Electronic Markets (2023) 33:52
https://doi.org/10.1007/s12525-023-00677-w

RESEARCH PAPER

Sanitizing data for analysis: Designing systems for data understanding

Joshua Holstein[1] · Max Schemmer[1] · Johannes Jakubik[1] · Michael Vössing[1] · Gerhard Satzger[1]

Received: 7 July 2023 / Accepted: 19 September 2023

Abstract
As organizations accumulate vast amounts of data for analysis, a significant challenge remains in fully understanding these datasets to extract accurate information and generate real-world impact. Particularly, the high dimensionality of datasets and the lack of sufficient documentation, specifically the provision of metadata, often limit the potential to exploit the full value of data via analytical methods. To address these issues, this study proposes a hybrid approach to metadata generation, that leverages both the in-depth knowledge of domain experts and the scalability of automated processes. The approach centers on two key design principles—semanticization and contextualization—to facilitate the understanding of high-dimensional datasets. A real-world case study conducted at a leading pharmaceutical company validates the effectiveness of this approach, demonstrating improved collaboration and knowledge sharing among users. By addressing the challenges in metadata generation, this research contributes significantly toward empowering organizations to make more effective, data-driven decisions.

Keywords Data understanding · Data governance · Metadata generation

JEL classification M15 · L6

Introduction

Over the last decades, an ever-increasing amount of data has been stored for analysis to tackle critical business and societal challenges. However, much of this data's potential remains unexploited. IDC (2020) stated that approximately 68% of the data is unused, as organizations struggle with extracting actionable insights (Ofe et al., 2023). One of the obstacles lies in the lack of understanding when it comes to the data's characteristics and how they are relevant to specific use cases (Abdel-Karim et al., 2021). This is a prevalent issue across various sectors, from manufacturing (Lenz et al., 2018; Voell et al., 2018) to banking (Vermeer, 2019) and healthcare (Dhayne et al., 2019). The consequent repetitive screening of data for different use cases results in repetitive effort for human experts. Depending on the number of features, this can be very time consuming (Duan et al., 2019), posing a substantial obstacle for the utilization of the data. Addressing this challenge requires effective strategies, among which data governance appears as a key factor, bridging the gap between data collection and utilization.

We recognize the critical role of data governance, and more specifically the provision of metadata, in addressing this obstacle. Data governance aims to ensure effective management and use of data resources (Khatri & Brown, 2010) by utilizing different elements, such as data quality and metadata. Metadata, defined as "information that describes data" (Singh et al., 2003, p. 1), not only reveal key attributes, such as data source, format, content, and meaning, but

Responsible Editor: Mark de Reuver

This paper belongs to the special issue on "Data economy in a globalized world: Opportunities and challenges for public and private organizations" (https://link.springer.com/collections/ihedfhbigd).

✉ Joshua Holstein
Joshua.Holstein@kit.edu

Max Schemmer
Max.Schemmer@kit.edu

Johannes Jakubik
Johannes.Jakubik@kit.edu

Michael Vössing
Michael.Voessing@kit.edu

Gerhard Satzger
Gerhard.Satzger@kit.edu

[1] Institute of Information Systems and Marketing, Karlsruhe Institute of Technology Kaiserstraße, 89-93, 76133 Karlsruhe, Germany

Published online: 09 October 2023

also provide a structured framework for understanding the data. Metadata therefore ease the data selection process and reduce the time and effort required by human experts. In turn, this facilitates informed decision-making by aligning data with the relevant use cases and allowing the data resources' full potential to be realized. The implications of this approach resonate in fields such as business intelligence and data analytics, where precise information extraction is critical for real-world impact.

Building upon the premise of metadata's importance, current methods for metadata generation, while effective, also have their limitations. Expert-driven approaches are contingent on the availability and expertise of the data creators, for example, datasheets (Gebru et al., 2021) and data nutrition labels (Chmielinski et al., 2022). This means that valuable metadata generation often faces obstacles when data creators are unavailable, especially when dealing with pre-existing datasets. Despite being efficient, automated approaches for metadata generation also have their shortcomings. One of the most significant issues is that these methods often lack the necessary context and domain-specific knowledge (Holland et al., 2018). They can capture basic statistical information, detect patterns, and identify dependencies within the data (Abedjan et al., 2015), but they may struggle to add context-specific details that are relevant to specific domains or use cases (Holland et al., 2018). The inefficiency of current metadata generation processes is not just a technical issue but has substantial implications for businesses and organizations. Inadequate or missing metadata can lead to inefficient data processing and decision-making (Kandel et al., 2012; Shankaranarayanan & Even, 2004), hindering the ability to leverage the data's full potential for operational and strategic tasks. As a result, there is a need for more efficient, yet comprehensive, metadata generation methods. Despite the potential benefits of metadata, there is a noticeable research gap in exploring effective and efficient methodologies for generating comprehensive metadata, particularly in the context of integrating domain-specific knowledge.

Considering these challenges, it is essential to explore novel methodologies that could streamline the process of metadata generation while optimizing the use of domain knowledge. Therefore, we formulate the following research question:

> RQ: *What design knowledge should guide the development of information systems that systematically improve metadata generation?*

Answering this question calls for a methodological approach that allows us to both create and evaluate innovative solutions in a real-world context. We therefore utilized a design science research (DSR) approach in our study (Kuechler & Vaishnavi, 2008), the specifics of which are detailed in the "Methodology" section, to derive design knowledge for systems that create metadata, thus improving data understanding for both domain and method experts. Based on two design requirements (DRs) derived from conducted interviews as well as a literature review, we propose two design principles (DPs): *data enrichment* (*semanticization*) and *benchmark dataset matching* (*contextualization*). While the former DP (i.e., data enrichment) complements the data with supplemental metadata, the latter (i.e., benchmark dataset matching) compares the existing dataset with benchmark datasets used for similar use cases. Our artifact addresses these DPs individually with two components, i.e., *data semanticization* and *data contextualization*, that support users in determining the meaning and relevance of features for specific use cases through the creation of metadata.

We worked with one of the world's largest pharmaceutical organizations that has gathered vast amounts of machine data from their automated production lines. The organization is currently working on multiple initiatives to enable its employees—with no formal training in data analysis—to utilize the available data for various use cases. Over nine months, we iteratively developed an information system that improves data understanding. Multiple real-world evaluations of the artifact demonstrated the proposed system's desirability, efficacy, and efficiency. Based on the qualitative analysis of the evaluations, we derive important insights, for example, that the proposed system allows domain and method experts to work collaboratively with the data, and that the proposed system therefore facilitates communication and knowledge sharing. This can result in the acceleration of the understanding of available datasets, to ultimately draw meaningful conclusions from the data.

The rest of the paper is organized as follows: In the next section, we elaborate on data governance and its components as the requirement for an in-depth understanding of the data. Furthermore, we outline related methods for the creation of metadata, which ultimately increase data understanding. Afterward, we explain the application context of our research and outline our "Methodology". In the "Findings" section, we deduct DRs from the conducted interviews, derive DPs, and instantiate the proposed system. In the following section, we discuss the results of the performed evaluation. Finally, we conclude with implications for research and practice, limitations of the conducted research, and avenues for future research.

Foundations and related work

In this section, we first outline the necessity of data understanding as the cornerstone of any data-related activities. As illustrated in Fig. 1, we contend that data understanding is crucial not only during the usage phase, but also during the

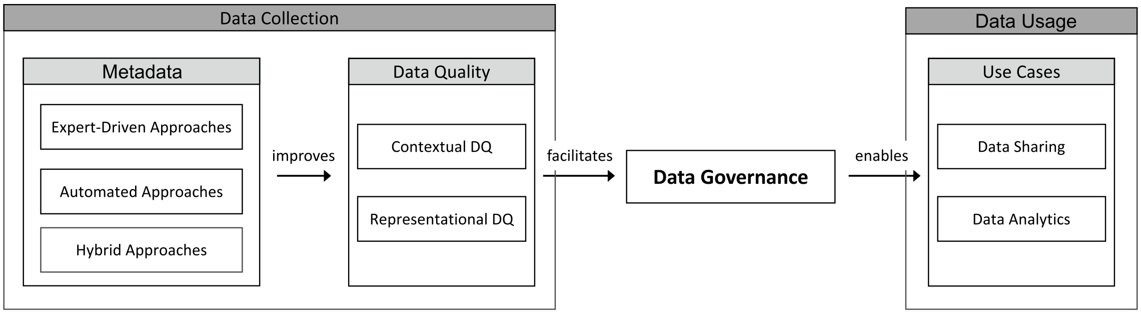


Fig. 1 Overview of theoretical pillars

preparation of comprehensive documentation, as required by standard data governance practices (Khatri & Brown, 2010). This documentation plays a vital role in enabling data consumers to utilize the information contained in the data effectively. However, achieving this goal necessitates both adequate data quality and high-quality metadata. Building upon these foundations, we outline commonly employed methods to generate such metadata in the related work. A thorough understanding of the capabilities and limitations of current methods for metadata generation builds the fundament for the challenges we address in our work and thus provides necessary context for our proposed solutions.

Data understanding and data governance

Data understanding involves, among other activities, becoming familiar with available (big) data, identifying data quality problems, and revealing insights (Wirth & Hipp, 2000). In the realm of big data, this understanding is particularly challenging given its five defining characteristics: volume, velocity, variety, veracity, and value (Ishwarappa & Anuradha, 2015). The inherent complexity of big data, further underscored by these dimensions, is highlighted by Jagadish et al. (2014). This complexity often stems from the numerous features contained within the data, each feature potentially offering valuable insights but also adding to the dataset's intricacy (Fan et al., 2014). The challenge lies in understanding and navigating this multitude of features effectively, as traditional database and software techniques often fall short in handling this magnitude of complexity (Fan et al., 2014). Considering these complexities, experts find themselves needing to dive deep into the data, rectifying data quality problems and selecting suitable methods for further data analysis. While existing research mostly focuses on the required understanding within the data's usage phase, for example, data exploration (Dimitriadou et al., 2016) and visual analytics (Cui, 2019), we argue that an adequate understanding of the data is also necessary to create comprehensible documentation, as required by data governance practices, about the dataset for later use cases like data sharing or data analytics projects.

