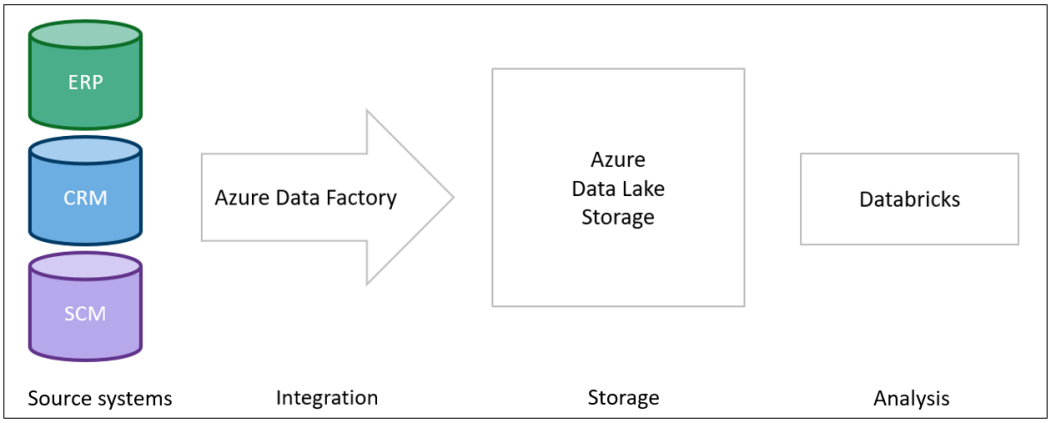

Figure 3.1: ELT data preparation process for analysis.

Data management

The data management process is automated by data engineers and done according to a daily schedule to always remain up to date. To keep the complexity reasonable, the most promising data sources are initially processed and the pipeline is then incrementally expanded with Continuous Integration / Continuous Deployment (CI/CD). After deployment, the processed data are stored in Azure Data Lake Storage where they can be used for future analysis with Azure Databricks. Data Lake also stores the backups of the prepared data and analysis results as well as other data such as protocols, quality metrics and credential structures. Writing and reading authorisations as well as plan versions also ensure that only the latest planning period can be processed so that the values from the past no longer change.

Exploratory data analysis

An important step in data preparation is the exploratory data analysis (EDA) where various statistics and visualisations are produced to start with. This results in an overview of the distributions, outliers and correlations in the data. The results of the EDA provide insights about characteristics to be taken into consideration for the next phase of the analysis. In the second step, Feature Selection and Feature Engineering are used to select the relevant characteristics or produce new features. A dimension reduction method such as a principal component analysis is applied for data with high dimensionality. The EDA provides information about the existing demand histories of SCHRAMME AG.

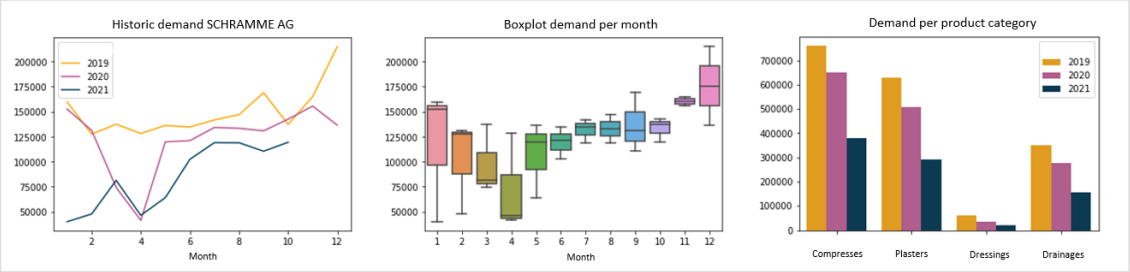

Figure 3.2: Example of results from exploratory data analysis.

3.3 Analysis

Identification of suitable analysis methods

The feasibility test at the beginning of the project made it clear that this project can and should really be solved with data science methods. The two data science employees involved initially provide an overview of the existing methods that are well suited for the existing problem.

This existing problem is part of the regression problem class in the supervised learning algorithms. Fundamentally, this is a type of time series analysis that can be expanded by additional factors or multiple regression. In connection with the key area of scientificity, the latest developments in research on comparable problems were examined. This showed that XGBoost, ARIMA, Facebook Prophet and LightGBM are frequently named methods for the problem class. A data scientist documents the corresponding advantages and disadvantages of each method and sorts them according to complexity and computational intensity. To receive the first indications on the modellability for products from SCHRAMME AG, simpler models are initially selected by the project team, which then adopts the classical exponential smoothing and ARIMA model family.

Application of analysis methods

Since multiple users are involved in the analysis process for this project, the team initially relies on a suitable notebook-based development environment in Databricks. Along the typical machine learning workflow, the code for the import and data cleaning is initially implemented. To ensure validity, the underlying dataset is ultimately divided into training, validation and test data by cross validation. The selected methods are then applied to training and validation datasets to optimise the model. In this context, attempts are also repeatedly made to optimise the parameters of processes and sensibly reduce the number of available dimensions, if need be.

The data scientists at SCHRAMME AG document the execution and validation results of the individual runs. The ARIMA family models fundamentally exhibit a better performance relative to the exponential smoothing, even if the target accuracy of 75% still cannot be achieved with a currently resulting value of 62.4%. The RMSE and MAPE metrics also show potential for optimisation.

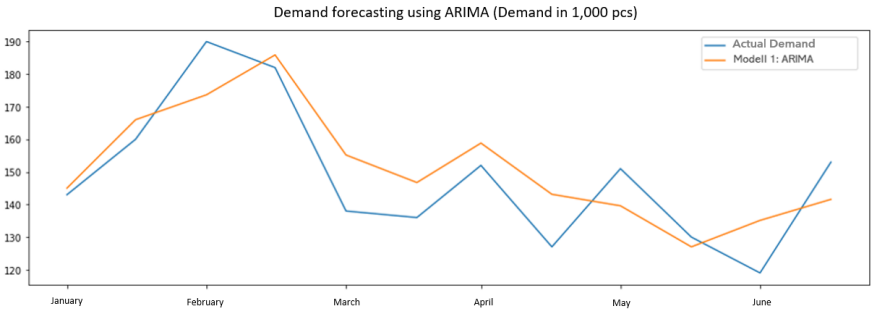

Figure 3.3: Comparison of the ARIMA forecast with actual need.

The parameter configurations and the basis for selecting the final model after the first application iteration are documented and prepared for the project manager and the head of the supply chain in a technically understandable way. What is seen in particular, is that some product groups have very unusual seasonality and certain products are generally very difficult to predict. Even if the product portfolio of SCHRAMME AG is affected somewhat less due to temporary closures (lockdowns) during the corona pandemic, a slight decline in demand for dressing products has been observed. It is assumed that less activity and transport as well as fewer accidents and injuries account for this drop. The trend can be modelled quite well in the analysis method used.

To improve the target accuracy, technically more complex methods are used in another experiment, with these methods proving to be relevant and applicable in the context of identifying suitable methods. After some iterations to optimise parameters and cross-validate, the Prophet and XGBoost methods demonstrated the highest validation results at 73.4% and 65.8%, respectively. The data scientists consider Prophet to be the most suitable method among the applied processes and determine the planning accuracy relative to the test time series. Even if the accuracy is slightly below the target value at 73.4%, a significant improvement in the planning accuracy is achieved. The MAPE is at 16.64% and the RMSE at 8,130, which implies a less absolute deviation in comparison to the RMSE in the XGBoost method (10,134). Similar to the first experiment, however, there are product groups that are very difficult to predict overall (37.2%) and negatively impact the cumulative accuracy.

