Panzner, Melina; von Enzberg, Sebastian; Meyer, Maurice; Dumitrescu, Roman

Article — Published Version

Characterization of Usage Data with the Help of Data Classifications

Journal of the Knowledge Economy

Suggested Citation: Panzner, Melina; von Enzberg, Sebastian; Meyer, Maurice; Dumitrescu, Roman (2022) : Characterization of Usage Data with the Help of Data Classifications, Journal of the Knowledge Economy, ISSN 1868-7873, Springer US, New York, Vol. 15, Iss. 1, pp. 88-109, https://doi.org/10.1007/s13132-022-01081-z

Journal of the Knowledge Economy (2024) 15:88–109
https://doi.org/10.1007/s13132-022-01081-z

Characterization of Usage Data with the Help of Data Classifications

Melina Panzner[1] · Sebastian von Enzberg[2] · Maurice Meyer[1] · Roman Dumitrescu[1]

Received: 2 November 2021 / Accepted: 23 September 2022 / Published online: 20 October 2022

Abstract

Comprehensive data understanding is a key success driver for data analytics projects. Knowing the characteristics of the data helps a lot in selecting the appropriate data analysis techniques. Especially in data-driven product planning, knowledge about the data is a necessary prerequisite because data of the use phase is very heterogeneous. However, companies often do not have the necessary know-how or time to build up solid data understanding in connection with data analysis. In this paper, we develop a methodology to organize and categorize and thus understand use phase data in a way that makes it accessible to general data analytics workflows, following a design science research approach. We first present a knowledge base that lists typical use phase data from a product planning view. Second, we develop a taxonomy based on standard literature and real data objects, which covers the diversity of the data considered. The taxonomy provides 8 dimensions that support classification of use phase data and allows to capture data characteristics from a data analytics view. Finally, we combine both views by clustering the objects of the knowledge base according to the taxonomy. Each of the resulting clusters covers a typical combination of analytics relevant characteristics occurring in practice. By abstracting from the diversity of use phase data into artifacts with manageable complexity, our approach provides guidance to choose appropriate data analysis and AI techniques.

Keywords Usage Data · Use phase data · Data understanding · Product planning · Data analytics

This article is part of the Topical Collection on *Intelligent Technical Systems*

✉ Melina Panzner
melina.panzner@uni-paderborn.de

[1] Heinz Nixdorf Institute, University of Paderborn, Fürstenallee 11, Paderborn 33102, Germany

[2] Fraunhofer Institute for Mechatronic Systems Design, Zukunftsmeile 1, Paderborn 33102, Germany

Introduction

Recent technical developments enable the collection and analysis of huge amounts of data from cyber-physical systems (CPS) in their use phase. These data can be fed back into product planning and development, where data analytics reveals valuable insights about product performance and usage patterns. This is already a common procedure in the software domain (software usage analytics) (Menzies & Zimmermann, 2013). However, it seems not to be common for CPS.

Data analytics is a field that is highly interdisciplinary in nature that has adopted aspects from disciplines such as statistics, machine learning (ML), pattern recognition, system theory, operations research, and artificial intelligence (Runkler, 2020). By integrating analytical insights into decision-making processes, existing and future products can be optimized. These concepts form the research area of data-driven product planning (Meyer et al., 2021).

However, the integration of data analytics into the decision-making processes of product planning and product development poses major challenges for companies (Hou & Jiao, 2020; Wilberg et al., 2017). When implementing data analytics, small- and medium-sized enterprises (SMEs), in particular, face problems and challenges, such as a lack of know-how, a shortage of qualified employees, the dominance of domain specialists, and a lower awareness of topics such as data analytics and AI (Coleman et al., 2016).

This also complicates effective use of machine learning in product planning and development. In the data analytics process (e.g., CRISP-DM), it requires comprehensive knowledge and understanding at several points. After understanding the problem, an essential step is to identify and understand the data and analyze the data sources to evaluate the relevant data for the defined analysis problem (Reinhart et al., 2017). One challenge arises during data discovery and collection: Often, companies or data analysts do not have an overview of the existing data. There is a lack of knowledge about where to start and where to get the right information from, as it is scattered throughout the company (Kayser et al., 2019; Menon et al., 2005). This quickly becomes a first hurdle in the application of data analytics.

In order to identify the data that is relevant to the defined use case or to better understand and use the existing data, those involved must also have knowledge and understanding of the data. "The most important knowledge a data mining engineer uses to judge workflows and models' usefulness: understanding the meaning of the data" (Kietz et al., 2010). A certain level of data competence is, therefore, required, although this is often lacking, as well as a systematic recording of metadata (Sternkopf & Mueller, 2018). These are challenges that must be overcome, especially in data-driven product planning, since numerous, heterogeneous data and an often complex system landscape must be taken into account in the use phase. Examples of data in the use phase of the product life cycle are maintenance data, fault messages, service data, and log and measurement data (Li et al., 2015).

After understanding the data, preprocessing and modeling can start. Here, different components or methods are used, e.g., cleaning data sets, preprocessing data further, extracting domain-specific features from the data, modeling them appropriately,

and post-processing the model output (Reinhart et al., 2017; Shabestari et al., 2019). Together, these components form a specific data analytics workflow.

In order to determine and select the appropriate workflows, characteristics of the data and the analytical methods must be always taken into account in addition to the objectives of the use cases (Nalchigar & Yu, 2018).

The versatility of data and the additional variety of machine learning methods used in data-driven product planning make the setup of a workflow an extensive task, which requires expert knowledge. Therefore, a simplification for companies is required to be able to implement solutions more resource efficiently or with fewer specialists. Our aim is to facilitate an easy entry to data analysis and workflow design for data-driven product planning by providing a pre-selection of relevant methods via sample workflows. The prerequisite for this is comprehensive knowledge of relevant data sources in an aggregated way.