Data governance refers to strategic decisions and accountabilities to leverage available data resources effectively(Khatri & Brown, 2010). The authors propose several decision domains, of which data quality and metadata are most relevant within our research scope, as they support data consumers in understanding the data's meaning and relevance:

1. *Data quality* can generally be defined as the data's "fitness for use" (Wang, 1996, p. 6). Based on the framework of Wang (1996), we focus on contextual and representational data quality. Contextual data quality aims at the usability of the data, considering the task at hand, and includes dimensions such as relevancy, timeliness, and completeness. In the context of creating value from data, Zeng and Glaister (2018) determined the capability to contextualize data as a primary driver to derive value from internal and external data sources. Furthermore, representational data quality describes how users can understand and interpret the data. In the age of big data, where vast quantities of data are gathered, it is crucial to consider the contextual and representational dimensions of data quality, particularly since not every feature may be relevant to every use case (Fan et al., 2014). This can be accomplished, for example, by utilizing metadata.
2. *Metadata* can generally be defined as "information that describes data" (Singh et al., 2003, p. 1). Accordingly, Khatri and Brown (2010) described the goal of the metadata decision domain as "establishing the semantics or 'content' of data so that it is interpretable by the user" (p. 149). Singh et al. (2003) further distinguished metadata into different categories, of which *domain-specific metadata* and *user metadata* are particularly relevant within this research. These two categories address the previously mentioned data quality dimensions. Domain-specific metadata refer to application communities' agreed-upon concepts of representations and therefore help interpret the meaning of features. Furthermore, user

metadata refer to annotations users made regarding the use cases for which the data was employed. Such metadata can be utilized for following data analytics projects to determine the required features for specific use cases.

Methods for metadata generation

Expanding on the significance of data understanding and governance emphasized earlier, it is crucial to examine the available techniques for creating metadata. Both traditional techniques, driven by expert knowledge, and modern, automated methods have their strengths and challenges. This section analyzes the intricacies of these methods and provides perspectives on their suitability and potential.

Expert-driven approaches include, for example, the creation of datasheets (Gebru et al., 2021) and data nutrition labels (Chmielinski et al., 2022). As proposed by Gebru et al. (2021), datasheets require the answering of 57 questions to gather relevant information on the dataset, for example, *the purpose of the data*, *what the instances represent*, and *how the data was collected*. This documentation of data resources has been established as a standard practice at data-driven companies (Gebru et al., 2021), a responsibility that often falls to those individuals most familiar with the data. While these individuals use their profound knowledge of the dataset to provide comprehensive and insightful documentation, several challenges may emerge. A significant issue arises with large dimensional datasets. Since the size and complexity of datasets increase, it can become an overwhelming, time-consuming task to generate detailed metadata manually, making traditional documentation methods less feasible (Wu et al., 2023). Addressing the scalability of metadata generation becomes a key concern when dealing with large and complex datasets. This problem is particularly noticeable when dealing with pre-existing datasets, where the original creator may be unavailable. In such situations, conventional methods of documentation may prove insufficient, necessitating innovative strategies for metadata generation.

Automated approaches for metadata generation vary widely in their mechanisms and offer different advantages. Data profiling is one such approach, examining data to extract basic statistics like minimum, maximum, and average values, or to detect dependencies between columns such as foreign keys (Abedjan et al., 2015). Data cataloging is another technique that often encompasses functionalities to automatically gather, organize, and index data, capturing metadata, such as data type, size, and origin, to create a searchable catalog (Ehrlinger et al., 2021) with the goal to enable a wider range of employees to work with available datasets (Labadie et al., 2020). However, while these tools facilitate a preliminary understanding of the data, their outputs often require further interpretation by human experts (Abedjan et al., 2015) to generate precise and contextually relevant information (Ehrlinger et al., 2021; Pal et al., 2019). Moreover, machine learning (ML) methods—programs that can learn from experience (Mitchell, 1997), where experience is often measured by the programs' access to training data (Padmanabhan et al., 2022)—are increasingly employed in automatic metadata generation. Supervised learning techniques use labeled datasets to predict metadata for new instances (see, e.g., Pepper et al. (2021) and Safder et al. (2020)). In recent years, natural language processing techniques have also emerged as promising approaches for automated metadata generation. For example, speech-to-text technologies transform audio into text, assisting in tasks like metadata generation from videos (e.g., Pal et al. (2019)). The choice of method depends on various factors, including data type, complexity, required metadata, and the feasible level of human intervention.

Methodology

We conducted a DSR approach, which seeks to combine scientific rigor and practical problem-solving, to create and evaluate innovative artifacts and thus advance existing knowledge to provide solutions to real-world problems (Hevner et al., 2004). More specifically, we followed the DSR framework proposed by Kuechler and Vaishnavi (2008) along its five steps: awareness of problem, suggestion, development, evaluation, and conclusion. We first derived design knowledge in the form of DPs, instantiated them in an artifact, and then evaluated the artifact's efficacy, general desirability, and efficiency in improving data understanding. In Fig. 2, we give an overview of the methods used in two design cycles. Following this approach, we initiated the first step of the DSR framework—building awareness of the problem.

To do so, we conducted semi-structured expert interviews (Whiting, 2008) with practitioners involved in the analysis and identification of potential use cases for machine data. We purposefully selected (Etikan, 2016) our interview partners by including three distinct groups: managers, domain experts, and method experts. This resulted in eight semi-structured interviews (lasting, on average, 28 min), which we transcribed subsequently and coded them inductively (Mayring 2000). An overview of the interview partners is shown in Table 1.

In the suggestion phase, we derived two DRs for information systems, that improve data understanding, with the DRs based on the inductively developed codes from the expert interviews. Afterward, we formulated two DPs, following the framework of Gregor et al. (2020), with the two DPs based on the DRs identified during the development phase. Next, we implemented the DPs in an artifact.

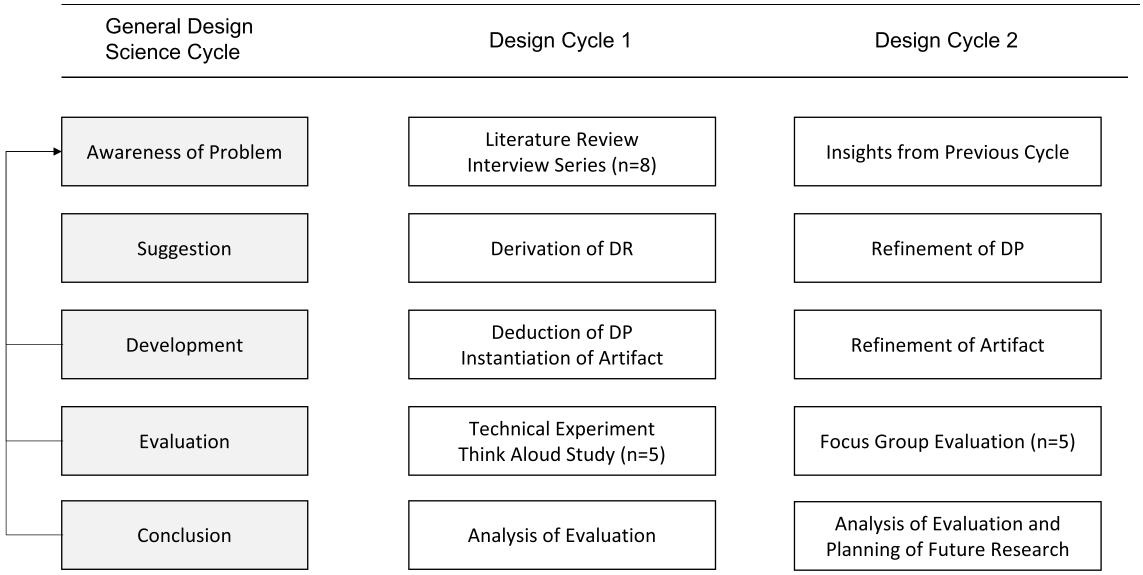

General Design Science Cycle
Design Cycle 1
Design Cycle 2
Awareness of Problem
Literature Review Interview Series (n=8)
Insights from Previous Cycle
Suggestion
Derivation of DR
Refinement of DP
Development
Deduction of DP Instantiation of Artifact
Refinement of Artifact
Evaluation
Technical Experiment Think Aloud Study (n=5)
Focus Group Evaluation (n=5)
Conclusion
Analysis of Evaluation
Analysis of Evaluation and Planning of Future Research

Fig. 2 Overview of design cycles and respective activities, based on Kuechler and Vaishnavi (2008)

For evaluation, we followed the "Framework for Evaluation in Design Science Research" (FEDS) of Venable et al. (2016). We applied the human risk and effectiveness strategy, which should be selected "if a critical goal of the evaluation is to rigorously establish that the utility/benefit will continue in real situations" (p. 82). Accordingly, we started with artificial and formative evaluations and then progressed toward naturalistic and summative evaluations. Furthermore, we utilized the framework of Venable et al. (2012) to formulate the goals of the evaluation periods. We progressed as follows:

First, we conducted a technical experiment to demonstrate the technical efficacy of our proposed model (see section entitled "Technical experiment"). We collected data from three manufacturing machines over one month, and domain experts categorized the data's features manually in manufacturing domain-specific categories (counter, status, failure code, setpoint, measurement). Using this data, we tested a range of ML algorithms (see section entitled "Technical experiment"), reserving 30% of the data for testing. To account for class imbalance, we used the macro-averaging F1-score, whereas the final model, a random forest, achieved a score of 97%. We validated the model's robustness successfully by testing performance consistency across individual machines and a second production site's data.

Second, we performed a field experiment using a think-aloud study (Ericsson & Simon, 1993) to evaluate the

Table 1 Job roles of the participants of the ideation and evaluation

Participant	Job role	Education	Years of industry experience	Awareness	Evaluation (Cycle 1)	Evaluation (Cycle 2)
Alpha	Senior Manager Digital Transformation Lead	PhD graduate	17	X		
Beta	Senior Manager Engineering	Graduate	7	X		
Gamma	Manager Project Portfolio	PhD graduate	7	X		
Delta	Senior Manager Operations	Under-graduate	12	X		
Epsilon	Global Manager: Manufacturing Digitalization & Analytics	Graduate	5	X		
Zeta	Data Scientist	PhD graduate	2	X	X	
Eta	Data Scientist	PhD graduate	1			X
Theta	Engineer	Graduate	2	X	X	X
Iota	Engineer	Graduate	1	X	X	X
Kappa	Engineer	Graduate	12			X
Lambda	Data Science Consultant	Graduate	3		X	
My	Data Analyst	Graduate	1		X	
Ny	Data Scientist	PhD graduate	1			X

artifact's general desirability, and identified areas for improvement (Venable et al., 2012). The think-aloud study included five participants, and the interviews lasted, on average, 43 min. Although we conducted interviews with managers in the awareness phase for ideation purposes, we intentionally prioritized the evaluation of the proposed system on actual users—domain experts and method experts. This decision was influenced by the interviewed managers confirming their reliance on aggregated reports instead of performing in-depth analyses of machine data themselves. Within this study, we refer to domain experts as individuals characterized by their extensive knowledge and experience within a specific field, despite not having received formal training in data analytics methods (Blair-Early & Zender, 2008). Conversely, method experts are those individuals who demonstrate proficiency in data analytics and ML techniques, although they may not have in-depth experience in or familiarity with a particular domain. After the interviews, we transcribed the recorded think-aloud studies and coded them inductively (Mayring, 2000). Based on the interviews, we learned the necessity of collaboration between domain and method experts to understand the data in more detail. We used this insight to refine our DPs and adjust the artifact accordingly.