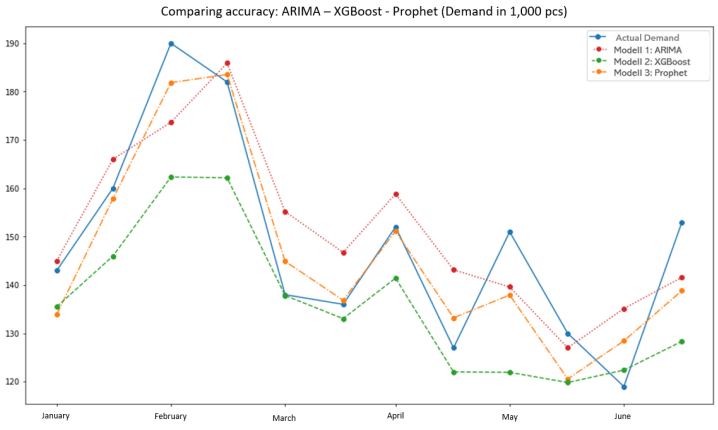

Figure 3.4: Performance comparison of various methods.

Evaluation

The results of the analyses are used as the basis for a logical evaluation and classification by the head of the supply chain and the analysts, which is organised and moderated by the project man-

ager. The adopted metrics for evaluation are the cumulative planning accuracy of all products defined in advance together with the common RMSE and MAPE metrics. It is very important for the department to have a realistic, trackable and reliable basis for determining requirements on the product level.

Table 3-1: Evaluation of the three best models.

	Accuracy	RMSE	MAPE
ARIMA	62,4 %	11.437,6	21,8%
XGBoost	65,8 %	10.134,2	19,1%
Prophet	73,4 %	8.130,8	16,6%

Table 3-2: Evaluation of the best model, distributed across product groups.

Prophet	Accuracy	RMSE	MAPE
Total	**73,4 %**	**8.130,8**	**16,6%**
Compresses	81,9 %	6.923,3	11,3 %
Dressings	86,3 %	6.027,9	9,3 %
Drainages	37,2 %	21.738,2	62,5 %
Bandages	70,7 %	8.345,4	18,4 %
Plasters	92,8 %	2.173,3	6,3%

The benchmark for planning accuracy is assumed to be the current (manually planned) median accuracy of 58% over the last two years. The evaluation of results shows that many product groups overall can be planned with a high degree of accuracy by using the data science model and vastly exceed the benchmark. However, there are also product groups that reflect similar accuracy with respect to the manual planning. It is necessary to discuss above all the product area of *drainage*, which sees much worse results with the model than in the manual planning and appears to be unsuitable for a statistical calculation of requirements with the methods used to date.

From a technical perspective, the head of the supply chain believes that it makes little sense to plan such product groups statistically since only limited planning accuracy is possible due to their specific seasonal and trend-based characteristics. She recommends the introduction of an error threshold value on a product basis to determine which products should be predicted with the model and which product groups will be removed from the modelling and still planned manually. A range slightly below the current benchmark seem to be as a suitable threshold value, since nearly as good accuracy with less manual effort from the perspective of the department is always an improvement on the way to achieving the project objective. The project leader documents the results of the evaluation with the decisions and measures adopted.

The required quantities of all selected products for the next 18 months can be documented as the analysis result after the first real modelling. This can now be utilised and integrated into the planning process of the teams.

3.4 Utilisation

The team now enters the utilisation phase of the DASC-PM for integration.

Technical-methodological preparation

It is possible to rely on the existing infrastructure for utilisation. The forecasts are loaded in the planning software IBM Planning Analytics where they are tested and reprocessed. The so-called

TurboIntegrator is used to automate the loading process that represents a central component of IBM Planning Analytics. The OLAP structure of Planning Analytics allows for the creation of flexible views where the users can personally choose their context (time reference, product groups, etc.) and adjust calculations in real time. Furthermore, the reporting software QlikSense is also integrated for more in-depth analyses. Here, the components of the time series (trends, seasonality, noise) can be visualised on the one hand and additional information such as outliers and median values can be displayed on the other hand. The final plans are loaded into the Data Lake after processing by the planning teams so they can be referenced in the future.

Ensuring technical feasibility

The forecasts themselves are automatically regenerated at the beginning of the month. The planners can make their corrections during the first four working days of the month and view the results in the planning system in real time. Since the algorithms work in a cloud environment, the computing power can be scaled, if need be. To get all processes to run automatically, changes in the data sources should be minimised. If there is a need for adjustment, the data engineer will be informed, and the interface document will be updated by recording all the information on data sources and connections. The planning and forecasting system is a mixture of the cloud (Microsoft Azure) and an on-premise system (Planning Analytics), with the planners only having active access to the on-premise structures. Credentials are awarded here so the local planners only have access to their areas, while the global planners can view all topics. After the end of the development phase, the support services are mainly handled by the IT department. In the case of complex problems, data scientists or data engineers are also consulted.

Ensuring applicability

Users of the solution are the local and global planning teams. Since members of the teams have less of a technical orientation, training sessions are held to help them interpret the forecasts and classify their quality. The user interface is also designed with a focus on clarity and understandability. Simple line and bar charts for processes and benchmarks are used, along with tables reduced to what is most important. The users are included in the development from the beginning to ensure the technical correctness and relevance and to ensure familiarity with the solution before the end of the development phase. In addition, complete documentation is drafted. The technical part of the documentation mostly builds on the interface document by demonstrating the data structures and connections, while the content part is jointly prepared with the users.

Technical preparation

To ensure that the new solution does not lose relevance or quality after a few months, work continues to be done on improvements after the completion of the first development phase, even if substantially less time is spent on it. The most important aspect of the ongoing improvement is the constant automated adjustment of the prediction model to new data. Other parts of the system still requiring manual work at the beginning are also automated over time. A change in various parameters such as the forecast horizon or threshold values for the accuracy of the prediction can be made by the planners themselves in Planning Analytics, with the model remaining flexible. Problems occurring after the release of the first version are entered via the IT ticket system and assigned to the data science area. At regular intervals, it is also checked whether the model still satisfies the expectations of the company or whether changes are necessary.

3.5 Use and summary

The transition to the use of the developed model means that the Data Science Process Model (DASC-PM) enters its last phase. As a whole, SCHRAMME AG was able to achieve the objectives it had set in the supply chain area by using the structured and holistic approach. Additional or new projects can now be derived from here. The planning processes were largely automated and supported by machine learning algorithms. The relevant stakeholders in management, finance and the supply chain were highly satisfied. After initial scepticism, the planning team itself is now also convinced by the reduction in workload and possible prioritization. However, it is also conceivable that weak points will surface during use and more iterations will be required in later phases.

The case study as a whole showed that non-linear process models in particular are advantageous for the area of data science. The DASC-PM is a suitable novel process that can also be transferred to numerous other domains and problems.

4 FLEMING Project – Predictive Maintenance for Central Components of the Medium Voltage Distribution Grid

Philipp zur Heiden

4.1 Introduction

In political circles and society, the energy and mobility transition are two of today's controversially discussed topics. The energy transition describes the shift from fossil fuels (above all coal and gas) to regenerative alternatives in the German electricity grid (Schiffer 2019; Renn and Marshall 2016). This also includes developing the distribution grid from a unidirectional grid to a bidirectional, decentralised grid with fluctuating quantities of energy generated (Bundesnetzagentur 2019). The mobility transition is understood as the shift from vehicles running on fossil fuels to alternative, no-emission drives that require a large amount of electrical energy. One of the main challenges for the distribution grid is therefore the strain on central components such as load-breaker switchgears. Here it is necessary to have an adapted maintenance strategy so that as few failures as possible occur in the medium voltage distribution grid (Hoffmann et al. 2020).

Three different maintenance strategies are popular in theory and practice (see Mobley 2002 below): reactive maintenance, preventive maintenance and predictive maintenance. Reactive maintenance means that important components are not maintained or serviced as long as they work. Only when they fail or there is a defect in the component are they replaced or repaired. While reactive maintenance exhausts the maximum potential in the service life of components, it results in failures that can mean an interruption above all in the supply of electricity in the energy grid and thus serious consequences for hospitals or the food industry, for example. Preventive maintenance calls for a regular maintenance strategy at the time when the components are checked on the basis of time periods or use metrics (e.g., number of switching operations). Some failures may be prevented, but failures are still possible and the costs are significantly higher than they are for a reactive maintenance strategy. Predictive maintenance combines the advantages of both strategies described. The permanent monitoring of condition and data science algorithms allow the status of components to be precisely mapped and predicted so that maintenance is planned when the maximum lifetime of components is achieved, but before defects occur and failures are triggered.