The fundamental research question in this context is as follows: Are there commonalities in data from data generated in the use phase or in product planning practice that suggest similar analytics processing and thus can be grouped together?

To answer this question and to build such data classifications, which can be mapped to specific analytics workflows, we use a design science research methodology approach. We first build upon an existing classification of use phase data from a product planning view to build a knowledge base of typical usage data. We then present a classification of usage data from the data analytics view, which delivers relevant data characteristics that can be used to describe the knowledge base data. Building on these artifacts, we propose a joint classification of data for data-driven product planning, whose artifacts can be assigned to appropriate workflows. Figure 1 summarizes these contributions and their connections.

The application in the context of data analytics in product planning is shown for one exemplary use case by means of a sample workflow.

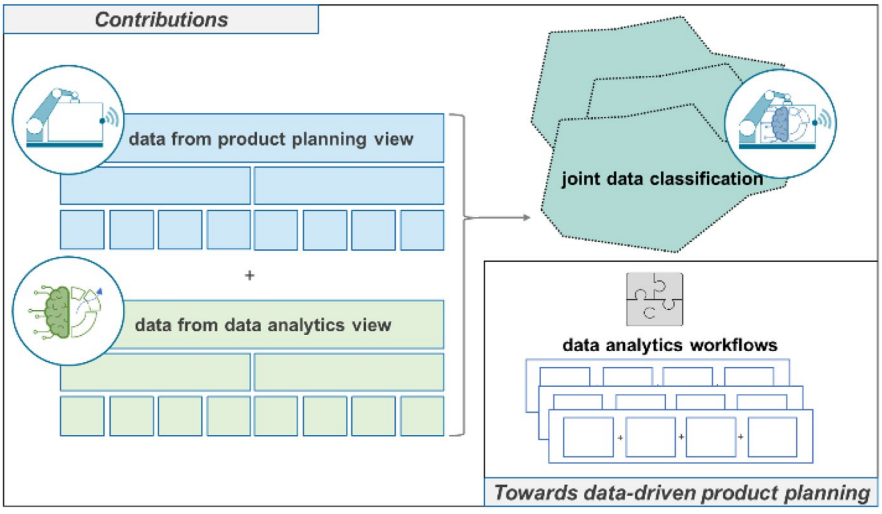

Fig. 1 Contributions

Foundations and State of the Art

In the following, we present the foundations for the data classifications and existing approaches.

Data from Product Planning View

Definitions

Various definitions for data exist (Awad & Ghaziri 2007; Bourdreau & Couillard, 1999; International Organization for Standardization, 1993; International Dama, 2017; Koohang et al., 2008; Morgenstern, 1997). In the following, we understand data as recorded interpretable signs and signals, which potentially provide information in a given context or for a specific purpose. In an industrial context, we speak of industrial data. It can be classified according to various properties.

Classifications of Industrial Data

Data in a production-oriented company can be divided into organizational and technical operational data (Kurbel, 2005). The organizational operational data includes order data and personnel data. Technical operational data are machine data, tool data, and material data. Machine data is differentiated into product and process data. The latter includes all data that is generated during the operation of a machine. Product data describes the condition of the manufactured part. In combination with process data, they encompass information about the production process as a whole.

According to Schäfer et al. data sources can be roughly divided into three groups according to the origin of the data: machine-generated and human-generated content and business data (Schäfer et al., 2012). Raffeiner proposes a classification, which distinguishes between created, received, paid, and public data (Raffeiner, 2019).

An additional subdivision of data, which is made in computer science as well as in management science, is a distinction regarding the time reference. With regard to this data constancy, a distinction can be made between “master data” and “transaction data.” The term master data refers to data that remains constant over a long period of time. This includes, for example, company data such as building or plants. In contrast to master data, transaction data is time related and changes according to known or unknown processes. Transaction data and movement data are usually related to master data (Spitta & Bick, 2008).

Another classification is offered by the automation pyramid, which represents the different levels of automation in a factory and allows the structuring of technologies into different functional layers of industrial manufacturing (Dumitrescu et al., 2015). Along these layers, data sources as IT-systems, such as sensory, PLC, SCADA, MES, and ERP, are categorized.

In addition, industrial data can be classified in terms of their occurrence in the functional areas service, marketing, work preparation, development, purchasing,

production, quality assurance, and IT (Gausemeier et al., 2009). A comprehensive product perspective is provided by categorizing data based on the product lifecycle phases product planning, design and development, production planning, production, use and support, and reuse and recycling (Kassner et al., 2015). Li et al. arrange data into the three main phases of product lifecycle management (PLM) BOL, MOL, and EOL (Li et al., 2015). Tao et al. propose another classification into management data, equipment data, user data, product data, and public data (Tao et al., 2018b). Table 1 summarizes the presented data classification approaches.

In data-driven product planning, the focus is often on the usage phase of the product life cycle and its data, often called usage data or field data (Kammerl et al., 2016; Kreutzer, 2019). Kreutzer refers to field data generated during the product or system usage phase after the point of sale (PoS) (Kreutzer, 2019). Edler defines field data as "[…] data that is generated in connection with the use of a product in the field or the use of a service by the customer. This include, in addition to errors, malfunctions, defects or failures, usage information such as machine running times, consumption of operating materials […], and the requirements expressed by the user for the next product generation." (Edler, 2001). With regard to the sources of field data, Kreutzer proposes the following classification for cyber physical systems: sensors and actuators, user data, and system data. Sensors are divided into shape and material measures, functional and process variables, and environmental interaction variables. For use in product planning, this classification is not sufficient, since data related to the CPS or the product is missing, such as service and customer data.