Finally, we incorporated our learnings in a second design cycle, which we concluded with a field study in the form of a confirmatory focus group interview (Tremblay et al., 2010). We evaluated the proposed artifact's efficiency to enhance the understanding of available data, by automatically aggregating supplementary domain-specific metadata. We focused the evaluation on the first component of our artifact: data semanticization. For the focus group analysis, we followed the recommendations of Tremblay et al. (2010). To account for the naturalistic evaluation, we asked the participants to work on a real-world task of the case company. According to the recommendations of Tremblay et al. (2010), we created a manipulation within the focus group and let the participants perform a given task once with and then without the proposed system. To test the task description and questioning route, we ran a pilot focus group with two students. Afterward, we conducted two focus group interviews with in total five participants (three domain experts and two method experts), which we recorded, transcribed, and finally coded using template coding (King, 1998).

Evaluation context

Our study was conducted in collaboration with a pharmaceutical manufacturing company that had undertaken several initiatives to leverage the data generated by their packaging machines. The case company is currently transitioning to Industry 4.0, which involves digitizing their production processes and utilizing data generated by novel production lines featuring automated machinery. These machines can generate up to 2000 features per machine, which should be used to inform decision-making and optimize production performance. This includes, for example, making data from the manufacturing machines available for data analytics projects or sharing it with external partners. However, converting existing data to value-added knowledge can be complex for digitized processes (Mohamed, 2018). This especially applies to data generated from manufacturing machines due to their inherent complexity (Wuest et al., 2016). Data of manufacturing machines are usually gathered from programmable logic controllers whose primary purpose is to control the machines (Lenz et al., 2018). Therefore, all existing signals generated by the machines, such as alerts, status, measurements, and setpoints, converge here and are then stored for later analysis. Moreover, the challenge of deriving insights amplifies, as machine data consist of time series having high interrelation both with the signal itself and also with other signals. Another obstacle is the lack of (data) standards for the manufacturing machines and the reluctance of the suppliers of the machines to share their knowledge about their data structure; for example, "A factor that hinders this development [collection of contextual information for training of ML algorithms] is that every manufacturer has their own standard and tries to degrade interoperability" (Bokrantz et al., 2020, p. 7). Combining the complexity of manufacturing data and the lack of knowledge of the data structure makes analyzing this data especially challenging.

Findings

In this section, we present our findings. We start with identifying DRs that aim to improve data understanding, with the DRs based on the interviews we conducted. Then, we describe the DPs derived from these requirements, offering design knowledge to improve data understanding. Lastly, we create a prototype to demonstrate these principles in practice, which we evaluate in the following section.

Design requirements

Next, we develop DRs, drawing from our detailed analysis of the existing literature and the coding of the expert interviews we had conducted. This exploration helps us understand the challenges that inhibit data understanding and metadata generation, i.e., the lack of information describing features' meanings and the opaque presence or absence of relevant features for future use cases.

Based on the analysis of the conducted expert interviews from the first design cycle (see Fig. 2), we learn that the lack of data understanding has various causes. First, there is an absence of appropriate information describing the meaning of the features. For this reason, method and domain experts currently work together in an iterative, effortful process to better understand the features. Second, the presence or absence of relevant features for upcoming use cases is opaque and requires a manual search. Accelerating this process (see the section "Focus group" for more details on its current implementation) is crucial for using such available datasets efficiently. Considering that there are different underlying factors contributing to the absence of data description and the determination of induvial feature relevance, and since these root causes therefore require individual solutions, we divide the need for a better data understanding into the following and specify their needs in more detail.

In our case, machine data that need to be analyzed are usually high dimensional, containing up to 2000 features per machine. Wuest et al. (2016) further added that manufacturing data can "contain a high degree of irrelevant or redundant information" (p. 29). As interviewee Zeta mentioned, "[...] there are cases that when I look at a machine [from producer A] at line X, it does not necessarily have the same features as the machine [from producer A] at line Y, although you would expect that. That's also due, for example, to the age of the machine and the applied updates [...]." This means that the analysis performed for a specific type of machine for one line cannot necessarily be transferred to the same type of machine on a different line. This is, for example, due to different ages, software versions, and vendors of the machines (Jaspert et al., 2021; Lenz et al., 2018). Furthermore, Bokrantz et al. (2020) highlighted that manufacturers have their own standard to degrade the interoperability. This addresses the different naming conventions among machine manufacturers and their lack of openness to share their knowledge regarding the data's structure. As a consequence, the availability of features within the dataset becomes inconsistent, requiring repeated screening and comprehension for each specific use case. Accordingly, scalability is hindered, and the process becomes time consuming. For example, interviewee Epsilon stated that "[...] how we currently do it is time intensive, and we do it more with, not really engineering, but we start with the feature that we want, and push it in the machine. Then it shows something somewhere, and then we know, ok, this feature means this." Another dimension of scalability addresses the limited number of available experts with knowledge of how to analyze data or with an in-depth understanding of the domain. To utilize available data resources efficiently, it is necessary to scale the understanding of data, meaning that more employees must be empowered to understand the data. We therefore formulate our first DR,

> **DR1 – Scalability of Data Understanding:** *The system should make data understanding scalable across different user groups.*

Domain experts usually have broad expertise in their domain but lack knowledge of analyzing data (Lenz et al., 2018). Since their expertise is crucial for identifying and engineering relevant features, the system needs to be intuitive and not rely on previous knowledge in the field of data analysis to make the data easily understandable. Interviewee Theta stated that he does not want "[...] overloaded information and [...] I do not want to look for it forever, like is currently the case, and I do not want to have a situation where I must first understand each piece of information." Therefore, we formulate the following design requirement,

> **DR2 – Simplicity of Data Understanding:** *The system should be easy to use for different user groups having varying levels of knowledge of data analysis.*

Design principles

Based on the identified DRs, we aim to develop a system to enhance data understanding. As the interviews of our first design cycle and the literature show, significant challenges exist in understanding data resources containing no metadata, making it difficult to understand whether useful features for specific use cases exist. We therefore propose to enrich data, taking their domain into consideration.

The first step toward making the available data sources more accessible is augmenting the metadata in line with domain-specific insights. Take manufacturing, for instance, where numerical features could be further distinguished into categories, including counters, measurements, or statuses. Generating the missing link between a feature and its corresponding category helps understand the feature's meaning. In addition, we propose to apply text mining to leverage the feature's names' inherent information. Despite difficult-to-understand names, text mining can help identify potentially hidden patterns in the names, with the patterns helping in understanding the relationship between features. Currently, domain experts' knowledge is leveraged to manually determine such information for requested features. Since there are not many domain experts with a rich knowledge of the machine and its representation on the data level, the current process constitutes a bottleneck in the scalability of data-driven use cases. We propose to address DR1 (scalability) by implementing functionalities to automate the generation of domain-specific metadata. Furthermore, since ML has shown its potential to uncover hidden relationships for high-dimensional datasets (Janiesch et al., 2021), we want to leverage this strength. Accordingly, we want to use ML to classify the features into domain-specific categories, for example, based on their distribution. By utilizing ML, we ensure the scalability (DR1) of our system and formulate the following DP,

> **DP1 – Data Enrichment (Semanticization):** *Provide the system with capabilities to enrich available data resources with domain-specific metadata to support users in understanding the data's characteristics.*

In addition to determining a feature's category, it is also necessary to identify whether the relevant features for a specific use case are included in the dataset, for example, "the extent to which data are applicable and helpful for the task at hand" (Wang, 1996). On the one hand, domain experts have a rich knowledge of existing systems and their influencing factors, which they can transfer to method experts by helping determine the relevance of features. On the other hand, data from similar use cases from the past, in the form of benchmark datasets, often exist. With the help of such user metadata, the inherent knowledge of those curated benchmark datasets can be leveraged to transfer knowledge (e.g., the relevance of particular features) from previous use cases to the present one. Therefore, we propose to compare existing features with benchmark datasets,

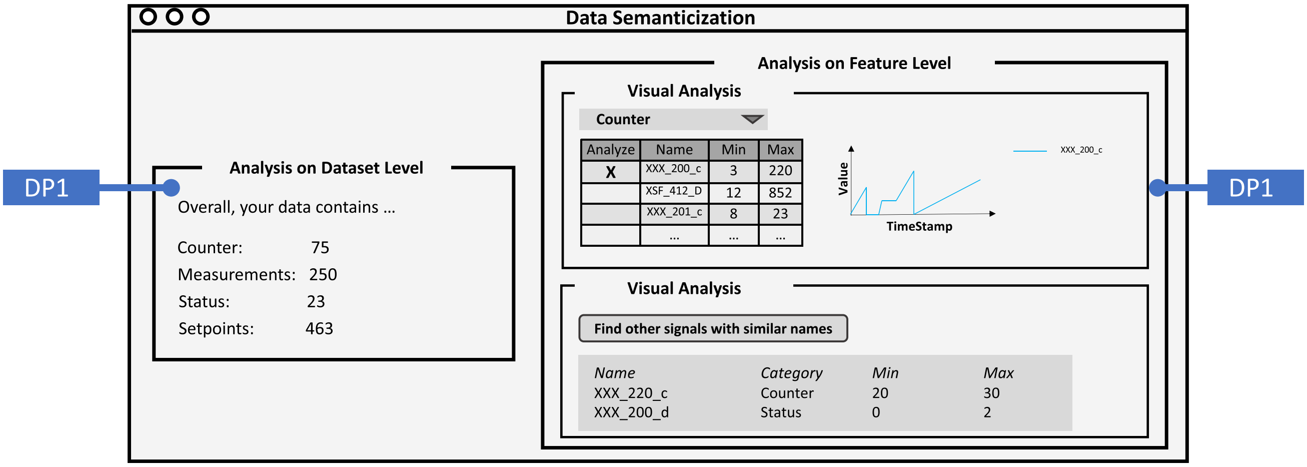
> **DP2 – Benchmark Dataset Matching (Contextualization):** *Provide the system with capabilities to compare different datasets in order for users to determine the relevance of their available features for a specific use case.*

Instantiation of design principles and solution description

In this section, we instantiate our DPs in a system and domain to enable effective evaluation. For insufficiently described data, we propose a system that integrates the knowledge of domain experts. Enriching and contextualizing the data semantically helps users increase their understanding. The system therefore supports obtaining an overview of the available data, understanding how relevant the present data is for a specific task, and understanding whether additional, important features are lacking. The system consists of two components: *data semanticization* (see Fig. 3) and *data contextualization* (see Fig. 5). The former enriches the data semantically through domain-specific metadata, thus instantiating DP1. The latter utilizes available user metadata to provide context to other datasets used for the same use case and thus instantiates DP2. To evaluate these DPs effectively, we instantiate them in a system and in a domain. In the following, we therefore first elaborate on the design of the system on a general level before we explain each component and how the DPs influence them in more detail.