The goal of the project introduced in this case study is the development of a complete predictive maintenance system for central components in the medium voltage distribution grid. The *FLEMING* research project (Flexible Monitoring and Control Systems for the Energy and Mobility Turnaround in the Distribution Network through the Use of Artificial Intelligence) consists of a consortium of project partners that handle diverse areas for the successful implementation of a data science project, including project managers, domain experts, data scientists, and data engineers. These include the research institutions of the Karlsruhe Institute of Technology (KIT), the FIR Association (e.V.)

at RWTH Aachen University, the Software Innovation Campus Paderborn (SICP) and the industrial partners ABB Research Centre in Ladenburg and Heimann Sensor GmbH.

The entire *FLEMING* project can be understood as a data science project, since predictive maintenance as a strategy for the permanent monitoring and prediction of condition requires a wide range of data and sophisticated analysis methods. In this case study, the *FLEMING* project is described according to the DASC-PM, with a special focus being placed on the *domain* key area. In the case of the medium voltage distribution grid, extensive parts of the *FLEMING* project unfold in a special application area for data science projects, which should be considered separately. The following sections therefore constitute a detailed subtopic of the *domain* key area in the DASC-PM: the problem and objectives, participants and stakeholders, project organisation and resources (combined in a section) and prior experiences. In particular, they will address the answers to the questions about the important project characteristics that are provided as a question catalogue in the DASC-PM. For data science projects, this question catalogue does not have to be filled out in the detail described here, but this is recommended for time-consuming and large data science projects in particular. Finally, an outlook is offered for the project results and future processes in the *FLEMING* project.

4.2 Problem and objectives

What problem should be solved in the project?

The energy and mobility transition will soon change the energy distribution grid from a unidirectional grid to a bidirectional, decentralised network with fluctuating amounts of energy (Bundesnetzagentur 2019). This means that the central components of the medium voltage distribution grid will be subject to much greater stress in the future. For example, this may be a problem for load-breaker switchgears, although they are usually designed for a service life of up to 40 years. The *FLEMING* project should solve the problem of determining the unknown current condition of the load-breaker switchgears, i.e. their functionality and probability of failure or malfunction, and not being able to predict a change in their condition, since diverse external and in part unpredictable factors influence the condition, e.g., temperature, air quality, humidity, animal and plant population. Consequently, it is hard to predict failures in the supply of electricity.

Which objectives are being pursued in the project?

The objectives pursued in the project can be divided into multiple categories. The first category is a description of the problem to be solved, namely the current condition of the load-breaker switchgears in the medium voltage distribution grid is unclear and cannot be predicted with its currently available data. Furthermore, operators of the switchgears should see a decline in time expended and costs of maintenance by no longer relying on failures or fixed maintenance cycles, but using predictive maintenance as a maintenance strategy. For manufacturers of switchgears, this could also mean the opening up of new business segments (e.g., provision-based switchgears) and target groups.

What results are expected?

It is expected that the consortium of research and application partners will develop a prototype for a complete predictive maintenance system. This includes the generation of data for analysing the current condition and predicting future conditions based on innovative sensors, the development

of data science models, the evaluation of results and the publication of articles at scientific conferences and in journals.

How will success be measured?

The success is not to be measured by KPIs or comparable metrics on account of the primarily exploratory objective of the research project. Instead, success in the project will be determined by progress in the results produced by the individual project partners and the findings gained in the joint meetings to coordinate the project, e.g., through prototype implementation and the demonstrator.

What is the motivation for addressing the problem in a data science project?

In the case of load-breaker switchgears, it is possible that arising problems will be noticed before failure, since various internal and external factors influence and cause failures or malfunctions (e.g., a rise in temperature, air polluted by soot, or animals in the system). This may make it possible to identify failures in advance and, ideally, prevent them. The monitoring of the condition is, however, very expensive because an operator's switchgears in the medium voltage grid are spread out over a large amount of space and can be prone to errors if handled manually. For this reason, new sensors will be developed in the project to allow for the monitoring of switchgears and their condition. These data on their condition, combined with other data (e.g., from adjacent components, associated geographic data such as air pollution and weather), allow for approaches in data science to supply targeted results.

Which project-related objectives should definitely not be pursued?

The project is designed as a pre-competitive research project and does not pursue the goal of providing a service on the market that is comparable to currently available services on the market. It should definitely not lead to any complete, fully implementable solution, but offer added value in the form of prototypes for the latest research in various disciplines that are involved in the project. This includes research on machine learning algorithms, sensors, load-breaker switchgears and the design for monitoring and maintenance systems.

4.3 Domain specifics

To understand the special aspects of the *FLEMING* project, rudimentary classification of the project and the underlying problem is necessary in the context of medium voltage grids. The German electricity grid can be fundamentally divided into four different voltage levels: extra-high voltage (220-380 kV), high voltage (60-110 kV), medium voltage (6-30 kV) and low voltage (230-400 V) (BMWi 2021; Kamper 2010). While power plants and centralized generation facilities produce energy at the extra-high and high voltage levels and only a few large industrial customers purchase such energy, a large number of different producers and consumers of energy are located at the medium voltage level due to wind power and photovoltaic plants as well as industrial and office buildings. Low voltage energy is also fed into the electricity grid (e.g., with solar roofs) and consumed by private customers.

Energy must be transformed to the low voltage level at transformer stations after it is produced in extra-high or high voltage. The medium voltage ultimately ends at the substation (Siemens 2018). A substation is a collection of buildings roughly the size of a garage where energy is transformed

from medium voltage to several low voltage phases. These substations have the central components that are particularly stressed by the increasing energy and mobility transition (Hofstätter et al. 2012). This also includes the load-breaker switchgear in medium voltage.

If a failure occurs in a local substation due to the increasing load on the components, this can also mean a failure for the low voltage energy consumers behind it, so that eventually the supply of electricity to small businesses and private households would be interrupted. As electric mobility progresses, the constant supply of electrical energy will become even more important than it is today.

4.4 Participants and stakeholders

What organisational units are involved?

The ABB Research Centre Germany is leading the *FLEMING* project as consortium leader. The research institutions involved in the project are the Institute for Electrical Energy Systems and High Voltage Technology at KIT, the FIR Association (e.V.) at RWTH Aachen University and the Chair of Business Information Systems (Prof. Dr. Daniel Beverungen) and the Intelligent Systems Group at the Software Innovation Campus Paderborn (SICP). Furthermore, Heimann Sensor GmbH is involved in the project as the developer and manufacturer of the sensors. The organisational units cannot be clearly divided into management, technical department, IT, data science team, and external parties; rather, these areas are less distinct in this research project.

What organisational units are responsible?

The various working areas in the research project as illustrated in Figure 4.1 are led and managed by various project partners or a group of project partners. It is necessary to ensure that the partner with the most expertise in the respective working area runs and assumes responsibility for the content of this area as well. Furthermore, each project partner must have responsibility for at least one working area.

Who commissioned the project?

The project was commissioned by the Bundesministerium für Wirtschaft und Klimaschutz (German Federal Ministry for Economic Affairs and Climate Action, abbreviated BMWK, formerly known as Federal Ministry for Economic Affairs and Energy, BMWi).

Which stakeholder groups, in addition to the project participants, serve as input providers on technical aspects?

In the project, three companies will provide additional input on technical aspects. The utility company Städtische Werke Überlandwerke Coburg GmbH (SÜC) operates medium and low voltage grids in the city and district of Coburg, so it can be used for the analysis of requirements, evaluation and prototypical implementation of the project. Additional associated partners are Westfalen Weser Netz GmbH, which provides regional medium voltage grids in the Westphalia region, and the WestfalenWIND Group, which operates renewable energies (especially wind farms) in Westphalia.

Who is supporting/funding the project?

The project is being funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK, formerly BMWi) on the basis of the call for funding “Operating resources and components in electricity grids” as part of the 7th Energy Research Programme initiated by the German federal government.