Data from the Data Analytics View

From a data analytics perspective, it is important to understand the (intrinsic) characteristics of the data in order to infer necessary or appropriate processing methods.

Table 1 Classification approaches for industrial data

business function	service	marketing	work preparation	development /construction	purchasing	production	quality assurance	IT
ressource/ system	sensor/actuator (field level)	PLC (control level)		SCADA/HMI (process control level)		MES (plant management level)		ERP/CRM (corporate management level)
constance	master data				transaction data			
access	created		received		public		paid	
origin	machine generated			human generated			business data	
PLM	beginning of live (BOL)			middle of life (MOL)			end of life (EOL)	
product lifecycle	concept & product planning	design & development		production planning	production		use and support	reuse and recycling
content	management data	equipment data		user data		product data		public data
others	organizational				technical			

Definitions

"Data characterization describes the data in ways useful to the miner and begins the process of understanding what is in the data—that is, is it reliable and suitable for the purpose?" (Pyle, 1999). To describe the nature of data, characteristics are needed (Kitchin & McArdle, 2016). In this context, there is also often a reference to meta data. Metadata ("data about data") refer to structured data that can be used to describe and specify facts about an information object (Dippold et al., 2005). Metadata are used to define data characteristics. This idea is common to the field of meta learning, where attributes relevant to the problem are of particular interest.

Classifications

In general, the following types of data characteristics can be distinguished: general measures (general information to the dataset at hand, such as number of instances and dimensionality) as well as statistical and information-theoretic measures (attribute statistics and class distributions, such as mean and standard deviation) (Bilalli et al., 2016).

Another approach of characterization is the characteristics of Big Data, such as quantity, variety, and speed (Zhang, 2016). According to Hildebrand et al. data can be described based on their characteristics based on six criteria. These criteria are divided into format, structure, content, stability, processing and the business object (Hildebrand et al., 2015). An important criterion in data analysis is the structure of the data, which is also hidden behind the term data variety in big data terminology. The degree of structuring determines the further processing. Quality differences and problems form another dimension (Corrales et al., 2015). In order to successfully prepare data for analysis, a large number of criteria must be taken into account. These include completeness (often a problem especially with textual data), consistency, and accuracy.

Existing classifications are usually not or only partially aligned with the data analytics requirements. Ziegenbein et al. provide a list of data set characteristics, which are related to machine learning procedures (Ziegenbein et al., 2018). Since this is not an exact fit for the requirements in data-driven product planning, a new classification is needed.

Research Methodology

In the last section, we motivated the need for structuring and concretization of data sources of the use phase, which we call usage data in the following, as well as suitable characteristics to describe them with the goal of data analysis. This is also the first activity in the design science research methodology (DSRM) presented by Peffers et al. that we followed to develop the classifications (Peffers et al., 2007). The research process is summarized in Table 2. In the following, we describe design and development of the research process for conceptualization of the classification from the product planning view, from the data analytics view, and the joint classification for data-driven product planning in more detail (see Fig. 2).

Table 2 The DSRM process

DSRM activity	Realization
Identify the problem and motivate	Data analysis implementations in data-driven product planning, especially the selection of suitable analysis workflows, require a deep understanding of the heterogeneous usage data. There is a lack of approaches that support data description and assignment to suitable analysis methods
Define objectives of a solution	The development of a complexity-reducing classification of usage data that combines the two views product planning and data analytics to select appropriate techniques more easily
Design and development	A mixed method approach was used to develop the classifications (artifacts) in an iterative way
Demonstration and evaluation	Finally, the resulting usage data classes were successfully applied in a scenario of data-driven product planning (see the "Toward Data-Driven Product Planning" section). Further evaluations are planned
Communication	Publication of research in academic papers

Data from a Product Planning View

In the context of the "Data from Product Planning View" section, different ways of classifying industrial and field data sources were introduced. For the usage data knowledge base, we used the classification according to Meyer et al. (2022), which introduces five categories of use phase data: 1. *usage data* (describe how a product is used by its customers and users), 2. *user behavior data* (summarize how users behave when utilizing the product), 3. *service data* (data dealing with problems and the quality of the product), 4. *product behavior data* (show how the product behaves and performs during operation) and 5. *status data* (describe the status and "health" of the product).

In order to extend the classification with further data objects, an intensive literature search was carried out, e.g., Li et al. (2015), Menon et al. (2005), Kassner

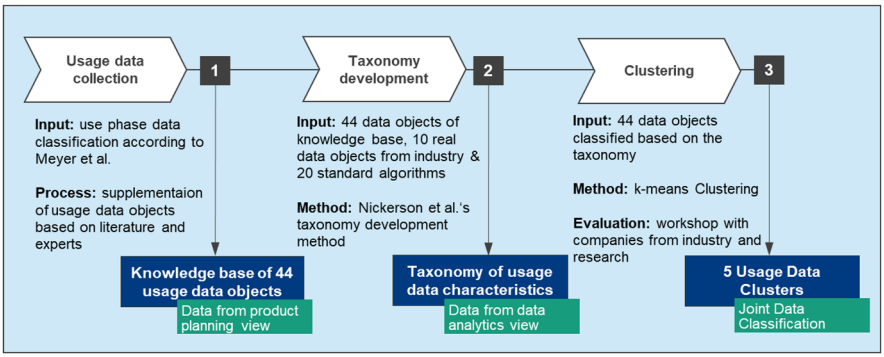

Fig. 2 Design and development of classification artifacts

et al. (2015), Kreutzer (2019), and Tao et al. (2018a). The results were enriched and validated by experts from industry and research within the research project *DizRuPt*.

