Shaping Hybrid Data Science Work: Investigating the Role of Domain Experts

Amelie Schmid, TU Dortmund University & Robert Bosch GmbH, amelie.schmid@tu-dortmund.de
Manuel Wiesche, TU Dortmund University, manuel.wiesche@tu-dortmund.de

Introduction and Motivation

During the last years, data science considerably influenced the nature of work, leading to several opportunities and challenges with respect to existing, well-established functions and tasks (Michalczyk et al., 2021). In this context, it is already acknowledged that a close collaboration between data scientists and domain experts is a success factor for data science implementation as well as the shaping of data science work, respectively (van den Broek et al., 2021). It can be explained by the fact that – in contrast – a forced technology implementation might generate a feeling of fear which leads to a rejecting attitude of the domain experts (Pachidi et al., 2021).

Nevertheless, current studies regarding the collaboration of domain experts and data scientists consider the role of domain experts as rather passive, with only limited specific contribution towards the emerging data science work (Sambasivan & Veeraraghavan, 2022). Instead, domain experts are asked to simply provide their domain knowledge and share their know-how in a way that the data scientists are able to make use of it in the further development process (Lebovitz et al., 2021). However, the rational and evidence-based behavior of data scientists might result in the exclusion of the valuable yet unproven insights provided by domain experts (Power et al., 2019). In addition, the current insights regarding data science work and the collaboration of domain experts and data scientists are predominantly grounded on studies within non-technical areas like in HR (van den Broek et al., 2021), sales (Pachidi et al., 2021) and consulting (Strich et al., 2021). These groups have in common that based on their professional background, only limited understanding on data science exists. Consequently, the resulting data science work is predominantly shaped by the data scientist and not by the domain expert.

Considering the high relevance of domain knowledge for successful data science work (van Giffen & Ludwig, 2023), the present work aims to expand the existing perspective of data scientists by focusing on the unique role of domain experts. Thus, domain experts which are closely related to data, e.g. engineers will be investigated. The formation of data science work based on the professional background of engineers will be studied, using the following research question: How do engineers shape data science work?
Methodology

The analysis is based on a multiple case study within one of the biggest suppliers within the automotive industry. We selected the automotive industry because it is a strategically relevant showcase for the complex digitalization process of traditional companies including its domain experts. The interpretative nature of the case study includes frequent visits to the field over the course of 12 months (Walsham, 1995). It helped us to understand how domain experts, specifically engineers at the automotive supplier, shape data science work.

Table 1. Overview of cases in the multiple case design

<table>
<thead>
<tr>
<th>Domain</th>
<th>Case 1: Supplier</th>
<th>Case 2: R&amp;D</th>
<th>Case 3: Manufacturing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional work design</td>
<td>Manual claim management of each individual material; shipment of the broken parts to the supplier; discussion with supplier ex-post, based on the physical product.</td>
<td>Product development by using domain knowledge; decision-making based on experience of the employees as well as analytical models, basic statistics/simulation.</td>
<td>Manufacturing problems are solved by reacting to emerging issues; evaluation of the situation by an operator at the line; non-systematic quality management.</td>
</tr>
<tr>
<td>Work design with data science application</td>
<td>Claim management based on manufacturing data from different production lines; no shipment of broken parts; exchange of product data.</td>
<td>Data-driven product development; decision-making based on data models by using manufacturing data as well as field data.</td>
<td>Systemized problem solving using big data and visualization; problem analysis based on manufacturing data; data-driven quality assurance.</td>
</tr>
</tbody>
</table>

We chose a multiple case study design, as its analytical conclusions are generally more powerful and robust compared to single-case designs. This is based on the fact that the generalizability of the findings is expanded by using varied circumstances for the case selections (Miles et al., 2020). Thus, case similarity and case variability served as guiding principles for the selection of the cases (Orlikowski, 1993). We sampled workers from different hierarchical levels, a varied degree of data science knowledge and working experience as well as a mixture of backgrounds and work roles. Data was collected using 30 semi-structured interviews (Miles et al., 2020).

Data analysis: We used open, axial and selective coding for analyzing the interview transcripts (Corbin & Strauss, 2015). In a first step, open coding was used to summarize the key message in a short phrase, in order to identify the general work characteristics in the context of data science application (Miles et al., 2020). During the data analysis process, we realized the persistence of the former engineering work practices to a large extent. That is why we used axial coding for the identification of connections between the subcodes (Corbin & Strauss, 2015). We were able to map the former engineering work practices as well as the new data science work practices to each work element. In selective coding, we further developed and combined the dimensions, resulting in the final consolidation.

Preliminary results

The results show that engineers shape data science work by integrating (i) the established work practices of the domain and (ii) the new work practices of data scientists.