Are there possible disruptors for the project?

Possible disruptors in the project are not known.

What are the fields of activity for an external service provider?

In the research project, no tasks are awarded to external service providers.

4.5 Project organisation

What project management method is planned?

The broad-based and spread-out consortium of partners in research and industry makes it necessary to ensure the smooth coordination and controlling of the project. This includes not only the aspects known from the DASC-PM for data science projects, but also the tasks of coordinating the time and content of the working areas, monitoring the overall project plan, coordinating reporting, balancing the interests of different partners, organising and holding project meetings and ensuring the exchange of information between the partners and working areas. No isolated project management methods can be selected for the execution of the indicated tasks; rather, different methods must be combined. In addition, the individual working areas (see Figure 4.1) are carried out and managed by the respective project partners, in part also in combination, so that the group’s own methods can be used here.

What roles are involved in the project?

Each project partner has their own role based on their work areas in the project. The role of the data scientist is assumed by employees from ABB and researchers from SICP. They also act as data engineers in combination with employees from ABB and Heimann Sensor and researchers from KIT for the thermal monitoring (WP 2) or switch drive monitoring (WP 3), the procurement of infrared sensors (WP 5), and the experimental collection of data (WP 6). ABB as a manufacturer of load-breaker switchgears and SÜC as distribution network operator are considered to be domain experts for special aspects of the application domain addressed in the *FLEMING* project. Furthermore, the FIR and SICP are to be viewed as the domain experts in the area of requirement analysis and operator expectations (WP 1) and the preparation of maintenance strategies and operating concepts (WP 8). The role of project manager for the entire project (WP 9) is handled by ABB.

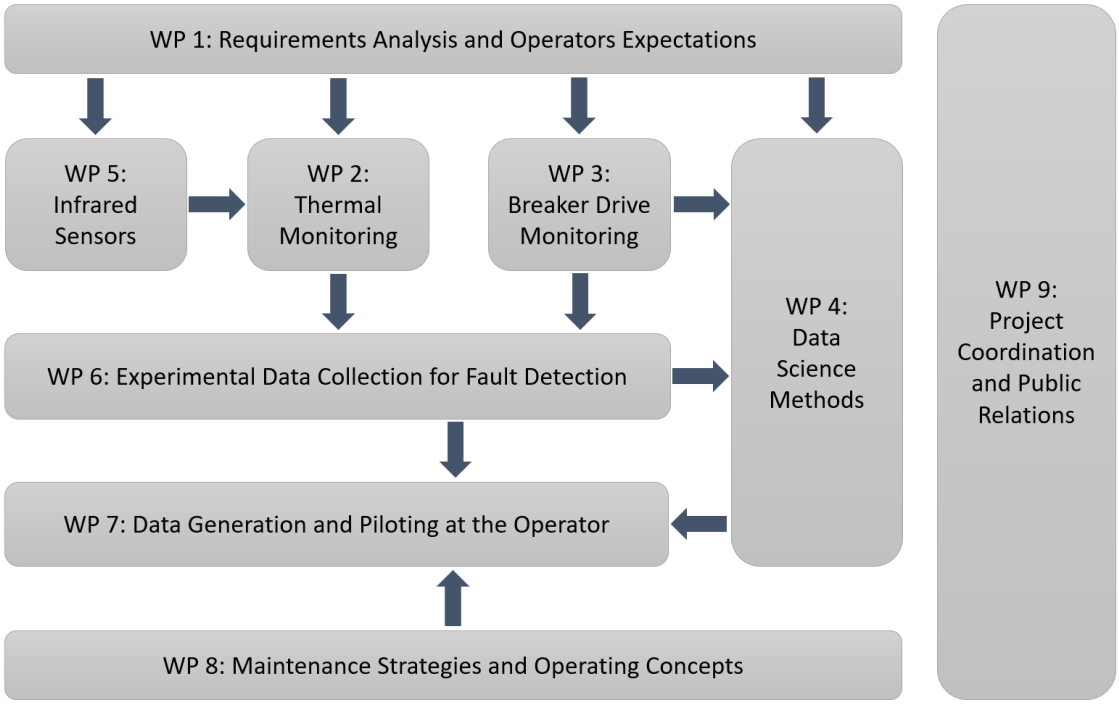

Figure 4.1: Overview of project organisation (work packages) of the FLEMING project.

What does the organisational form of the project look like?

The organisation of the *FLEMING* project is depicted schematically in Figure 4.1. The figure shows the different working areas and the connections between the working areas.

How much time is planned from the start of the project to the presentation of the result?

The project will be carried out over a period of three years. It is necessary to consider that the respective working areas (see Figure 4.1) have more detailed schedules for their individual work packages. For example, WP 1 starts right away, since the results are very relevant for all the following working areas. Above all, it is important to know what requirements the operators of switchgears will be expected to meet in a system for predictive maintenance. By contrast, the maintenance strategies and operator concepts (WP 8) are to be prepared when the requirements and operator expectations
(WP 1) are known, since they build on them.

What competencies do the project members have?

The industry and research partners are experts in their respective areas and have the competencies necessary to handle the project. This is evident in the scientific partners from their research projects and publications, while the industrial partners have gained acceptance on the market with their growth.

What financial conditions are there?

The project is commissioned by the Federal Ministry for Economic Affairs and Climate Action (BMWK, formerly BMWi). The funds to be used are predefined by each project partner and broken down by person-months, IT infrastructure, travel funds, and other items.

How much lead time exists until the project must begin?

The project was approved by the Federal Ministry for Economic Affairs and Climate Action (BMWK, formerly BMWi) in June 2019, and work began in September 2019. The lead time of three months is less relevant, however, because the time required for orientation in each topic is planned for most working areas.

4.6 Prior experience

What solutions are already available?

Predictive maintenance is not a new technology in the area of energy distribution grids, although it is being used for the first time in this project. In the energy industry, the potential that predictive maintenance can offer as a maintenance strategy has already been successfully deployed. For example, there is already research work on the use of predictive maintenance on wind turbines (Canizo et al. 2017) and in nuclear power plants (Hashemian 2011). Therefore, the focus in this project is also on viewing existing data-driven methods and selecting suitable methods for the application domain of load-breaker switchgears in medium voltage. The first solutions for the monitoring of switch actuators and infrared sensor technology already exist, but they are not optimised for the aforementioned purpose either.

What experiences were collected in similar projects in the past?

Some individual project partners have gained experience in technical areas through previous projects that were partially funded publicly and partially handled internally. To date, no joint project has been worked on in the consortium.

Where were the difficulties in past projects?

Since there have been no joint projects in the past, it is not known what difficulties will affect the entire project or multiple partners.

What organisational units have prior experience with data science?

The ABB Research Centre in Ladenburg, especially the groups involved in AB 4, have already gained intensive experience with data science in internal and external projects. To date, the SICP has been able to contribute significant and diverse experience to the project in the area of data science, in particular with the Intelligent Systems working group. The FIR consists of experts for data-driven business models so that it can provide advice on data science tasks, as can the Chair of Business Information Systems at the SICP. Heimann Sensors has already gained some experience in data science methods such as person detection and hotspot detection through applications based on sensors. KIT, which is involved in the *FLEMING* project, has less previous experience in data science activities than the other project partners. It is similar with SÜC as an operator of distribution networks.

4.7 Conclusion and outlook

The *FLEMING* project constitutes a large data science project involving multiple industrial company and research institutions. In many places, the boundaries between the classic roles in the DASC-PM are not clearly evident and they overlap. Nonetheless, the DASC-PM offers a sensible way to

execute and structure the entire research project. The questions asked in the course of this case study show the challenges that were identified in the project description and incorporated into a project with a diversified team.

As in other data science projects, there may be special circumstances that substantially complicate the execution of the project. In the case of the *FLEMING* project, the COVID-19 pandemic should be mentioned in particular, as it was not foreseeable at the beginning of the project. The pandemic greatly complicated the purchase of load-breaker switchgears for the experimental generation of data (AB 6) and prevented the working areas dependent on these data from generating the results as planned. Since data science projects cannot produce meaningful findings without data as a rule (see the key area of *data*), particular caution is required here.