Data from a Data Analytics View

The quality of insights for product planning generated from usage data highly depends on the correct usage of analytics techniques, which—in turn—is highly dependent on smart classification of the data characteristics. The type of data determines which tools and techniques can be used to analyze the data (Tan et al., 2016). So, in the following, we will attempt to answer the question "What are the key characteristics/what is the nature of usage data?". The characteristics are identified and organized using the method for taxonomy development suggested by Nickerson et al. (2013). Often used synonymously with terms such as framework, typology or classification taxonomies are empirically and/or conceptually derived groupings in terms of dimensions and characteristics (Puschel et al., 2020). Nickerson et al.'s method includes the following steps: determination of a meta-characteristic, determination of objective and subjective ending conditions, and the iterative choice of approach until all ending conditions are met. For the choice of approach, Nickerson et al. propose empirical-to-conceptual and conceptual-to-empirical approaches. In the empirical-to-conceptual approach, real-life objects are selected, characteristics are induced, given conceptual labels, and assigned to dimensions. In the conceptual-to-empirical approach, researchers first propose dimensions and characteristics before dimensions and characteristics are examined by classifying objects. This leads to an initial or revised taxonomy. Figure 3 summarizes the taxonomy development method suggested by Nickerson et al. (2013).

In line with our research question, out meta-characteristic was *analytics relevant characteristics of usage data.* We distinguished between general data set-describing characteristics, which we assume are similarly pronounced for usage data, and very individual characteristics, which are company and infrastructure dependent. We used

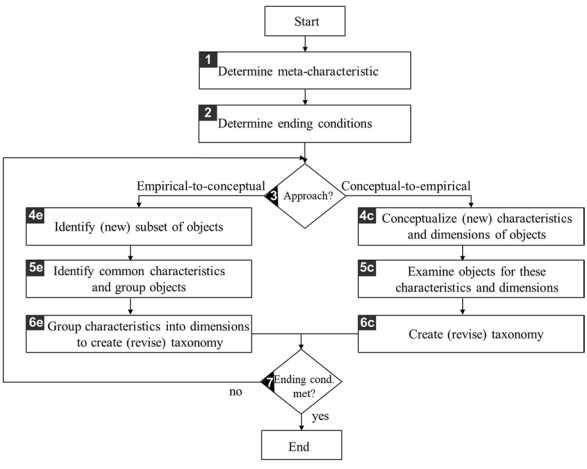

Fig. 3 Taxonomy development method according to Nickerson et al. (2013)

the objective ending conditions proposed by Nickerson et al.: every characteristic is unique in its dimension, every dimension is unique and not repeated, at least one object is classified under each characteristic of each dimension, and no new dimensions or characteristics have been added in the last iteration. Subjectively, the method will end when the taxonomy is determined by all the authors to be concise, robust, comprehensive, extendible, and explanatory. In Table 3, details of all iterations are shown. In the first iteration, we chose the conceptual-to-empirical approach to conceptualize dimensions and characteristics based on standard literature and expert knowledge by the authors. As a starting point, we chose the popular big data characteristics, which we filtered with respect to our meta characteristic. To evaluate the initial taxonomy, we used the first 22 data objects of the knowledge base (see Fig. 4). In the next iterations, we applied the empirical-to-conceptual approach. In summary, we used additional 22 data objects from the knowledge base to infer new characteristics or other constellations and 10 real usage data sets from industry to challenge the individual dimensions and characteristics. To cover the perspective of the analytics side even better, we used descriptions of 20 algorithms from the literature in the last iteration to find out if the taxonomy was final.

Joint Data Classification

The goal of this research step was to identify usage data with similar general characteristics and narrow down possible combinations of characteristics to a fixed set of artifacts. For this purpose, we combined the product planning and analytics view by using the classified data objects of the knowledge base according to the taxonomy. The assignments were again challenged with experts from research and industry who frequently work with usage data and, therefore, know their characteristics well. In the end, we obtained binary vectors that acted as input for the automated clustering. We chose a prototype-based algorithm, the well-known and most widely used clustering algorithm *k-means*, which determines a prototype for each cluster and forms clusters by assigning data objects to the closest cluster prototype (Wu, 2012). To determine the optimal number of clusters k, we used the graphical "elbow" method. That resulted in five clusters. The interpretation of these revealed each cluster could be reasonably interpreted standalone and in relation to the other clusters.

In the second step, the generated clusters were combined with possible forms of the individual characteristics to obtain a comprehensive list of usage data classes.

Approach of Data Classification for Workflow Assignment

Classification for Usage Data from a Product Planning View

Figure 4 presents the knowledge base relying on the classification by Meyer et al. (see the "Data from a Product Planning View" section). It lists 44 relevant data sources or data objects of the use phase.