In contrast to method experts, domain experts often have only limited knowledge of data analysis and are generally novices in this field. Schenk et al. (1998) explained that novices require higher mental efforts to retrieve information than experts. Our system should therefore help minimize necessary mental efforts by guiding the users. Blair-Early and Zender (2008) proposed presenting information linearly, for example, step by step, to increase the user's existing knowledge. Besides the structure, the system is designed to be both easy to use and understand for data novices (e.g., through visual data exploration).

Data semanticization. This component aims to complement the features with domain-specific metadata in two steps. First, only non-numeric features are categorized based on their datatype, for example, booleans or strings. In addition, we assume numeric features with only two unique values similar to booleans, for example, door open or closed, coded with 1 and 0 (see Fig. 4, left: 1.1). Second, numerical features are examined. Within a domain, there are usually different categories of features that have common behavior within the group but distinct patterns between groups, for example, measurements vs. counters in manufacturing. We leverage the characteristics of those groups to engineer new variables based on a feature's distribution (see Fig. 4, right: 1.2). Those engineered variables are used to train an


Data Semanticization
DP1
DP1
Analysis on Dataset Level
Analysis on Feature Level
Visual Analysis
Counter
Value
TimeStamp
XXX_200_c
Visual Analysis
Find other signals with similar names

Overall, your data contains ...

Counter:	75
Measurements:	250
Status:	23
Setpoints:	463

Analyze	Name	Min	Max
X	XXX_200_c	3	220
	XSF_412_D	12	852
	XXX_201_c	8	23
	...	...	...

Name	*Category*	*Min*	*Max*
XXX_220_c	Counter	20	30
XXX_200_d	Status	0	2

Fig. 3 Schematic overview of data semanticization

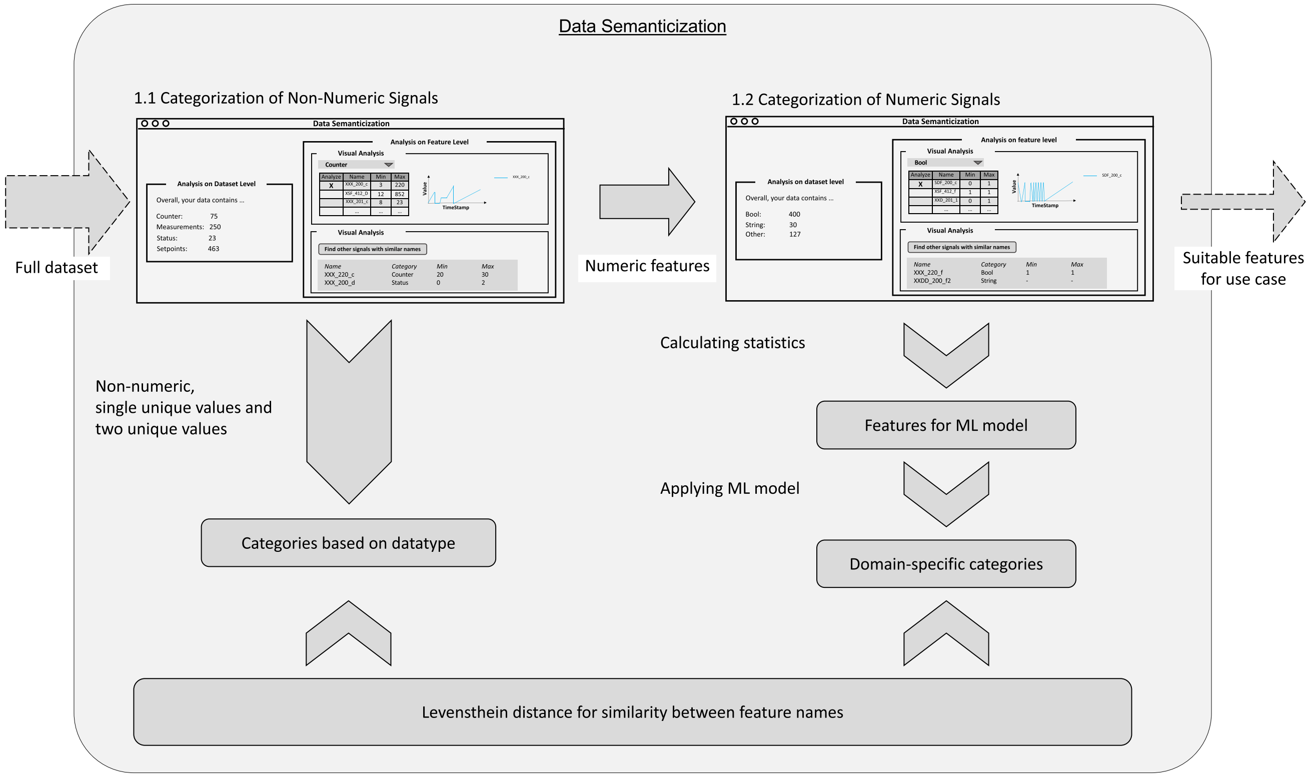

Data Semanticization
1.1 Categorization of Non-Numeric Signals
Data Semanticization
Analysis on Dataset Level
Overall, your data contains ...
Counter: 75
Measurements: 250
Status: 23
Setpoints: 463
Analysis on Feature Level
Visual Analysis
Counter
Analyze | Name | Min | Max
X | XXX_200_c | 3 | 220
XSF_412_D | 12 | 852
XXX_201_c | 8 | 23
... | ... | ...
Value
TimeStamp
XXX_200_c
Visual Analysis
Find other signals with similar names
Name | Category | Min | Max
XXX_220_c | Counter | 20 | 30
XXX_200_d | Status | 0 | 2
Full dataset
Numeric features
1.2 Categorization of Numeric Signals
Data Semanticization
Analysis on dataset level
Overall, your data contains ...
Bool: 400
String: 30
Other: 127
Analysis on feature level
Visual Analysis
Bool
Analyze | Name | Min | Max
X | SDF_200_c | 0 | 1
XSF_412_f | 1 | 1
XXD_201_1 | 0 | 1
... | ... | ...
Value
TimeStamp
SDF_200_c
Visual Analysis
Find other signals with similar names
Name | Category | Min | Max
XXX_220_f | Bool | 1 | 1
XXDD_200_f2 | String | - | -
Suitable features for use case
Non-numeric, single unique values and two unique values
Calculating statistics
Features for ML model
Applying ML model
Categories based on datatype
Domain-specific categories
Levensthein distance for similarity between feature names

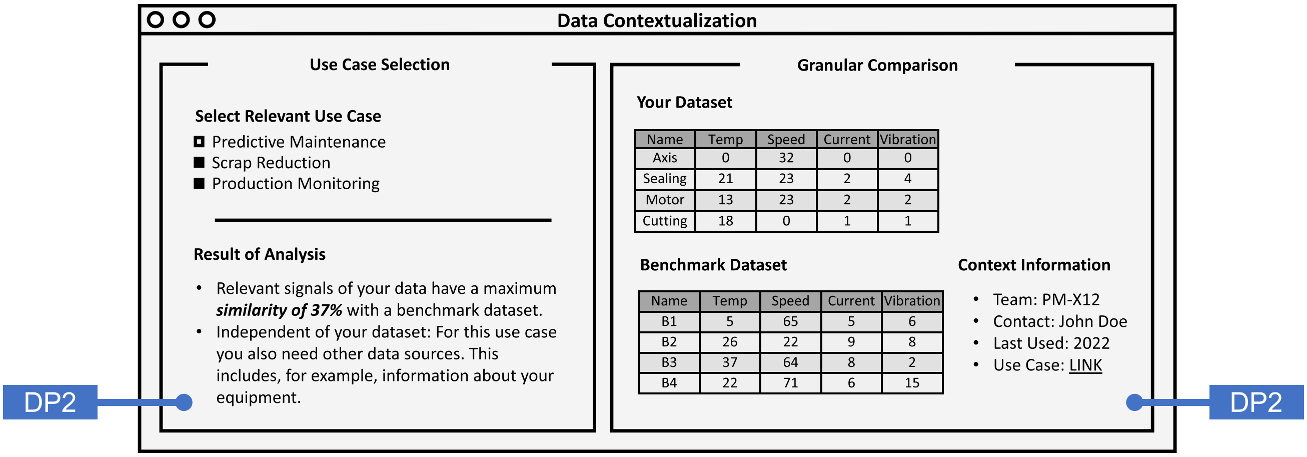
Fig. 4 Schematic overview of the dataflow throughout the data semanticization

ML model to classify numerical data into domain-specific categories, for example, whether a feature represents a measurement or a counter (see Fig. 3: Analysis on dataset level). These metadata empower users to explore, understand, and interpret the data more easily. We provide users with the possibility to plot the data (see Fig. 3: Visual analysis), as visual data exploration enables faster data exploration and does not require previous knowledge of mathematics or statistics (Keim, 2002).