The results achieved in the project to date range from an analysis of the operator requirements (zur Heiden und Priefer 2021) to the further development of AutoML as an analysis method for calculating remaining lifetimes of components (Tornede et al. 2021) and a complete design of the predictive maintenance system (zur Heiden et al. 2022). The project was still running at the time this article on the case study was composed, so the final results are open and it is not possible to provide an evaluation of whether the project achieved the objectives laid out at the beginning.

4.8 Funding information

This article describes the FLEMING research project, which is funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK, formerly BMWK) under Grant no. 03E16012F. We would like to thank the BMWi and the project manager Jülich (PtJ) for their support.

Literature

ABB (2021): Luftisolierte Schaltanlagen. Available online at https://new.abb.com/medium-voltage/de/mittelspannungs-schaltanlagen/luftisolierte-mittelspannungsschaltanlagen, last updated on 8 November 2021, last checked on 8 November 2021.

BMWi [Federal Ministry for Economic Affairs and Energy] (2021): Das deutsche Stromnetz. Available online at https://www.bmwi.de/Redaktion/DE/Infografiken/Energie/abbildung-das-deutsche-stromnetz.html, last checked on 26 August 2021.

Bundesnetzagentur (2019): Bericht - Zustand und Ausbau der Verteilernetze 2018. Bonn.

Canizo, Mikel; Onieva, Enrique; Conde, Angel; Charramendieta, Santiago & Trujillo, Salvador (2017): Real-time predictive maintenance for wind turbines using Big Data frameworks. In: 2017 ieee international conference on prognostics and health management (icphm). IEEE, pp. 70–77.

Hashemian, H. M. (2011): State-of-the-Art Predictive Maintenance Techniques. In: *IEEE Trans. Instrum. Meas.* 60 (1), pp. 226–236. DOI: 10.1109/tim.2010.2047662.

Hoffmann, Martin W.; Wildermuth, Stephan; Gitzel, Ralf; Boyaci, Aydin; Gebhardt, Jörg & Kaul, Holger et al. (2020): Integration of novel sensors and machine learning for predictive maintenance in medium voltage switchgear to enable the energy and mobility revolutions. In: *Sensors* 20 (7), p. 2099.

Hofstätter, Frank; Weber, Thomas & Rabanus, Sebastian (2012): Intelligente Ortsnetzstationen als Alternative zum Netzausbau. In: *etz elektrotechnik & automation* (4).

Kamper, Andreas (2010): Dezentrales Lastmanagement zum Ausgleich kurzfristiger Abweichungen im Stromnetz: KIT Scientific Publishing.

Mobley, R. Keith (2002): An introduction to predictive maintenance: Elsevier.

Renn, Ortwin & Marshall, Jonathan Paul (2016): Coal, nuclear and renewable energy policies in Germany: From the 1950s to the “Energiewende”. In: *Energy Policy* 99, pp. 224–232. DOI: 10.1016/j.enpol.2016.05.004.

Schiffer, Hans-Wilhelm (2019): Zielvorgaben und staatliche Strategien für eine nachhaltige Energieversorgung. In: *Wirtschaftsdienst* 99 (2), pp. 141–147. DOI: 10.1007/s10273-019-2408-x.

Siemens, A. G. (2018): Planung der elektrischen Energieverteilung-Technische Grundlagen. Erlangen, Germany.

Tornede, Tanja; Tornede, Alexander; Wever, Marcel & Hüllermeier, Eyke (2021): Coevolution of remaining useful lifetime estimation pipelines for automated predictive maintenance. In: Francisco Chicano (ed.): Proceedings of the Genetic and Evolutionary Computation Conference. GECCO '21: Genetic and Evolutionary Computation Conference. Lille France, 10 07 2021 14 07 2021. ACM Special Interest Group on Genetic and Evolutionary Computation. New York, NY, United States: Association for Computing Machinery (ACM Digital Library), pp. 368–376.

zur Heiden, Philipp & Priefer, Jennifer (2021): Transitioning to Condition-Based Maintenance on the Distribution Grid: Deriving Design Principles from a Qualitative Study. In: Michael H. Breitner, Sebastian Lehnhoff, Astrid Nieße, Philipp Staudt, Christof Weinhardt und Oliver Werth (eds.): Energy Informatics and Electro Mobility ICT. Community Workshop Proceedings, Pre-Conference 16th International Congress on Wirtschaftsinformatik Duisburg-Essen University: BIS-Verlag der Carl von Ossietzky Universität Oldenburg, pp. 72–87. Available online at https://oops.uni-oldenburg.de/5084/.

zur Heiden, Philipp; Priefer, Jennifer & Beverungen, Daniel (2022): Utilizing Geographic Information Systems for Condition-Based Maintenance on the Energy Distribution Grid. In: *Proceedings of the 55th Hawaii International Conference on System Sciences*.

5 The Road to the Project

Florian Schwade, Heiko Rohde

5.1 Background and description of the case study company

ACM Technology AG does business as an automotive supplier (A), manufacturer and supplier of consumer electronics (C) and medical technology equipment (M). The company has roughly 3,500 employees and manufactures primarily in the categories of *air filters*, *ventilators* and *membranes*. The products are used in automotive air conditioning systems (automotive industry), fans and coolers for servers, personal computers and laptops (consumer electronics) and medical respirators (medical technology), for example. On the basis of the broad business segment, there are challenges for the company in the assignment of materials and material requirements planning. Especially in the medical technology business segment, there are frequently express orders whose processing must be given preference, so that the planning of production and material requirements can also change within one day. A sharply fluctuating number of orders from the automotive industry and their frequent cancelation make the planning more difficult. In part due to delivery bottlenecks for important components such as chips and other raw materials, ACM Technology AG has seen, for quite some time, a major need to improve material requirements planning and the allocation of (scarce) materials to satisfy the needs of all business segments. At the end of 2020, the company started to make attempts at optimisation by using the processes in operations research and data science methods without prior experience or expertise. Above all, the complexity and the management of this data science project was underestimated so good approaches could not be implemented successfully. There were challenges due to a lack of expertise on the project team, insufficient support in the organisation and for technological aspects. This project was classified as a failure and the reasons for this were addressed. As a result, a new and carefully picked project team was put together.

The project should now be implemented in a structured manner on the basis of the DASC-PM. The persons involved agreed on the formulation of the questions and the objectives associated with them. This first step was clearly documented in the form of a project description. The preparation of the project description is the subject matter of this case study and is based on the project description phase of the DASC-PM. This can be viewed as the trigger for this data science project.

To start the definition of the project more successfully, the first measure was to identify competence profiles (based on the dimensions of *mathematics/statistics*, *information technology*, *application area*, *communication*, *strategy and management*) while putting together the project team to ensure that the necessary expertise and roles (especially domain expert, data scientist, data engineer and project manager) were represented in a balanced way in the project team. A consequence of this was that not only data scientists were involved in the project. Rather, it also included experts or product owners of the most important products with expertise and domain knowledge in the business segments, target markets and production. Since the CIO and COO were direct product sponsors, the support of management was ensured. The project team consisted of a total of nine participants:

Project manager

- SCRUM master
- Product owner – breathing apparatus (domain expert)
- Production manager – 3D manufacturing (domain expert)
- Head of sales (domain expert)
- Head of procurement (domain expert)
- Data scientist
- Head of IT (technical support)
- Compliance officer (compliance support)

5.2 Development of the use cases

After the project team was put together, the members were invited to a collective focus group. The focus group was headed by an employee who is familiar with the details of the project, but was not involved. Since this employee has a PhD and the focus group interviews had already been conducted as a research method, she had the appropriate expertise in their design and execution. The decision to elect a focus group as a method for fleshing out the details of the project was made, since focus groups aim for diversity in ideas and opinions and regard different levels of knowledge or expertise equally. Furthermore, the participants in a focus group were ideally suited on account of the diversity of their roles and eleven participants are considered to be the ideal size of a focus group.