Table 3 Iterations of the taxonomy development process

# It.	Approach	Actions/changes	Ending conditions	Real-life objects
1	Conceptual to empirical	- Starting with filtered (relevant for the selection of data analytics techniques) big data characteristics ("10 V's") from literature as dimensions: *variety*, *volume*, *velocity*, *veracity* (individual dimension), *viscosity*; further individual dimension: *variable type* with characteristics *categorical/qualitative* and *numerical* (Khan et al., 2018; Kitchin & McArdle, 2016) - Starting with//Addition of characteristics in *veracity: measurement and data collection errors, technical noise, outliers, missing values, inconsistent values, duplicate data* (Coleman et al., 2016; Tan et al., 2016) - Addition of characteristics *structured data, semi-structured,* and *unstructured* in dimension *variety; small, medium,* and *big* in dimension *volume,* concretization of dimension *velocity* with *real-time behavior* and characteristics *live* and *not live (batch)* - Replacement of *viscosity* by dimensions *dimensionality, distribution*, and *complexity* and addition of characteristics (Tan et al., 2016) - Replacement of *variety* characteristics by *structured data, text, image, audio*, and further subtypes (Dong & Liu, 2018; Tan et al., 2016)	*X* subjective conditions violated: - Structuring from dimension *variety* not concise and detailed enough as analytics methods require data type such as sequential data or tabular data - Not explanatory as the term *viscosity* is very general	22 data objects from knowledge base

Table 3 (continued)

# It.	Approach	Actions/changes	Ending conditions	Real-life objects
2	Empirical to conceptual	- Addition of sub-characteristic *Signals* and *Others* after *time series* - Addition of *text* sub characteristics: *unstructured* and *semi-structured* - Removal of characteristics *image*, *audio*, and *graph-based data* - Addition of characteristics *special/hybrid form – binary* and *date/time* in dimension *variable type* - Renaming of dimension *veracity* into *quality problems*	*X* subjective conditions violated: Not concise and detailed enough as object *vibration data* require special preprocessing *X* Objective conditions violated: - Empty characteristics (i.e., *graph-based data*, *image*, and *audio*) as they seem not to be relevant data types of usage data - Dimension name *veracity* not explanatory	Additional 22 data objects + 10 real usage datasets from industry
3	Empirical to conceptual	- Final fine-tuning of labels of characteristics to enable unambiguous and intuitive classification	✓ All objective and subjective conditions met: Every dimension is unique and not repeated, at least one object can be classified under each characteristic of each dimension. All authors agreed that the taxonomy was concise, robust, comprehensive, extendible, and explanatory	20 algorithms from standard literature, e.g., Tan et al. (2016), Liu (2011), and Alpaydın (2004)

Status data	Product behavior data	Usage data	User behavior data	Service data
• built-in physical elements (complex system, little variations) • built-in physical elements (simple system) • hardware configuration • hardware status (local protocol via sensory) • hardware status (transmitted protocol) • hardware status (local storage of states) • hardware status (via human/protocol) • factory settings • version numbers • current licenses • installed updates (updateprotocol) • software status (condition, configuration) • warning message (via human/protocol) • warning message (from software) • error messages • standstill message • runtime • operating mode • time and location	• actuator data for single actor (selectively controlled) • actuator data for single actor (continuosly controlled) • actuator data for complex overall system • sensor data (e.g. temperature, humidity, pressure, proximity, level, acceleration • vibration (sensor) data • energy consumption • disturbance times and downtimes • production quanitity • good quantity • scrap quantity • workload	• order and job • user activities aggregated (e.g. usage of functions)	• user activity protocol/log • usage process/interaction path • activity data via user interfaces • personal employee data • user login	• service reports (automated) • repair protocol • maintenance protocol • warranty case • customer complaints • customer reviews/ ratings • customer suggestions

Fig. 4 Usage Data Knowledge Base

Classification for Usage Data from a Data Analytics View

Figure 5a, b show the taxonomy for general and individual usage data characteristics and possible indicators for easier classification of data objects. In the following, all dimensions and characteristics are described in more detail.

General Dimensions and Characteristics

– The data set group (variety): This dimension examines data in terms of its variety, i.e., data set type and degree of structuring. Characteristics on the first layer are *tabular data* (*structured)* and *text data*. These can be broken down further. *Record data* assumes a dataset as a collection of records with a fixed set of data fields (variables). Table or matrix form is common. Generally, there is no explicit relationship among records, and every record has the same set of variables. *Graph-based data* considers data with relationships among objects or data with objects that are graphs (if objects contain sub objects that have relationships). For *ordered data*, the attributes have relationships with a temporal or spatial order. Ordered data can be grouped further into *sequential transaction data* (each transaction has a time associated with it), *sequence data* (the dataset that is a sequence of individual entities–positions instead of time stamps), *time series data* (each record is a series of measurements taken over time) with *signals* and *no signals*, and *spatial data* (spatial attributes, such as positions or areas). For text, a distinction can be made between *structured* and *semi-structured* text data. Image-, audio-, and graph-based data are grayed out because the procedure in the "Data from a Data Analytics View" section showed that they are not relevant

	Dimension	Characteristics				Indicator (exemplary)
Data characteristics (general)	Data set group (variety)	tabular data (structured)	record data (e.g. table, data matrix)			fixed number of data fields, no explicit link between entries, itemsets
			graph-based data			links, object relations
			ordered data (order in time or space)	sequential transaction data		itemsets + time specification (no particular frequency)
				sequence data		ordered without timestamp
				time series data (each record is a time series, regular intervals)	signals	measurable physical parameters
					no signals	measurements in general taken over time
				spatial data		positions, areas
		text	unstructured			missing format (no separations of the information)
			semi-structured			tags, meta data
		image				meta data
		video				meta data
	Dimensionality	small dimensional				significantly fewer features than observations
		high himensional				more features than observations, > 1000 features
	Distribution	sparse				many gaps in the data – e.g., only changes become visible (infrequent recording of data points)
		dense				continuously recorded values
	Complexity	(auto-)correlated				data points have a high dependence on their neighbors
		uncorrelated				data points have no dependence on their neighbors
	Real time behavior (velocity)	Real time / live				delivery of the data promptly after its creation
		static (batch) or not-live				the data is transferred only once per day/ the data must be tapped from the machine via usb stick
	Volume (per day)	small				rather hundred data rows
		medium				rather thousand data rows
		big				rather a hundred thousand data rows and more