In addition, we use the Levenshtein distance (Levenshtein, 1965) to compute the similarity among feature names, allowing us to uncover hidden patterns and relationships (see Fig. 3: Textmining). This identifies features with similar names (and thus presumably also similar behavior) that could potentially be of interest, for example, to scale the use case further. For instance, the features V720_R_SenPlW_State and V750_R_SenDis_State are more similar to each other than they are to V400_MD_VC_Mspe (see Fig. 3). To summarize, the instantiation of DP1 complements the available data with metadata (semantics) by grouping them into domain-specific categories, which allows the discovery of previously unknown relationships. Finally, this allows users to explore the data efficiently and easily.

Example: Figure 3 shows an analysis of a dataset comprising 811 non-numeric features. By utilizing information about the features' distribution, the ML model was able to categorize them into domain-specific classes, such as counters, measurements, status, and setpoints. To analyze the available counters, the system user selected the category counter and chose to visually inspect the feature XXX_200_c from the list of all available features classified as counter. In addition, the user can select other features for visualization and comparison. Finally, the user requested to find signals with similar names and identified XXX_220_c and XXX_200_d as features with similar names.

Data contextualization. This component intends to support users in determining the relevance of their dataset for a specific use case. The users therefore select a use case from a domain's list of potential use cases, for example, predictive maintenance or scrap reduction in manufacturing. To evaluate the selected use case's feasibility with the dataset at hand, we leverage available user metadata by comparing the dataset to benchmark datasets that have already implemented a specific use case successfully. For this purpose, public datasets can be used, as well as datasets from the organization, once similar use cases are successfully applied. For example, for a predictive maintenance use case, various measurements are required. Therefore, we make use of benchmark datasets that employed a predictive maintenance use case (see, e.g., Arnab (2020), Axenie and Bortoli (2020), or Matzka (2020)).



Data Contextualization

Use Case Selection

Select Relevant Use Case

▣ Predictive Maintenance
■ Scrap Reduction
■ Production Monitoring

Result of Analysis

- Relevant signals of your data have a maximum ***similarity of 37%*** with a benchmark dataset.
- Independent of your dataset: For this use case you also need other data sources. This includes, for example, information about your equipment.

Granular Comparison

Your Dataset

Name	Temp	Speed	Current	Vibration
Axis	0	32	0	0
Sealing	21	23	2	4
Motor	13	23	2	2
Cutting	18	0	1	1

Benchmark Dataset

Name	Temp	Speed	Current	Vibration
B1	5	65	5	6
B2	26	22	9	8
B3	37	64	8	2
B4	22	71	6	15

Context Information

- Team: PM-X12
- Contact: John Doe
- Last Used: 2022
- Use Case: LINK

Fig. 5 Schematic overview of data contextualization

To evaluate the fitness of the data for the intended use case, the available features of the own dataset and of all available benchmark datasets are identified by mapping substrings to use case-specific categories, for example, *temp* is matched to the category *temperature*. A weighting, developed by domain experts, is then applied to the categories to calculate a simple matching score. This score is based on the number of times features occur within each category, which is compared to the own dataset and each benchmark dataset. For example, in Fig. 5, we see four benchmark datasets (B1, B2, B3, B4) and our own dataset with four subgroups (e.g., axis, sealing, motor, and cutting, in Fig. 5). After the names of the signals are mapped to the categories, we can observe that each machine part of our dataset contains less signals for every category compared to the benchmark datasets. To find the subgroup of the data that contains the most relevant signals with a benchmark dataset, we calculate the score per subgroup and benchmark dataset and utilize the highest score. Furthermore, to avoid scores larger than one, for example, if the dataset contains more signals for a category than a benchmark dataset does, we employ the minimum function. Thus, we limit the matching score to a maximum of one. To summarize, the matching score is calculated as follows, where C denotes the set of categories:

$$\text{matching\_score} = \sum_{c \in C} \text{weight}_c \, * \, \min\left(1, \frac{\text{occurrences}_{c,\text{subgroup}}}{\text{occurrences}_{c,\text{benchmark}}}\right)$$

Finally, the highest matching score is displayed together with, if applicable, further use case–specific requirements. While the score considers features' presence, it does not take their course over time or distribution into account. However, the absence of features relevant to the use case, and thus also a low matching score, allows domain and method experts to semi-automatically exclude unpromising use cases and instead focus on potentially suitable ones. The understanding of the data is improved as the system evaluates the relevance of available features for the task at hand. Furthermore, the comparison between existing features in the own dataset and the benchmark dataset helps users determine which additional features other experts used in the past, that are not available in the dataset at hand. This allows users to adjust and, if necessary, extend the collected features to build a suitable foundation of features to realize novel use cases. To summarize, the instantiation of DP2 complements the available data with user metadata (context) by comparing the data to benchmark datasets. Finally, this allows users to determine the relevancy of the dataset for a specific use case and, if necessary, to expand the collected data with additional required features.

Example: In the second step of the analysis, as shown in Fig. 4, the user determines the relevance of the features for specific use cases. Here, the user aimed to develop a predictive maintenance use case using his or her dataset. An automated analysis revealed that only 37% of the features in the user's dataset were also employed by others for the same use case. To gain a better understanding of the available features, the user received a granular comparison between his or her dataset and the features employed by others. This comparison revealed that others often used features such as current or vibration, which are not present in the user's dataset.

Evaluation

In this section, we elaborate on the conducted evaluations presented in Fig. 2 and analyze the findings. We start with an explanation of the technical evaluation and the qualitative think-aloud study from the first design cycle, which helped us refine the DPs and thus build our artifact. Afterward, we discuss the final qualitative confirmatory focus group evaluation.

Technical experiment

During the first design cycle, the focus was on a thorough evaluation of the technical effectiveness of the ML model. A structured technical experiment was performed based on established methodologies to evaluate the capabilities of the model, and to refine the artifact further based on empirical data. The main objective of the experiment was to analyze the model's ability to process and interpret data from the manufacturing machines. This approach allowed for a comprehensive analysis that involved feature engineering, performance metrics, and model validation.

Data collection. We collected all available data from three manufacturing machines over the time horizon of one month. The data included 3906 features in total, of which 3415 were non-numeric features, such as string and boolean, and 491 were numeric features, respectively. To clean the data, we used the information about the connection quality provided by the machines (True and False) and dropped all instances having a bad quality. Since this data lacks appropriate metadata, domain experts labeled the numeric features manually into five manufacturing domain-specific categories: counter, status, failure code, setpoint, and measurement. When questions about the correct class arose, the domain expert consulted another domain expert, and together they determined the class, using the physical machines via reverse engineering. A description of these categories is shown in Table 2.

Feature engineering. As previously mentioned, we leverage the characteristics of these categories, such as the ongoing incremental nature of a counter and the infrequent resetting, to create new features, and we achieve this by analyzing the distribution of signals. Hereby, we leverage domain and analytics knowledge to identify suitable features that describe the distinct categories: *frequency of occurrence*, *frequency of sign changes*, *mean*, *standard deviation*, *interquartile range*, *number of unique values*, *maximum and minimum*.

Performance metric. Since the distribution of the classes is imbalanced, we used the macro-averaging F1-score as our performance metric. The F1-score is the harmonic mean of precision and recall, providing a balance between these two metrics. The macro-averaging approach further enhances its effectiveness in our context by calculating these metrics for each class independently and averaging them afterward. This method gives equal weight to each class, irrespective of its proportion within all classes, thereby ensuring that our performance evaluation is not biased toward the majority class.

Algorithm selection and hyperparameter optimization. We use the prepared data for the initial training and testing of the ML algorithms, employing the Scikit-learn library in Python for this purpose. We partitioned the data randomly, utilizing 30% as a test set. To fine-tune the models, we applied grid search for each algorithm to optimize the hyperparameters via a fivefold cross-validation. A comprehensive overview of the utilized algorithms and their corresponding results is presented in Table 3. Among all algorithms tested, the random forest outperformed the rest, achieving an F1-score of 97%.

Model validation. Finally, we test the robustness of the performance in two ways. First, we look at the performance of our model for each machine separately to ensure that the model's performance does not vary substantially between machines performing different processes. Second, we collect data from a second production site and apply our model to their data. The machines at this site provide less data—only 33 features for five machines of a manufacturing line. For this reason, experts have already identified their meaning in the past. Our model can classify all 33 features correctly such that we assume sufficient robustness of the performance.

Think-aloud study

In the first design cycle, we evaluated the general desirability of the artifact within a think-aloud study and identified areas for improvement (Venable et al., 2012). Here, we asked the participants to analyze a real-world dataset from their site. We provided them with two tasks: (1) identify whether the data contains product counters for a scrap monitoring use case and (2) determine the relevance of the dataset for a predictive maintenance use case. The participants provided rich feedback that demonstrated overall satisfaction with the system. Especially domain experts stressed that the proposed artifact opens new opportunities for a wider range of employees to

Table 2 Domain-specific categories of manufacturing and their frequency in the dataset

Category	Description	Example	# of occurrences
Counter	Counts the occurrences of an event	Number of good products	225
Measurement	Measures a process parameter	Temperature	157
Failure code	Represents a code for an error	High temperature	7
Status	Represents the status of a machine (part)	Running or down	46
Setpoint	Represents a setpoint	Target speed	225

Table 3 F1-score of selected algorithms

Algorithm	F1-score (macro)	Accuracy	Recall (macro)
Perceptron	0.21	0.49	0.23
Naïve Bayes	0.39	0.24	0.516
Linear Support Vector Classification	0.50	0.73	0.5
Ridge Classifier	0.62	0.74	0.61
Logistic Regression	0.73	0.84	0.73
K-Nearest Neighbors	0.87	0.86	0.87
Random Forest	0.97	0.95	0.96

work with the data as a domain expert. Kappa mentioned: “In any case, it makes the work with the tags [data] much more accessible, and easier.” For example, the knowledge of available and missing data allows them to prioritize the use cases more easily. Similarly, method experts commented that the system makes their work easier because the meaning of single features is (partially) revealed without having to rely on the knowledge of domain experts each time. To summarize, domain and method experts benefit from the system by being able to analyze their data (more effectively) on their own.