Combining a reactive and a predictive approach. On the one hand side, the reactive work practices are well established since engineers need to support in problem solving during the manufacturing. Short reaction times are needed, as internal errors cause a high amount of internal defects cost. This explains that the collaboration between R&D and manufacturing requires to stay reactive at specific circumstances: “We get a lot of verbal
feedbacks from the plants regarding their problems. They are saying: ‘[...] last night we had 10 pieces of failures consecutively, please support us with this issue.’” (#28)

Nevertheless, the identification of such incidences serves as an important baseline to learn for the future. Thus, the reactive working approach is strengthened by the predictive perspective through the application of data science. As an example, the proactive integration of the developed solution into future products by the workers form R&D ensure a reduction or even elimination of specific failures. Such a mindset of forward-thinking in R&D helps to introduce fundamentally new ideas: “With much more detail we can clearly identify the root cause in comparison to a hypothesis [...] Now we can understand: ‘If I do this, we will have a good result at the end of the line, because I understand what is happening in between.’ [...] And this allows to make better designs from scratch.” (#27)

Analytical diagnosis based on knowledge and data. One respondent highlights the benefits of knowledge application, like an intense and autonomous analysis of any problem. For that purpose, a deeper involvement into the field is necessary to deal with the situation and to deduce a solution simply based on domain knowledge: “The bottom line of my work is to build up knowledge. And I believe that this will remain the same and that you will continue to need the expertise, the experience.” (#22)

However, in some situations the exact identification of failures and error patterns is complex. This can even lead to an entire lack of understanding on occurring issues. To tackle that data science is strongly needed. As an example, a detailed visualization of the manufacturing data can help to get a general feeling for the respective circumstances: “Even the development who is very far away from production can check and see that: ‘, we have a problem at this and this station and that press fit force has been increased for the past two hours.’ [...] And then I see: ‘ok, the force increase is caused by this supplier batch of the pins.’ And this is making it much easier to detect the root cause of the problem.” (#26)

Using event-driven and systematic problem categorization. Case 3 shows that single, specific events must be analyzed in order define a solution for the specific problem. In particular, this holds if the occurring issue needs to be solved fast. Timing is another crucial factor, which explains an apparently randomized way of working: “The difficulty is: I always have the problems in the plants at different point in times. That’s tricky, because today plant A calls, tomorrow plant B calls.” (#09)

Such an event-driven procedure has to be guided by a systematic, all-time solution of the problem. As an example, the systematic usage of data provides a new basis for collaboration with external partners (case 1): “It is possible to make a claim of the performance of the components based on production data. This means that not only individual cases can be claimed, but also the global performance of different plants as well as differences in the performance of different suppliers.” (#29) The systematic problem categorization can also help to implement continuous improvement methods within the international production network. It enhances the detection of similar occurrences at the different locations: “We share all information in an international lesson learned network so that everyone in the organization can make use of a holistic approach and transfer it to his cases.” (#03)
Decision-making grounding on physical model and data model. Engineers build a basis for their decisions through the application of models which are referring to physical laws. Thus, decisions based on physical models do not necessarily require large amount of data for verification, like the selection of the appropriate material in the product design: “If I change the material from copper to iron, I would like to know the influence on my failure mode. My physical model tells me that stress and strain can cause cracks in the solder joint [...] and that iron promotes this more than copper. I have not recorded a single data point; I have derived analytically from physical models.” (#02)

By using the data for decision-making, it is possible to identify issues which cannot be explained solely by a physical model. One example is the correlation between the results and a certain supplier, which enables insights into their respective performance and creates a better basis for negotiation: “Based on our data we see that the range of the current defined specification is too big, and we can reduce it.” (#28). It is important to mention that the examples of the decision-making aspects illustrate how the physical and data model approach interact with each other. In the first glance, decisions by data models based on probabilities can be taken, which do not require any domain knowledge. However, the physical reason behind these failures is still lacking and unclear: “Of course, based on the data we would only see that the workpiece carrier is causing the issue and when we check it physically, we know this is being caused by the loose screw.” (#26)

Figure 2. Hybrid practice of data science work

Potential contributions
For data scientists, it is already well known that they utilize data to construct new ways of working (Muller et al., 2019). In contrast, domain experts were so far considered as being passive in the sense of shaping data science work (van den Broek et al., 2021). In the present work, we broaden this limited view and illustrate a new work setting which is created by the engineers themselves. Consequently, data science work is shaped by domain experts which leads to the fact that engineers question their existing boarder criteria. Moreover, the workers even calibrate and verify their established work results by means of data science. Among others, this can be emphasized e.g. in the sense of decision-making. Here, the solely application of data science, results in uncertainty and missing underlying explanations (Lebovitz et al., 2022; Provost & Fawcett, 2013). Nonetheless, work practices of engineering like the understanding of physical causalities remain relevant. We are planning to further concentrate on this and study the use of data science and AI in shaping new work practices of domain experts.
References