a

	Dimension	Characteristics		Assesment	Indicator (exemplary)
Data characteristics (individual)	Data quality problems	systematic errors (data error caused by e.g., miscalibration)		to be neglected	constant systematic errors (vendor specification or domain knowledge) where only relations are of interest
				to consider	systematic errors that can be corrected
				dominant	intolerable sensor failures
		random noise (noise distribution does not follow a known statistics/model)		to be neglected	random errors that are rare enough to have an impact
				to consider	random errors
				dominant	random errors dominating the data
		outliers		to be neglected	none or isolated (not relevant) outliers
				to consider	outliers can be clearly identified
				dominant	faulty values predominate
		inconsistency (frequencies, unit (e.g. kg/h), value range)		to be neglected	frequencies and units are uniform
				to consider	variables have significantly different value ranges, frequencies vary
				dominant	information on units and value ranges is missing
		missing values		to be neglected	no or only occasionally missing values
				to consider	missing values are in the minority
				dominant	missing values outweigh
		duplicate data		yes	duplicates can be clearly identified
				no	no duplicates
	Variable type	categorical/qualitative	nominal		no natural order, names
			ordinal		data have natural order
		numerical			natural order, quantifiable distances
		special/hybrid form	binary		two values (1/0, yes/no)
			date/time		year, month, day, hour, minute, second, microsecond

b

Fig. 5 **a** General data characteristics; **b** individual data characteristics

as a characteristic for usage data. However, since these formats may well play a greater role in the future, they are also listed.

– Dimensionality: Dimensionality is another important factor that can play a crucial role for the selection of an adequate analytics technique, e.g., too many dimensions cause every observation in a dataset to appear equidistant from all

the others (curse of dimensionality), which is a big problem for clustering algorithms. Hence, the characteristics are *small dimensional* and *high dimensional*.
- Distribution: Some general aspects of distributions often have a strong impact, which can make modeling difficult. Sparsity is such a special case, where most attributes of an object have values of 0. Some data mining algorithms, such as the association rule mining algorithms, work well only for sparse data (Tan et al., 2016). On the other hand, some algorithms such as random forests work best on dense data.
- Complexity: Complexity in data can be expressed by, e.g., *(auto-)correlation*, which is important to know, as e.g., one of the assumptions of regression analysis is that the data has no autocorrelation. Therefore, other methods may have to be used. Correlation and multicollinearity in data may have an impact on the performance of the model, too. Algorithms, such as logistic regression or linear regression, are not well suited in that case so that it should be fixed before training.
- Real-time behavior (velocity): In data analytics or machine learning real-time or online ML (training of a model by running live data through it to continuously improve the model) can be distinguished from traditional training, where a batch of historical data is used. The former requires different procedures than the latter.
- Volume: Regarding the volume, a data object or dataset can have *small*, *middle*, or *big size*. To evaluate this, the amount of data generated per day is certainly important. The volume affects the analysis to the extent that some methods are better able to handle few training samples, e.g., support vector machines, or some algorithms are better suited to process large volumes of data.

Individual Dimensions and Characteristics

These individual characteristics are not only important mainly for the selection of the right preprocessing techniques but also play a role in the modeling algorithms (Banimustafa & Hardy, 2012).

- Data quality problems: Data quality has a major impact on data analysis, for example, some techniques are more tolerant to missing values, outliers, and unusual data distributions. Some data pre-processing procedures (e.g., outlier elimination, normalization, phasing, data reduction) may be necessary to address the quality issues and tailor the data for modeling. Characteristics are *random noise, systematic errors, outliers, inconsistency, missing values,* and *duplicate data*.
- Variable type: To describe individual data objects, the variable type is suitable. Basically, *categorical* (qualitative) and *numerical* (quantitative) attributes are distinguished here. Qualitative attributes lack most of the properties of numbers and should be treated more like symbols. Here, again, nominal and ordinal types can be distinguished. Quantitative attributes are represented by numbers and have most of the properties of numbers. *Binary* and *date* variables can be both categorical and numerical and are sub characteristics of *special/hybrid form*.

To evaluate or determine the quality characteristics and to better estimate preprocessing actions, we propose the use of a three-level scale "to be neglected," "to consider," and "dominant." Quality constraints in the context of measurement quality may be negligible, for example, if the dataset contains constant systematic errors but only relations are of interest, or if random errors are present that are rare enough to have an impact. Systematic errors that can be corrected would be to be considered in the context of preprocessing. "Dominant" is intolerable sensor failures or random errors, which dominate the data. The final evaluation of course needs also to consider the use case.

Joint Data Classification

As mentioned in "Joint Data Classification through clusters", the data basis for the clustering is the assignment or classification of the data objects from the knowledge base to the general feature characteristics according to the taxonomy by experts (see Fig. 6).We inferred five clusters or categories covering combinations of general usage data characteristics that typically occur together. The clusters are illustrated in Table 4 where we highlighted the most frequent characteristics per dimension. The names of the clusters are shaped by the most distinctive characteristics.

Cluster 1: Sequential Sparse Real-Time Data

This cluster is characterized by the dataset group *ordered data*, more specifically *time series* and *sequential transaction data.* Data objects in this cluster are, in most cases, generated in real time; data size is small to middle, and, mostly, there is no obvious correlation. Furthermore, the cluster is marked by low dimensionality and sparsity. Sensor data, in some cases also actuator data, hardware, and software states as well as warning and error messages, can often be classified here.