Participants found the comparison with benchmark datasets helpful in determining the dataset’s relevance for a specific use case, i.e., predictive maintenance. However, the displayed matching score led to different interpretations among the participants. For example, some participants drew references to an internal project which determined the dataset’s infeasibility for the specific use case and thus correctly concluded that the dataset was not relevant to the use case. Other participants, however, assumed that enough promising features were present and that the use case was consequently possible with the available features. Interviewee Lambda, for example, mentioned: “But since it suggests something to me, [...] I assume you could utilize [the data for the use case].” Assessing the relevance of features for solving a particular use case is a highly complex task. We also observe this in our evaluation. Users seem to need additional information beyond the displayed matching score to judge the use case’s feasibility. One of the natural options to facilitate this would be a staged system that categorizes the matching score to give a clear recommendation, such as a traffic light system. The benefit of such an approach would be its simplicity and practicability. However, the question that arises concerns how to estimate the thresholds for distinguishing certain categories. Answering this question certainly requires domain knowledge and empirical validation afterward. While assessing the relevance of features for solving a particular use case is an inherently complex task, providing information about the overlap between the dataset at hand and data required for a specific use case is the first step to facilitate the method experts’ and data experts’ process of use case selection.

Focus group

In the second design cycle, we utilized a confirmatory focus group to account for the naturalistic and summative evaluation (Venable et al., 2016) and we also tested the system’s efficiency. In the beginning, we let the participants identify relevant features for a scrap monitoring use case, first without the proposed system and afterward with the help of the system. In the following, we provide an overview of the current process without using the system.

Without the proposed system, domain and method experts work separately to identify relevant features. The method expert explains the use case and the data requirements, in terms of needed features, to the domain expert. Since domain experts often do not have knowledge of advanced data analysis, they rely on simple methods to analyze available data, for example, a list of all potentially available features. However, only selected features from this list are stored in the database to minimize costs. Based on this list, the domain expert then selects those features that sound potentially helpful while mentioning that this is always just a “best guess“ (Kappa) and there is a high degree of uncertainty about the relevant features. If possible, the selected features and a short explanation are summarized and sent to the method expert. The method expert then obtains an overview of the features by analyzing them. Afterward, she or he tries to assess the features’ relevance. It frequently happens that the domain expert selected not enough, wrong, or nonexistent (not stored) features. The method expert then reports the analysis results to the domain expert, who restarts the search for relevant features. Since the sub-processes take a considerable time and only require the specific knowledge of either the domain or the method expert, the process is designed iteratively.

Next, we performed the same task, using the proposed system. The system fosters collaboration between method and domain experts by enabling them to combine their skills. The advanced analytics provided by the system enabled a more informed decision-making process, reducing the reliance on “best guess” approaches. This is because it is “easier for you to verify whether it works or not” (Theta) and because it “provides more opportunities to improve the confirmation [of features] after all” (Iota). Furthermore, participant Eta highlighted that the system “give[s] the advantage of working on the same basis [and having] everything at a glance, sort of like a standard.” This standardization merges the different decision-making foundations (i.e., the list of feature names and the data itself) into a single source of truth.

During the evaluation, participants recognized that some features, previously identified by the domain expert through a list, were not present in the data. Conversely, with the proposed system’s assistance, participants identified features that described the same target variable in a more granular way. Domain expert Kappa noted the potential for discrepancies with the traditional method, stating, "I work with excel files [containing the features' names], which can be wrong. I believe more in what comes out of the machine."

Ultimately, the system allows participants to share their specific knowledge about the data. The domain expert, for instance, explained to the method expert certain patterns in the features' names, which is something she previously did not know. At the same time, the method expert could share observations from the data, that the domain expert was previously unaware of. Using their combined knowledge and skills, participants were subsequently able to identify patterns that they had been unable to identify individually. In contrast to the existing process, the use of the proposed system reduces the iterative nature of the process significantly. It minimizes the necessity for the domain experts to repeatedly reassess their feature selection, thereby enhancing efficiency and saving valuable time. This manifests the system's potential not only in facilitating a more comprehensive understanding of the data, but also in streamlining the process.

Importantly, participants observed the system's capacity to help identify and eliminate irrelevant data, potentially leading to reduced data storage. For instance, interviewee Iota described that the system might help, as they "have no idea yet what large amounts of data we have […] and we have a suspicion that many [features] are also not so helpful. We therefore suspect that we have a lot of data that would not need to be stored." This capability optimizes the feature selection process by distinguishing necessary data from the irrelevant, streamlining the approach to database management.

In conclusion, the proposed system considerably improves the metadata generation process for data analysis and therefore helps experts identify relevant features. By facilitating collaboration and knowledge sharing between experts, the instantiated systems reduce the iterations of the current process and allow for more efficient decision-making. The findings align with our initial goal of designing a more efficient and accurate system for metadata generation for high-dimensional datasets. These findings reaffirm the need for advanced analytics tools in data-intensive fields and provide valuable insights for future developments.

Discussion

As stated, organizations generate vast amounts of data every day with the intention to extract relevant information to address business and societal challenges. However, to derive accurate information, it is often not sufficient to apply complex ML models to high-dimensional datasets (Alt, 2021). Instead, organizations need to understand the collected data in depth to be able to unlock relevant insights.

In this context, we chose to focus on the pharmaceutical manufacturing industry for our study. This industry is highly regulated and subject to rigorous quality standards, leading to the production of a vast amount of data from diverse sources. Hence, it offers an ideal setting to test and demonstrate the efficacy of the ML model in a data-rich environment. Furthermore, the complexity of pharmaceutical manufacturing processes and the critical nature of the industry's operations highlight the importance of understanding and classifying machine data accurately. Moreover, larger enterprises such as the case company often lead the way in the adoption and development of new technologies (Lee & Xia, 2006), for example, due to their extensive resources (Thong & Yap, 1995). In the context of Industry 4.0, the technological advancements pioneered by larger enterprises become even more critical (Sommer & Sommer, 2015), as their diffusion to smaller and medium-sized enterprises (Nooteboom, 1994) is key to driving industry-wide innovation and efficiency. This potential broadens the impact of our research and underscores the value of testing our approach in a large-scale, complex environment like a pharmaceutical manufacturing company.

However, the application of these ML models in such a complex and data-intensive environment also presents certain challenges. While these organizations have access to a wealth of data, understanding and contextualizing this data effectively is a hurdle that needs to be overcome. This difficulty is not limited to the pharmaceutical industry—it extends to any organization attempting to extract insights from high-dimensional datasets. In the related work, we have highlighted that traditional methods often fall short when it comes to contextualizing high-dimensional datasets accurately. This deficiency poses a significant obstacle in data governance, where the utilization of metadata and high data quality is crucial for ensuring the usability of the collected data (Khatri & Brown, 2010). To close this gap, our research provides design knowledge for systems that aim to leverage the value of these insufficiently understood yet promising datasets for specific use cases.

Making use of our outlined design principles, we have demonstrated the effectiveness and desirability of our developed artifact, thereby paving the way for a deeper understanding of datasets through the generation of metadata. Our system handles the intricacies of high-dimensional datasets by generating automated metadata for domain and method experts to review collaboratively, addressing the shortcomings inherent in current methodologies. More specifically, we semanticize numeric features of the dataset by leveraging characteristics of the distinct classes and afterward compare the occurrences of features in other datasets to set the dataset in context with potential use cases. This process contextualizes the dataset within the landscape of potential use cases. Data, when enriched both semantically and contextually through metadata, may encourage data consumers to explore and harness their data assets more comprehensively. This, in turn, may lower the barriers for utilizing datasets and catalyze the process of value creation from data, for example, by utilizing it internally or by sharing it within their ecosystem.

Sharing datasets in the ecosystem

Building upon these findings, previous research has emphasized the lack of sufficient metadata and methodologies to identify available data as significant technological challenges (blinded for review), which often leads organizations to refrain from sharing their data with their ecosystem (Choi & Kröschel, 2015; Van Den Broek & Van Veenstra, 2015; Van Panhuis et al., 2014). To close this gap, it is necessary to develop novel methods to enhance metadata and thus improve data understanding. Our research can help organizations address these challenges by refining existing data governance processes. By improving experts' understanding of the meaning and (ir)relevance of high-dimensional data (DP1 & DP2), organizations can, for example, reduce barriers of data sharing, for instance, the fear of accidentally sharing competitive knowledge (Fassnacht et al., 2023).

Despite the potential benefits of data sharing, such as fostering inter-organizational innovation (Enders et al., 2022), it is essential to counterbalance these aspects with the potential risks and ethical dilemmas. The nature of data sharing is not inherently advantageous or disadvantageous; its implications largely hinge upon the conditions of its deployment and handling. A case in point is the widely publicized incident involving Facebook and Cambridge Analytica (Isaak & Hanna, 2018), where the unauthorized use and dissemination of personal data for politically oriented advertising sparked a maelstrom of ethical and privacy disputes. Our study fundamentally emphasizes the necessity for a comprehensive understanding of the data, its accompanying metadata, and possible applications prior to its sharing. Such insight could mitigate associated risks like ethical issues (Reer et al., 2023) and unintentionally sharing competitive knowledge (Fassnacht et al., 2023). Hence, responsible data sharing should align with the principles of data privacy, informed consent, and security, fortified by robust data governance processes.

Leveraging datasets in data analytics

It is important to note, however, that the challenges presented by insufficient data understanding go beyond the scope of data sharing and hold significant implications for broader data analytics processes, such as the application of ML. While ML models are a promising means to extract insights from data (Han et al., 2012), applying such models to insufficiently understood data can fail due to various reasons. For example, it might go unnoticed that models may rely on spurious correlations instead of meaningful relationships, or biased models might accidentally disadvantage minorities when prescribing medical treatments in healthcare or approving loan applications (Fabris et al., 2022). As Boyd (2021) showed, understanding the relevant characteristics of datasets, for example, the features' meaning, can help mitigate biases in models and thus prevents models from leveraging sensitive information and, finally, disadvantage minority groups.

One of the traditional approaches to identifying biased models is using inherently interpretable ML models or, in the case of more complex models, explainable artificial intelligence. However, despite the importance of explainability, the effectiveness of these methods in supporting users is uncertain when the data used to train ML models is insufficiently understood. This is particularly challenging when the meaning of individual features is unclear, highlighting the need for comprehensive data understanding for the application of explainable artificial intelligence methods. Our research contributes to avoiding such adverse consequences of employed datasets by collaboratively analyzing the features' meaning with domain and method knowledge (DP1). Nevertheless, obtaining this understanding also has implications for the training of ML models in general. Recent literature has highlighted the importance of this systematic preparation (and understanding) of data to train more accurate ML models. For example, one of the solutions is leveraging domain knowledge to create inherently meaningful features (Jakubik et al., 2022). The higher degree of information contained within the features can potentially enable experts to utilize simpler and inherently interpretable models, ultimately aiding in the avoidance of unintentional biases within the models and training more accurate ML models.