The main objective of the focus group was to work out the use cases in detail and identify the actual problems and challenges, so that they could work towards defining the project. Since the project involves multiple subprojects, the characteristics of the "triggers" should be addressed in greater detail and questioned. The focus was on the *technical purpose*, *objectives*, *application frameworks* and *complexity* in particular. The issue of the required data was also addressed. However, that issue is not addressed in this case study.

The results of the focus group's most important findings for the *project description* phase are summarised below. The creative elements in the focus group identified a number of possible use cases. This case study focuses on the three aspects of material requirements planning, setup time optimisation and logistics. The other possible use cases are disregarded.

One of the greatest challenges for ACM Technology AG is the material requirements planning. As described above in the background, the material requirements planning is based on the changing requirements of customers and the current widespread scarcity of resources (e.g., microchips). The scarcity of resources often prevents the full processing of all orders immediately. This frequently results in triage. The company must decide which orders have to be prioritised and which orders can be handled at which volume. The challenge is made even greater by the ongoing COVID-19 pandemic, which has caused a sharp rise in express orders at short notice in the area of medical technology (e.g., ventilators). Such express orders make it necessary to adjust the production planning for the day in real time at short notice, sometimes within a few hours. In turn, the consequence

of this is that the production must be reprioritised, which was not foreseen in the capacity utilisation planning. Another part of the project is the planning, calculation, prediction and simulation of capacities (human, machine, raw materials) with changing parameters.

Setup time optimisation was identified as one of the main new problems for ACM Technology AG. A few months ago, the company converted large parts of the production lines to 3D printing production and has still not gained much experience with the optimisation of the product sequence. A particular challenge here is changing the necessary filaments and partially necessary downtimes due to cleaning. They are hoping data science will achieve improvements in production, primarily with regard to the production sequence and setup optimisation to use the machines as efficiently as possible and minimise downtimes.

Finally, another possible objective is to optimise logistics. In this area, the members of the focus group are looking for improvements in the logistics, especially in the commissioning of goods. The consumption of packaging is a secondary concern. The most important factor is when deliveries should be shipped, especially in the case of express orders. The main questions here are: Should individual products for express orders be sent individually after completion or consolidated as partial or complete deliveries? Since the previously mentioned aspects are major challenges, it was decided that the logistics optimisation should be postponed to a follow-up project.

After a summary of the focus group results, management was encouraged to let a small part of the project be implemented in a focussed manner and as proof of concept, since big potential for an immense competitive advantage was seen in improvements. The proof of concept was supposed to allow for the quick development of a first prototype with few resources, which would then be gradually developed into a lighthouse project throughout the company. After a final discussion of the possible risks, it was decided that a clean prototype and no “quick & dirty” solution should be developed.

In the preparation of the focus group, the moderator was also included in the project team to represent the perspective of science even more prominently in the team.

That is why the team now approached the problems by figuring out whether they should primarily have a strategic perspective or an operative perspective, whether they are purely data-driven and how much they impact the market environment or achieve an impact. The classification was supposed to clarify the question of whether it is fundamentally a data science project. Two aspects were examined in particular here and verified with external data scientists – first the data availability of suitable datasets in the company, or from connected data sources, and second whether the use of data science appears to be superior to classical analytics methods. In addition, a clear objective should be defined this time, so that a result will be achieved even in the event that a problem does not lead to the desired result.

The first use case in material planning was initially viewed operatively, since a strong operative impact was noticed as a result of production adjustments at short notice. However, the domain experts in the focus group were able to quickly show that the few raw materials for all production processes are the basis so that the end material is mainly dependent on the manufacturing process with its recipes and less on special materials. There are additives, but they hardly play a role in total. This would move the topic more into the area of “market environment and strategy”, since the questions here refer to the supply chain, the strategic selection of suppliers and the long-term securing of the availability of production raw materials.

The second use case for production optimisation in 3D printing also has a clear strategic focus and very innovative dimensions. Considering the process should not only facilitate process optimisation, but also open up new business fields such as production at the customer.

The third topic, which should be addressed in a separate project, is to be viewed as an operative topic. Logistics optimisation is related to the customer and has a certain degree of relevance for the market, but more and more companies are relying on sustainability and also expect this from their suppliers. The combining of deliveries is not viewed as a major risk, provided that the delivery targets are still met. The domain experts also see additional potential here and a potential impact extending all the way down to production planning.

5.3 Suitability test and ensuring feasibility

The accompanying *suitability test* was very exciting in this project. The suitability of the problems was viewed critically by management, since the first part of the project was not successful and the problems only slightly differ from those in the first attempt.

In the preliminary study for the project, it was necessary to ensure the feasibility of the use cases. To gain a better picture, comparable challenges, as described in publications and online sources, should be compared with the problems in the project to find out whether ACM is subject to restrictions that play a role in the execution of the project.

The probably most important domain-specific requirements are also correlated with the problems. It is necessary to have a high level of technical process knowledge for the problems in the area of production for medical products, since batch tracking, material requirements through regulation and special purity processes are used here. For example, air filters for the automotive sector cannot simply be converted into medical products without paying attention to the specific manufacturing parameters, even if the products range from very similar to identical. However, this technical knowledge is available from the many years of experience in the company, so that it can be directly integrated into the project. The risk of having too little expertise for a project can be ruled out here.

Another domain-specific aspect resulted from the COVID-19 pandemic: political pressure on companies in the medical sector. Management was regularly invited by politicians responsible for health policy and sought out for help to quickly supply, together with other manufacturers, the necessary equipment such as air filters. Here, too, ACM was caught between different interest groups that were important to serve. The consequences of another intensification of the situation could not be ruled out, but the direct project risk can be judged as low or the management views data science as an opportunity here.

However, in general, the manufacturing process in 3D printing technology is much more exciting than the special manufacturing in medical technology. This was classified as strategic, in the hope that production could be sustainably optimised in all segments. In the technical department, they are even considering the outsourcing of production to a customer location so that the product will be printed locally.

5.4 Procedure for selecting use cases

The next step is to evaluate the use cases to identify project-relevant ones and determine suitability in general. The DASC-PM distinguishes between the two accompanying tasks of *suitability test* and *ensuring feasibility*, but the stakeholders in this project agreed to use a nine field matrix to incorporate all the results of the sub-tasks into an evaluation. The assessment of feasibility is depicted on the x-axis, and the relevance for the company on the y-axis, each in the categories of low, medium and high (see Figure 5.1).

Although this process does not follow any strictly scientific standards, the participants in the focus group were able to show which use case was to be prioritised over another. Furthermore, this matrix offers the opportunity to evaluate aspects such as IT infrastructure, the assessment of risks or possibly the later integration in applications in the current corporate context, which follows the accompanying task of *ensuring feasibility*. Acceptance in the company, from an organisational point of view, can also help with feasibility. In this context, the focus group's initial resistance had to be overcome with experts. With the support of management, the focus was placed on the technical infrastructure and data availability.

The relevance for the company was defined through the degree of innovation in a solution and the opportunity to increase revenue or save costs, since the greater influence of management here was noticed, which resulted in an internal prioritisation of the use cases.