Cluster 2: Highly Structured Historical Data

This cluster predominantly contains structured data, which can be stored in relational databases. The data volumes are rather small, also because the data are rather

		variety (data set group)									velocity (real time behavior)		volume			
		record data	graph-based data	sequential transaction data	time series data	time series data - signals	sequence data	spatial data	text unstructured	text semi-structured	real time / live	not live	small	medium	big	
1	sensor data (e.g. temperature, pressure)	0	0	0	1	0	0	0	0	0	1	0	0	1	0	
2	vibration sensor data (e.g. acceleration)	0	0	0	0	1	0	0	0	0	1	0	0	0	1	
3	Actuator data for single actor (selectively controlled)	0	0	0	1	0	0	0	0	0	1	0	0	0	1	
4	Actuator data for single actor (continuosly controlled)	0	0	0	1	0	0	0	0	0	1	0	0	0	1	
5	Actuator data for complex overall system	0	0	0	1	0	0	0	0	0	1	0	0	0	1	
6	hardware configuration	1	0	0	0	0	0	0	0	0	0	1	1	0	0	
7	built-in physical elements (complex system, little variations)	1	0	0	0	0	0	0	0	0	0	1	1	0	0	
8	built-in physical elements (simple system)	1	0	0	0	0	0	0	0	0	0	1	1	0	0	
9	hardware status (local protocol via sensory)	0	0	0	1	0	0	0	0	0	1	0	0	1	0	
10	hardware status (transmitted protocol)	0	0	1	0	0	0	0	0	0	1	0	0	1	0	
11	hardware status (local storage of states)	1	0	0	0	0	0	0	0	0	1	0	0	1	0	
12	hardware status (via human/protocol)	1	0	0	0	0	0	0	0	0	0	1	1	0	0	
13	factory settings	1	0	0	0	0	0	0	0	0	0	1	1	0	0	
14	version numbers	1	0	0	0	0	0	0	0	0	0	1	1	0	0	
15	current licenses	0	0	0	0	0	0	0	1	0	0	1	1	0		
16	installed updates (updateprotocol)	0	0	1	0	0	0	0	0	0	1	0	1	0		
17	software status (condition, configuration)	0	0	0	0	0	0	0	0	1	1	0	1	0		
18	software status (condition, configuration)	0	0	1	0	0	0	0	0	0	1	0	1	0		

Fig. 6 The data basis for clustering (excerpt)

Table 4 Usage data clusters

	Cluster				
	Sequential sparse real time data	**Highly structured historical data**	**Mixed-structured, high-dimensiona real-time data**	**Real-time time series data**	**Text data**
Examples	Sensor data, control signals for individual actuator, hardware states, software status, warning and error message	Hardware configurations, factory settings, warnings, complaints, ratings, login data, location data	Actuator data, activity data	Vibration data, hardware status, software status, production quantity, workload, runtime, energy consumption	Licenses, protocols (service, maintenance)
Data set group	Time series, sequential transaction data	Structured (relational database schema)	Mixed	Time series (signals)	Semi-structured to unstructured text
Real-time behavior	Real time	Not real time	Real time	Real time	Not real time
Volume	Small to middle data	Small data	Middle to big data	Small data	Small data
Complexity	Uncorrelated	Uncorrelated	Uncorrelated	(Auto-)correlated	Uncorrelated
Dimensionality	Low dimensional	Low dimensional	High dimensional	Low dimensional	Low dimensional
Distribution	Sparse	Sparse	Dense	Dense	Sparse

sparse. Examples of data objects are hardware configurations, factory settings, warnings, ratings, and login data.

Cluster 3: Mixed-Structured, High-Dimensional Real-Time Data

This cluster includes data from different data set groups. Objects that can be classified here often have semi structured format, but can just as well be sequential or structured data. Other characteristics of this data are its real-time behavior and middle to big data sizes. Often, they are also high dimensional and dense.

Cluster 4: Real-Time Time Series Data

Real-time time series data are characterized by a time series format or even signal characteristics. They are generated in real time and mostly small data. Signal data such as vibration, on the other hand, often appear in large datasets. Since the focus is on time series, they are often characterized by autocorrelation, but tend to be low dimensional and dense. Vibration data, hardware, and software status, runtime, and energy consumption can be classified into this cluster.

Cluster 5: Text Data

The last cluster is characterized by an unstructured or structured text format. The amount of data is rather small. Sparsity is given by the format. Examples are licenses and various protocols.

The resulting classes can be combined with the respective data quality assessment and variable type feature (see Fig. 3b). Since this results in more than 12,000 possible combinations, these must be reduced to a few representative classes. To obtain classes that are relevant in practice, if possible, we asked six industry and research institutes in a workshop in which of the five clusters they classify their usage data and in which quality characteristics their usage data are available. Some key results are summarized in Table 5 and the resulting final data classes in Table 6.

Table 5 Workshop results

Data information	Data cluster	Quality issues	Variable type
Status data	1	- Negligible data quality problems	Binary
Sensor data (acceleration, pressure)	4	- To be considered systematic errors and random noise - Negligible missing values, inconsistency	Numerical
Log-data (temperature, configuration, error message)	3	- Negligible outliers and duplicates - To be considered missing values, inconsistency	Mixed
Service reports	5	- Dominant missing values	Nominal

Table 6 Final data classes

Class nr	Class description/characteristics
1	Sequential sparse real-time data with random noise and inconsistency
2	Highly structured historical data with duplicates and missing values
3	Mixed-structured, high-dimensional real-time data with missing values and inconsistency to be considered
4	Real-time time series data with systematic errors and random noise to be considered
5	Text data with partly dominant many missing values