To integrate the required domain knowledge, for example, for feature engineering, into the data analytics process, a collaboration of domain and method experts is essential. In our think-aloud study, we found that the proposed methods for analyzing the data already help the domain and method experts, for instance, generate ideas about the meaning of individual features. However, they often require additional knowledge from the complementing expert to identify the features' true meaning. The instantiation of the proposed design knowledge thus creates a standard that allows the experts to discuss the data on a shared basis. Similarly, observations have been made in the field of data democratization, where a key enabler is collaboration and knowledge sharing to enable a wider range of employees to draw value from datasets (Lefebvre et al., 2021). Through our evaluation, we learned that while the domain expert usually prefers simpler-to-understand analyses, the method expert often needs further analyses to extract additional knowledge from the data, for example, correlations or feature importance.

Contribution to theory and practice

Regarding theory, we address the challenge of creating contextually and semantically meaningful metadata to increase the data understanding and, ultimately, to enable data consumers to leverage their datasets for different use cases. Furthermore, we discuss how the insufficient understanding of available data may potentially impact the application of data analytics projects, for example, inaccurate or biased models or noninterpretable models. To mitigate these undesirable consequences, we propose to leverage decision domains of data governance to generate the necessary data understanding. More specifically, we increase the data quality through the generation of metadata. Domain-specific metadata allow domain and method experts to jointly explore their data by implementing elements of data democratization. Furthermore, we show how existing user metadata, in the form of benchmark datasets (DP2), can be used to estimate the potential of datasets and their features' relevance for specific use cases. More generally, we describe the challenging characteristics of insufficiently documented datasets and the impacts of employing such datasets. Finally, we suggest a hybrid method that improves data governance and, as a result, prepares organizations to use their datasets internally or to collaborate with their ecosystem partners in today's data economy.

Our research also has implications for practitioners: The instantiation of our system tackles a real-world problem by generating an overview of the meaning and relevance of existing data and thus facilitates data understanding. This improved understanding not only simplifies the communication between domain and method experts, but also enhances their capabilities. First, instead of repetitively relying on the domain expert to determine suitable signals for a specific use case, method experts can identify potentially relevant signals that need further investigation. Second, domain experts are empowered to explore the data and its relevance for use cases more easily. In the process, the recurring task of screening data for available features is reduced, effectively removing a major obstacle to extract insights from available data. This not only streamlines data analytics projects, but also unlocks the latent value inherent in organizations' data sources more efficiently, thereby promoting a more effective utilization of data assets.

Limitations and future research

Our research certainly comes with limitations. Primarily, our study is restricted to machine data from the pharmaceutical manufacturing industry, introducing potential bias due to the specific constraints and conditions that are characteristic of this industry. Furthermore, the manual labeling of numeric features into domain-specific categories by domain experts introduces the risk of bias. This is because their personal understanding or interpretation might have influenced the categorization. Even though a second expert was consulted when there was uncertainty, this process still relies heavily on human judgment, which could be influenced by personal experience, education, or inherent biases. Next, the current study validated the model's performance, using data from a second production site. However, this introduces potential validation bias if the second site is not representative of the broader spectrum of manufacturing settings, i.e., the second production site does not reflect the variability in other sites or industries sufficiently. Consequently, it remains an open question as to how the ML model might behave with datasets from other sectors, marking an area for future research.

While the ML model was developed and tested in the context of pharmaceutical manufacturing, the principles and methods underlying the model's development could be applicable in other industries and domains. The methodology of using domain knowledge for manual feature labeling and the application of ML models can be adjusted and applied to different settings, given the domain-specific knowledge is available. However, each domain or industry has unique characteristics, constraints, and quality of data. The key is to adapt the approach to each specific domain. In light of this, especially the engineered features are vital, as they need to be engineered to the specific characteristics of the domain-specific categories. The method of labelling, the choice of engineered features, and the algorithm selection should all be adjusted according to the specific domain. The differences between domains could be reflected in the metadata by encoding domain-specific information, thus enabling the model to learn and adapt to different domains.

Another obstacle lies in the initial effort to categorize existing features manually. To minimize these required efforts for labeling the training dataset, further research should consider options like active learning and unsupervised learning, that may identify additional relations in the data. Similarly, the availability of public datasets that are suitable for the benchmark dataset matching is still limited, restricting the applicability of the contextualization component. However, we argue that this challenge will diminish with the rise of open data, as more data will be made publicly available from public and private organizations alike (Enders et al., (2022), Janiesch et al., 2022). The last obstacle is identifying which signals describe a particular real-world object within *benchmark dataset matching* (e.g., that a measurement represents a temperature). Thus far, this is solely based on the signal names, which limits its general applicability. However, we argue that if the features' names do not contain any information, *benchmark dataset* *matching* is difficult not only for ML models, but also for human experts. Future research should therefore investigate the possibility of comparing the features' relevant characteristics with the benchmark datasets' features.

While the proposed system design, thus far, mainly leverages simple (visual) analysis to accommodate the domain expert's capabilities (DP1), great potential lies in further analyses to complement the gained knowledge. A potential avenue for further analysis lies in unsupervised methods, which are often employed for exploratory data analysis (Ferreira de Oliveira & Levkowitz, 2003). For example, clustering methods could be utilized to identify similar features within or beyond the domain-specific groups. Finding such groups can then contribute to extracting knowledge more efficiently from the data and thus deriving promising real-world value from the datasets (Fürstenau et al., 2021).

In light of the presented challenges, the benchmark dataset matching can be a starting point for future research to develop more advanced methods to compare datasets for specific use cases. In future research, we plan to extend the focus group interviews with more experts from other organizations to generalize our findings beyond the case company and the domain. This generalization then helps build systems that improve data understanding to ultimately leverage the potential of datasets to address various organizational and societal challenges.

Conclusion

This paper addresses the timely challenge of the insufficient understanding of existing real-world datasets. Our contributions are as follows: We create design knowledge for systems that enhance data understanding and demonstrate the efficacy, desirability, and effectiveness of the proposed solution through multiple evaluations. The creation of design knowledge is based on a range of expert interviews in one of the world's largest pharmaceutical manufacturing organizations, resulting in our DRs. Building on the previously deducted DRs, we identify *data enrichment* (*semanticization*) and *benchmark dataset matching* (*contextualization*) as DPs. Using these DPs, we instantiated and evaluated a system that improves data understanding. The evaluation specifically showed that both user groups—domain experts and method experts—can accomplish their tasks more efficiently when using the system. Importantly, the true potential of the datasets is realized through the collaboration of both. Finally, generating this understanding of the data can help in unlocking the real-world value of data-rich industries for specific use cases and their data ecosystem in general.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Abdel-Karim, B. M., Pfeuffer, N., & Hinz, O. (2021). Machine learning in information systems - A bibliographic review and open research issues. *Electronic Markets, 31*(3), 643–670. https://doi.org/10.1007/s12525-021-00459-2

Abedjan, Z., Golab, L., & Naumann, F. (2015). Profiling relational data: A survey. *VLDB Journal, 24*(4), 557–581.

Alt, R. (2021). How to organize for AI? An interview with Yao-Hua Tan. *Electronic Markets, 31*(3), 639–642. https://doi.org/10.1007/s12525-021-00497-w

Arnab. (2020). *Microsoft Azure Predictive Maintenance | Kaggle*. https://www.kaggle.com/datasets/arnabbiswas1/microsoft-azure-predictive-maintenance

Axenie, C., & Bortoli, S. (2020). *Predictive maintenance dataset*. https://doi.org/10.5281/ZENODO.3653909

Blair-Early, A., & Zender, M. (2008). User interface design principles for interaction design. *Design Issues, 24*(3), 85–107.

Bokrantz, J., Skoogh, A., Berlin, C., Wuest, T., & Stahre, J. (2020). Smart Maintenance: A research agenda for industrial maintenance management. *International Journal of Production Economics, 224*, 107547. https://doi.org/10.1016/j.ijpe.2019.107547

Boyd, K. L. (2021). Datasheets for datasets help ML engineers notice and understand ethical issues in training data. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–27. https://doi.org/10.1145/3479582

Chmielinski, K. S., Newman, S., Taylor, M., Joseph, J., Thomas, K., Yurkofsky, J., & Qiu, Y. C. (2022). The dataset nutrition label (2nd gen): Leveraging context to mitigate harms in artificial intelligence. *arXiv*. https://doi.org/10.48550/arXiv.2201.03954

Choi, S. T., & Kröschel, I. (2015). Challenges of governing interorganizational value chains : Insights from a case study. *ECIS Proceedings*.

Cui, W. (2019). Visual analytics: A comprehensive overview. *IEEE Access, 7*, 81555–81573. https://doi.org/10.1109/ACCESS.2019.2923736

Dhayne, H., Haque, R., Kilany, R., & Taher, Y. (2019). In search of big medical data integration solutions - A comprehensive survey. *IEEE Access, 7*, 91265–90.

Dimitriadou, K., Papaemmanouil, O., & Diao, Y. (2016). AIDE: An active learning-based approach for interactive data exploration. *IEEE Transactions on Knowledge and Data Engineering, 28*(11), 2842–2856. https://doi.org/10.1109/TKDE.2016.2599168

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). Artificial intelligence for decision making in the era of Big Data – Evolution, challenges and research agenda. *International Journal of*

Information Management, 48, 63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021

Ehrlinger, L., Schrott, J., Melichar, M., Kirchmayr, N., & Wöß, W. (2021). Data catalogs: A systematic literature review and guidelines to implementation. *Communications in Computer and Information Science*, *1479 CCIS*, 148–158. https://doi.org/10.1007/978-3-030-87101-7_15/TABLES/3

Enders, T., Satzger, G., Fassnacht, M., & Wolff, C. (2022). Why should I share? Exploring benefits of open data for private sector organizations. *Pacific Asia Conference on Information Systems*, 1.

Ericsson, K. A., & Simon, H. A. (1993). Protocol analysis: Verbal reports as data. *Protocol Analysis*. https://doi.org/10.7551/mitpress/5657.001.0001

Etikan, I. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics, 5*(1), 1. https://doi.org/10.11648/j.ajtas.20160501.11

Fabris, A., Messina, S., Silvello, G., & Susto, G. A. (2022). Algorithmic fairness datasets: The story so far. *Data Mining and Knowledge Discovery, 36*(6), 2074–2152. https://doi.org/10.1007/s10618-022-00854-z

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review, 1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032

Fassnacht, M., Benz, C., Heinz, D., Leimstoll, J., & Satzger, G. (2023). Barriers to data sharing among private sector organizations.