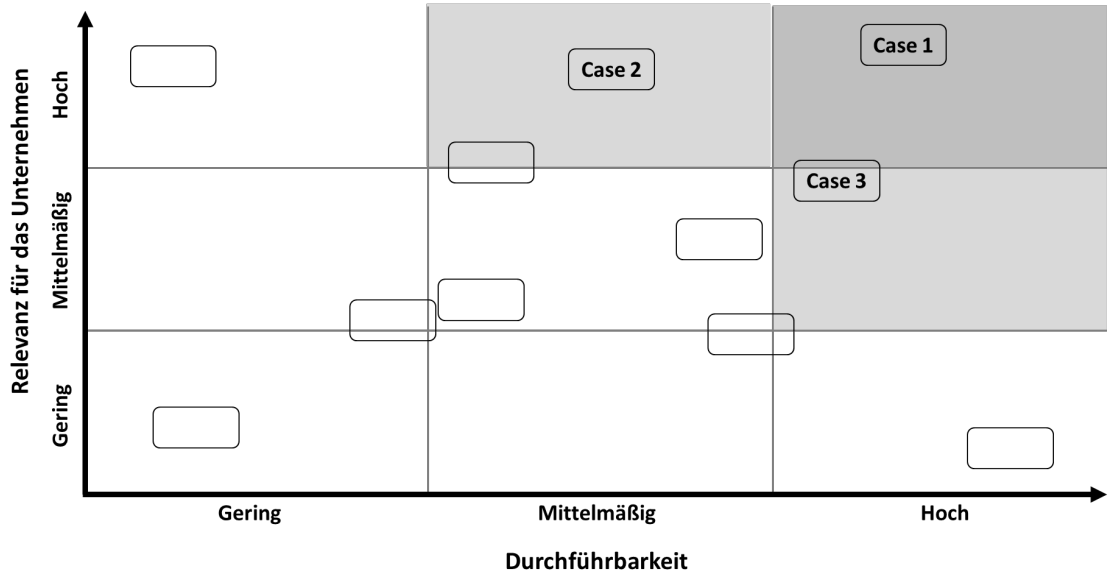

Figure 5.1: *Classification of use cases according to feasibility and relevance*

For this project, the use cases could be defined, along with other potential ones for future projects. A simple matrix representation could have been supplemented by additional scientific methods, if a trivial selection had not been possible.

5.5 Project design

After the use cases were selected and the tests were performed, the following project preparation report with a further description of the project design was submitted for internal presentation to the two project sponsors (CIO & COO).

Project report

Project title:
Material requirements planning & setup optimisation through data science

Motivation:
The company frequently faces the challenge of having to prioritise orders. This applies in particular to express orders in the medical technology sector. In this context, scarce resources are also a challenge, such as the ongoing shortage of chips. Therefore, the company would also like to be better equipped for resource shortages in the future. Closely tied to the challenges is setup optimisation. The company is a new player in 3D printing production and would like to improve the setup times for 3D printers due to the changing of filaments.

Objectives:
The project objectives are therefore the improvement of material requirements planning and setup times for 3D printers. The objectives can be measured with the following key performance indicators:

- Quantity of used materials in relation to production output
- Quantity of scrap in relation to resources deployed
- Utilisation of production capacities (people and machines)
- Downtimes
- Changes in production after receipt of express orders

Innovation:
For ACM Technology AG, the innovation in the project consists of optimising the material requirements planning and the setup times for printers and machines in production by applying data science. Especially with regard to setup times and the optimised changing of filaments, a competitive advantage is expected relative to competitors. Another aspect of the innovation is the expansion of the data science competencies in the company, which will be beneficial for future data science projects.

Possible uses:
The primary possible uses of the project results can be found in the outlined areas of material requirements planning and setup time optimisation. The content-related, technical and methodological data science competencies gained can be expanded to other areas after the end of the project. It is expected that analysis processes developed in the project can also be applied for the improvement of the sales and operations planning (S&OP). Since the aspect of setup time optimisation for the pilot project is specifically focused on 3D printing to start with, it is expected that especially the analyses and findings in this regard can be expanded to all production steps and locations throughout production.

Milestones and schedule:

The plan is to achieve a "quick win" with the project within one month in the form of a demonstrated prototype. The prototype should demonstrate the technical feasibility and make it possible to gain the first rudimentary findings. The total period for the project is set to six months. To achieve these goals, the project uses the DASC-PM and continues to refer to agile methods for the execution and management of the project. The milestones result from the key areas of the DASC-PM. The project should start on 3 January 2022. The milestones outlined below focus on the achievement of the proof of concept after one month. Since an agile approach is adopted for work in the project, it is also possible to work iteratively on already completed milestones after the development of the prototype, e.g. if the findings from the application of the prototype show that new data sources must be included or data must continue to be prepared.

Table 5-1: Milestones in the Project

Milestones	**Schedule**
Definition of the project	31 December 2021
Identification and selection of the original data sources	7 January 2022
Data preparation	24 January 2022
Identification and evaluation of suitable analysis methods	7 January 2022
Development of analysis methods	31 January 2022
Ensuring technical feasibility	31 January 2022
Technical-methodological preparation	31 May 2022
Technical preparation	30 June 2022

Technical requirements:

The technical requirements for the project cannot be fully assessed in advance, since they depend in part on the selected analysis methods and the datasets. It was decided that the prototype will be implemented in Python, so that a development environment must be present in Python. Furthermore, the members of the project team need high-performing laptops for this. Later, the cloud can be used to ensure better availability and integration in the IT landscape. However, we will refrain from doing this for the time being, because no cloud architecture is available.

Project team:

The project team consists of nine people. The project involves the following participants and roles:

- Project manager
- SCRUM master
- Product owner – breathing apparatus (domain expert)
- Production manager – 3D manufacturing (domain expert)
- Head of sales (domain expert)
- Head of procurement (domain expert)
- Data scientist
- Head of IT (technical support)
- Compliance officer (compliance support)

The selection of participants and the assignment of roles is based on the recommendations of the DASC-PM. The composition of the project team shows that the project was not designed as an IT project. Importance was attached in particular to using the domain expertise of the involved technical departments to extract as much value as possible from the project. Although the "business" roles have a disproportionately high weighting, the core role in the project is assigned to the data scientist.

Cost estimate:

The project costs can be broken down into staff and costs of materials. The purchase of the development environment and the necessary hardware is required for the cost of materials. This should not be a major investment. Furthermore, external consulting costs for a total of about 80 project days are planned. The first cost estimate here totals €100,000-120,000. Above all, external data science know-how is necessary for the project and must be purchased externally, until the required internal resources are trained.

150 project days are planned for internal resources. They are broken down as follows:

Table 5-2: Composition of the project team incl. person-days

Project team	**Days**
Product management / owner	30 days
Scrum master	15 days
Domain experts (3 persons, in addition to product owner)	40 days
Data scientist (1 person, without external parties)	50 days
Support roles (IT, Compliance)	15 days

A crude estimate of the internal costs is that they will amount to €50,000 for the project period of 150 days.

Project outline

Scrum was chosen for the internal project management after approval of the project application to organise the project management as part of the DASC-PM.

In this regard, a Sprint cycle of two weeks was initially chosen in the POC phase and then expanded to three weeks afterwards in order to carry out all phases of the DASC-PM once for the implementation of an increment, i.e. a finished, useable case. The Scrum events were connected with the parts of the DASC-PM. In a review, the definition of the project must be checked again and possibly adapted so that a new cycle of the DASC-PM can start with the Sprint planning.

The agile work method should facilitate the identification of the biggest problem in a data science project at an early stage, namely the possibility that no useable result is possible after the technical definition of the objective. The reasons for this may be diverse, but they can be identified more quickly through short cycles.

However, a minor deviation to the classical scrum was implemented. The definition of done should say that the result, i.e. the increment, of a sprint should be possible to compile and execute, but additional "definitions of done" were created for the stories or tasks, which should relate to the individual phases of the DASC-PM.

For example, the *data procurement* phase requires the data to be extracted from the data source, transformed and the quality checked by passing through technical validation. Two definitions were

prepared for the *analysis* phase, which can be used independently of each other. If test data are available, then a test of the model is possible against a validation dataset. Alternatively, the analysis artefact is validated by technical areas or stakeholders. No other definitions were initially developed for the utilisation and use. Others could be developed with the characteristics of the DASC-PM phases.

Some backlog items were generated for the project start, ranging from the technical process definition to the extraction and transformation of data as well as the possible analysis method. Since a complete backlog would be too long here, a few examples for the material requirements planning are provided. The current process orders from the ERP system should be extracted to compare them with the historical sales. Furthermore, it is necessary to see whether there were events in the past that have led to replanning and may exhibit a certain regularity or make early identification possible on the basis of characteristics. The domain experts should fundamentally describe the sales flows in a better procedural way and identify the data characteristics. The discussions in the previous project showed, for example, that the type of order may cause the materials to differ. A long-term contract is not as susceptible to changes as a triggered order, although there may be material-specific deviations that can flow into optimisation.

In summary, the DASC-PM can be well combined with agile methods, even if the methods must be coordinated with each other. The most important parts of a data science project are a motivated team, support in the organisation, a coordinated and methodical approach (e.g., DASC-PM) and an achievable use case that does not start by “reaching for the stars”.

6 Face Mask Detection

René Theuerkauf, Tony Franke

6.1 Starting Point

The company BIG KAUF AG would like to order software to identify the wearing of a face covering so it can efficiently design the entry check for its stores. The practice to date has consisted of having employees in the stores monitor the entrances of the building during all opening hours. The number of entry points had to be substantially reduced because of the available staff resources. By using software, they should again increase the number of entrances, if need be, while maintaining the same level of staff. If this is implemented, it will be necessary to ensure sufficient reliability on account of the requirements imposed by the competent health authority. The company should be given an application that automatically triggers/prompts an action in the event of violations and allows for remote monitoring. With use of the DASC-PM, the development of the software should be carried out. A project horizon of three months has been agreed. The budget includes financial resources for an 0.25 full-time project manager, 2 full-time data scientists and 0.5 full-time data engineers. The project will be accepted by the customer.

6.2 Introduction

Due to the ongoing corona pandemic and the associated hygiene requirements, branches with public traffic are required to check the wearing of a face covering by their customers, e.g., in the form of a medical mask, and ensure other compliance measures. The corresponding check can be done upon entry into the building. For companies with larger branches that have multiple entrances, this is an expensive requirement. It may entail much higher staff deployment for the implementation of these controls. To counter this, it is possible to use software solutions in the area of data science. This case study describes the preparation of a software prototype based on the DASC-PM, which was implemented by students as a research project in a seminar of the Martin Luther University Halle-Wittenberg. The five successive phases “project order”, “data provision”; “analysis”, “deployment” and “application” of the process model were planned, and the first four were implemented. The analysis phase is exemplified. Following, the phases are described in detail below and the applicability of DASC-PM is demonstrated on real economy projects.

6.3 Project Order

The company BIG KAUF AG would like to order a software solution for automated mask recognition in live video feeds. This will be used at the entrances to the company’s branches and instruct customers automatically to comply with the current hygiene requirements if they are not wearing a medical mask. The budget includes financial resources for 0.25 full-time project manager, 2 full-time data scientists and 0.5 full-time data engineers for a three-month horizon. The project objective was defined as follows by BIG KAUF AG. The prepared software should be 80% accurate at differentiating between customers with a mask and customers without one or using one incorrectly.

A visual and acoustic signal should warn the customer about this in the event that false use is identified.

6.4 Data Provision

The data were procured in two steps during the project. The first step was to search for a free data source or database that included pictures already labelled as a person in the picture wearing a mask or a person without a mask. Since such a basis for the data could not be found, the required pictures had to be obtained by the project participants. The pictures were collected by using various search engines. The search engines Google, Bing and Pexels were used. In all the search engines, the following search terms were entered: “person with medical mask”, “person with mask”, “covid masked face”, “face”, “person”. Due to the cost constraints, application programming interfaces provided by Google and Bing could not be relied on. Instead, a Selenium-based web crawler was used so the pictures could be searched for and downloaded automatically in these search engines.
However, the interface provided by Pexels could be used for its search engine. With the used search terms, the pictures found were automatically classified into the respective categories (mask, no mask). This classification was then checked manually and adjusted as necessary.
A total of 500 pictures were collected. Then the data were expanded by duplicating the pictures with defined adjustments. These included mirroring, rotating, changing saturations, and adding Gaussian noise to previously collected images. This process made it possible to improve the underlying data by incorporating the changing circumstances of lighting and movement in a store. This resulted in a dataset of 5,000 pictures, which forms the basis for the training of the method described in the following.

6.5 Analysis

The next step, the analysis, consisted of developing a process for classifying the data on whether a person is wearing a mask or not from the live video feed.

Identification of suitable analysis methods

At the beginning of this phase, possible procedures, and methods for the realisation of the classification are reviewed and evaluated. Experiences from previous data science projects were primarily relied on to select the processes to be evaluated. Another evaluation was whether a new method must be developed, or an already existing concept can be used.

Application of analysis methods

In the viewing of potential methods, the Support Vector Machine, Random Forest, XGBoost, Naive Bayes, Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks were initially examined more closely and implemented on a test basis. Common implementations with standard parameters were used. The evaluation was handled with statistical quality criteria such as accuracy, sensitivity, and F1-Score. After the evaluation of the tests, CNN was selected as the most promising candidate. Then the existing networks for object identification based on images were viewed. The second step was to test five already trained networks provided by Google. These are DenseNet, SqueezeNet, ResNet_V2, Mobilnet_V1 and Mobilenet_V2. All networks were tested for their time per classification process and TOP 5 accuracy. Table 6-1 shows the results of the tests.

Table 6-1: Evaluation of image classification models.

CNN	Classification time	TOP 5 accuracy
DenseNet	195 ms	85%
SqueezeNet	36 ms	72%
ResNet_V2	526 ms	93%
Mobilenet_V1	1.7 ms	70.2%
Mobilenet_V2	17.5 ms	90.6%

Evaluation

To determine the suitable network for the project, an initial selection was made based on the TOP 5 accuracy. The two networks with the highest accuracy were selected: ResNet_V2 (93%) and Mobilenet_V2 (90.6%). Then the classification time was considered in both networks. This is relevant because the company would like to use mask recognition in connection with a live video feed. ResNet_V2 offers the greatest accuracy, but a classification requires around 0.5 seconds. That is not practical when using live video feed. Therefore, Mobilnet_V2 was used as the basis for mask recognition. It offers slightly worse accuracy, but the classification time is only 17.5 milliseconds and is therefore suited for mask recognition in video feeds with HD quality. The existing Mobilenet_V2 was then trained again and adjusted to the application case through another layer. The model was evaluated on the basis of sensitivity and the F1-Score of the classification results on the test data.

6.6 Deployment and Application

In the context of utilisation, the technical and methodological approach was made available and the applicability was ensured. To ensure this, two additional components were created. Firstly, the classification result with the desired error tolerance was made recognisable on the live video feed by using different coloured boxes around the face of the person to be classified. The box around the face of a person with a mask was rendered in green. Red signalled that the algorithm decided no mask was being worn. To simplify use, an application was designed and implemented for the customer. This includes both a graphic user interface and the possibility of manually adjusting the threshold of the classification decision of the developed method. In the utilisation step, data from each classification decision are saved in a database. These data are intended for a review of the results and may be used in the next run as new or additional training data for the continuous improvement of the method's performance.

Then the software solution was included in the IT infrastructure of BIG KAUF AG in connection with the domain expert. In this connection, the use of classification results for controlling the audio-visual signalling was implemented. For this purpose, a combination of warning light and sound along with a monitor were installed at each entrance to the branch. If a person who triggered a negative classification result passes the previously defined area, the audio-visual signal is set off and the current camera picture is displayed on the monitor for a short time. As a result, the person should be automatically reminded to comply with the current hygiene requirements. Furthermore, a central monitoring of the application by a combination of live video feed with stored database information in the form of a dashboard was provided with aggregated information.

A follow-up order drawing on the implementation so far is conceivable in the future. The project objective could be defined as greater accuracy of roughly 95%, for example. For realisation of this

objective, two procedures could be relevant. Firstly, the algorithm could be retrained with continuously collected pictures that contain the classification results. These data must be checked manually for the correct execution of the classification. Furthermore, it is possible that additional layers will be included in the CNN structure. In addition, it might be possible to not only send out a signal, but also execute an action. For example, it might be possible to expand the actions by including a rotating door that is controlled and only opens if the classification result is satisfactory. This might ensure that access is only granted to persons wearing a medical mask. This is not achievable in the current implementation without additional effort by staff. Entry could be made more restrictive in a follow-up project. It is also possible that other potential requirements will be implemented in the context of the hygiene regulation. An adjustment of the underlying data may allow for a differentiation between the use of medical masks and FFP2 masks, for example.