Ferreira de Oliveira, M. C., & Levkowitz, H. (2003). From visual data exploration to visual data mining: A survey. *IEEE Transactions on Visualization and Computer Graphics, 9*(3), 378–394. https://doi.org/10.1109/TVCG.2003.1207445

Fürstenau, D., Klein, S., Vogel, A., & Auschra, C. (2021). Multi-sided platform and data-driven care research. *Electronic Markets, 31*(4), 811. https://doi.org/10.1007/s12525-021-00461-8

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Gregor, S., Chandra Kruse, L., & Seidel, S. (2020). The anatomy of a design principle. *Journal of the Association for Information Systems, 21*, 1622–1652. https://doi.org/10.17705/1jais.00649

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Data Mining: Concepts and Techniques. https://doi.org/10.1016/C2009-0-61819-5

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly: Management Information Systems, 28*(1), 75. https://doi.org/10.2307/25148625

Holland, S., Hosny, A., Newman, S., 4, J. J., & Chmielinski, K. (2018). *The dataset nutrition label: A framework to drive higher data quality standards*. http://datanutrition.media.mit.edu/2 http://datanutrition.media.mit.edu/demo.html

IDC. (2020). *Put more of your business data to work-from edge to Cloud*. https://www.seagate.com/files/www-content/our-story/rethink-data/files/Rethink_Data_Report_2020.pdf

Isaak, J., & Hanna, M. J. (2018). User data privacy: Facebook, Cambridge Analytica, and Privacy Protection. *Computer, 51*(8), 56–59. https://doi.org/10.1109/MC.2018.3191268

Ishwarappa, K. S., & Anuradha, J. (2015). A brief introduction on Big Data 5Vs characteristics and Hadoop technology. *Procedia Computer Science, 48*, 319–324. https://doi.org/10.1016/j.procs.2015.04.188

Jakubik, J., Vössing, M., Kühl, N., Walk, J., & Satzger, G. (2022). Data-centric artificial intelligence. *arXiv preprint* arXiv:2212.11854.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. (2014). Big data and its technical challenges. *Communications of the ACM, 57*(7), 86–94. https://doi.org/10.1145/2611567

Janiesch, C., Dinter, B., Mikalef, P., & Tona, O. (2022). Business analytics and big data research in information systems. *Journal of Business Analytics, 5*(1), 1–7. https://doi.org/10.1080/2573234X.2022.2069426

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets, 31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

Jaspert, D., Ebel, M., Eckhardt, A., & Poeppelbuss, J. (2021). Smart retrofitting in manufacturing: A systematic review. *Journal of Cleaner Production, 312,* 127555. https://doi.org/10.1016/j.jclepro.2021.127555

Kandel, S., Paepcke, A., Hellerstein, J. M., & Heer, J. (2012). Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics, 18*(12), 2917–2926. https://doi.org/10.1109/TVCG.2012.219

Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics, 8*(1), 1–8. https://doi.org/10.1109/2945.981847

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Commun. ACM, 53*(1), 148–152. https://doi.org/10.1145/1629175.1629210

King, N. (1998). Template analysis. In G. Symon & C. Cassell (Eds.), *Qualitative methods and analysis in organizational research: A practical guide* (pp. 118–134). Sage Publications Ltd.

Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: Anatomy of a research project. *European Journal of Information Systems, 17*(5), 489–504. https://doi.org/10.1057/ejis.2008.40

Labadie, C., Legner, C., Eurich, M., & Fadler, M. (2020). FAIR enough? Enhancing the usage of enterprise data with data catalogs. *Proceedings of the IEEE 22nd Conference on Business Informatics CBI 2020*, *1*, 201–210. https://doi.org/10.1109/CBI49978.2020.00029

Lee, G., & Xia, W. (2006). Organizational size and IT innovation adoption: A meta-analysis. *Information & Management, 43*(8), 975–985. https://doi.org/10.1016/J.IM.2006.09.003

Lefebvre, H., Legner, C., & Fadler, M. (2021). Data democratization: Toward a deeper understanding. *ICIS 2021 Proceedings.*

Lenz, J., Wuest, T., & Westkämper, E. (2018). Holistic approach to machine tool data analytics. *Journal of Manufacturing Systems, 48*, 180–191. https://doi.org/10.1016/j.jmsy.2018.03.003

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics. Doklady*, *10*, 707–710. https://api.semanticscholar.org/CorpusID:60827152

Matzka, S. (2020). Explainable artificial intelligence for predictive maintenance applications. *Proceedings of the 3rd International Conference on Artificial Intelligence for Industries, AI4I 2020*, 69–74. https://doi.org/10.1109/AI4I49448.2020.00023

Mayring, P. (2000). Qualitative content analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research [On-Line Journal], 1.*

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.

Mohamed, M. (2018). Challenges and benefits of Industry 4.0: An overview. *International Journal of Supply and Operations Management, 5*, 256–265. https://doi.org/10.22034/2018.3.7

Nooteboom, B. (1994). Innovation and diffusion in small firms: Theory and evidence. *Small Business Economics, 6*(5), 327–347. https://doi.org/10.1007/BF01065137/METRICS

Ofe, H., De Reuver, M., Nederstigt, B., & Janssen, M. (2023). Data analytics platforms: Value propositions and adoption challenges for small hospitality businesses. *Proceedings of the 56th Hawaii International Conference on System Sciences*, 3964–3973.

Padmanabhan, B., fang, xiao, Sahoo, N., & Burton-Jones, A. (2022). Machine learning in information systems research. *Management*

Information Systems Quarterly, *46*(1). https://aisel.aisnet.org/misq/vol46/iss1/4

Pal, S., Pramanik, P. K. D., Majumdar, T., & Choudhury, P. (2019). A semi-automatic metadata extraction model and method for video-based e-learning contents. *Education and Information Technologies, 24*(6), 3243–3268. https://doi.org/10.1007/S10639-019-09926-Y/TABLES/5

Pepper, J., Greenberg, J., Bakis, Y., Wang, X., Bart, H., & Breen, D. (2021). Automatic metadata generation for fish specimen image collections. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, *2021-September*, 31–40. https://doi.org/10.1109/JCDL52503.2021.00015

Reer, A., Wiebe, A., Wang, X., & Rieger, J. W. (2023). FAIR human neuroscientific data sharing to advance AI driven research and applications: Legal frameworks and missing metadata standards. *Frontiers in Genetics, 14*, 1086802. https://doi.org/10.3389/FGENE.2023.1086802/BIBTEX

Safder, I., Hassan, S. U., Visvizi, A., Noraset, T., Nawaz, R., & Tuarob, S. (2020). Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. *Information Processing & Management, 57*(6), 102269. https://doi.org/10.1016/J.IPM.2020.102269

Schenk, K. D., Vitalari, N. P., & Davis, K. S. (1998). Differences between novice and expert systems analysts: What do we know and what do we do? *Journal of Management Information Systems, 15*(1), 50. https://doi.org/10.1080/07421222.1998.11518195

Shankaranarayanan, G., & Even, A. (2004). Managing metadata in data warehouses: Pitfalls and possibilities. *Communications of the Association for Information Systems, 14*(1), 13. https://doi.org/10.17705/1CAIS.01413

Singh, G., Bharathi, S., Chervenak, A. L., Deelman, E., Kesselman, C., Manohar, M., Patil, S., & Pearlman, L. (2003). A metadata catalog service for data intensive applications. *ACM/IEEE SC 2003 Conference (SC'03)*, 33. https://doi.org/10.1145/1048935.1050184

Sommer, L., & Sommer, L. (2015). Industrial revolution - Industry 4.0: Are German manufacturing SMEs the first victims of this revolution? *Journal of Industrial Engineering and Management, 8*(5), 1512–1532. https://doi.org/10.3926/jiem.1470

Thong, J. Y. L., & Yap, C. S. (1995). CEO characteristics, organizational characteristics and information technology adoption in small businesses. *Omega, 23*(4), 429–442. https://doi.org/10.1016/0305-0483(95)00017-I

Tremblay, M. C., Hevner, A. R., & Berndt, D. J. (2010). Focus groups for artifact refinement and evaluation in design research. *Communications of the Association for Information Systems*, *26*. https://doi.org/10.17705/1CAIS.02627

Van Den Broek, T., & Van Veenstra, A. F. (2015). Modes of governance in inter-organisational data collaborations. *23rd European Conference on Information Systems, ECIS 2015*, *2015-May*.

Van Panhuis, W. G., Paul, P., Emerson, C., Grefenstette, J., Wilder, R., Herbst, A. J., Heymann, D., & Burke, D. S. (2014). A systematic review of barriers to data sharing in public health. *BMC Public Health, 14*(1). https://doi.org/10.1186/1471-2458-14-1144

Venable, J. R., Pries-Heje, J., & Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. In: Peffers, K., Rothenberger, M., Kuechler, B. (eds) *Design Science Research in Information Systems. Advances in Theory and Practice. DESRIST 2012*. Lecture Notes in Computer Science, vol 7286. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-29863-9_31

Venable, J. R., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems, 25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Vermeer, R. (2019). *Are you ready for data driven banking?*

Voell, C., Chatterjee, P., & Rauch, A. (2018). Closing the lifecycle loop with installed base products. *IFIP Advances in Information and Communication Technology*, 540. https://doi.org/10.1007/978-3-030-01614-2_32

Wang, R. Y. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

Whiting, L. S. (2008). Semi-structured interviews: Guidance for novice researchers. *Nursing Standard (Royal College of Nursing (Great Britain) 22*(23), 35-40.

Wirth, R., & Hipp, J. (2000). *Crisp-dm: Towards a standard process modell for data mining.*

Wu, M., Brandhorst, H., Marinescu, M.-C., Lopez, J. M., Hlava, M., & Busch, J. (2023). Automated metadata annotation: What is and is not possible with machine learning. *Data Intelligence, 5*(1), 122–138. https://doi.org/10.1162/DINT_A_00162

Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: Advantages, challenges, and applications. *Production and Manufacturing Research, 4*(1), 23–45. https://doi.org/10.1080/21693277.2016.1192517

Zeng, J., & Glaister, K. W. (2018). Value creation from big data: Looking inside the black box. *Strategic Organization, 16*(2), 105–140. https://doi.org/10.1177/1476127017697510

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.