Toward Data-Driven Product Planning

An exemplary use case from data-driven product planning shall illustrate the application of the joint data classification for linking to a suitable sample workflow. A popular application to improve products is failure detection and diagnosis or root cause analysis. For example, a company wants to detect frequently occurring errors on its production machine and uncover possible causes. To do this, it can use machine sensor data, such as pressure, speed, and motor current on the one hand, and service reports on the other hand, which contain error information and possible causes for some processes**.** Machine data can be categorized into data class nr. 4 from Table 6 **(**real-time time series data with systematic errors and random noise to be considered). Service reports belong to data class 5 (text data with partly dominant many missing values). A possible data analytics workflow for data with these characteristics is shown in Fig. 7 and could look as follows:

1. Selection: For detecting failures, all machine information is helpful (pressure can indicate valve damage, RPM can indicate motor damage or bearing damage, and motor current can indicate bearing damage or blockage). Since service reports have very few failure cases documented and causes are usually missing, they are more suitable for validating failure detection.

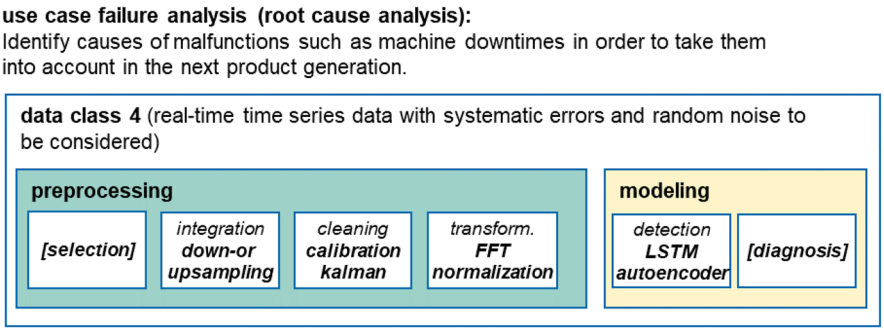

Fig. 7 Sample workflow for exemplary data class

2. Integration: The three sensor measurement data can be combined for multivariate analysis. For this, it is important that the time stamps and the sampling rates match. Here, if necessary, down- or upsampling can be used.
3. Cleaning: Due to the data class, the data suffers from measurement inaccuracies and random noise. These can be resolved by calibration or setting an offset and a filter such as Kalman.
4. Transformation: For time series, it is often worth transforming to the frequency domain to get a different perspective on the data. With respect to the detection method selected in the next step (LSTM autoencoder), the data require normalization between 0 and 1 and must be reshaped into a three-dimensional tensor.
5. Detection: Since we want to detect failures, we can frame the problem as an anomaly detection task. Since numerical time series are involved, statistical approaches or unsupervised or semi-supervised models, since mainly normal states are known, come into question. Methods, which are able to monitor several features or time signals in parallel (multivariate), are, e.g., clustering methods like DB-SCAN or K-means, ARIMA, or autoencoder. We propose an LSTM autoencoder due to its suitability for temporal data.

After detection, the diagnosis part would start. Suitable techniques can again be provided for this task.

Conclusion and Future Research

We have presented three classification schemes for data in data-driven product planning. The first classification looks at usage data from a product planning view. The resulting knowledge base lists typical data of the usage phase and offers an overview about relevant data for data-driven product planning use cases. The second classification looks at data from a data analytics view by summarizing characteristics that are relevant to preprocessing and data analytics algorithm selection. Finally, these two approaches were combined by assigning the characteristics to the typical data sources and doing a cluster analysis on it. The resulting data classes can be used in data-driven product planning to match to sample workflows. This greatly simplifies the task of understanding data and selecting appropriate analytics techniques. We illustrated the utilization of the classes and that the data classes are useful to derive abstracted, generally valid sample analytics workflows for data-driven product planning, with an application example. The development of such workflows requires future work. This can only take place for selected data classes. For this purpose, the important classes and their most frequent quality ratings must be identified. Furthermore, these workflows can only provide initial impetus, since not all factors to be considered, such as concrete domain knowledge, can be covered.

Funding Open Access funding enabled and organized by Projekt DEAL. This work is funded by the German Federal Ministry of Education and Research (BMBF).

References

Alpaydın, E. (2004). Introduction to machine learning, adaptive computation and machine learning series. *MIT Press*.

Awad, E., & Ghaziri, H. (2007). *Knowledge management*. Pearson Education India.

Banimustafa, A., & Hardy, N. (2012). A strategy for selecting data mining techniques in metabolomics. *Methods in Molecular Biology (Clifton, N.J.)*, *860*, 317–333.

Bilalli, B., Abelló, A., Aluja-Banet, T., & Wrembel, R. (2016). Towards intelligent data analysis: The metadata challenge, *undefined*.

Bourdreau, A., & Couillard, G. (1999). Systems integration and knowledge management. *Information Systems Management, 16*(4), 24–32.

Coleman, S., Goeb, R., Manco, G., Pievatolo, A., Tort-Martorell, X., & Reis, M. (2016). How can SMEs benefit from big data? Challenges and a path forward: S. Coleman et al. *Quality and Reliability Engineering International*, *32*.

Corrales, D. C., Ledezma, A., & Corrales, J. C. (2015). A conceptual framework for data quality in knowledge discovery tasks (FDQ-KDT): A proposal. *JCP, 10*(6), 396–405.

Dippold, R., Meier, A., Schnider, W., & Schwinn, K. (2005). *Unternehmensweites Datenmanagement: Von der Datenbankadministration bis zum Informationsmanagement; Zielorientiertes Business-Computing,* 4, überarb. und erw. Aufl, Vieweg, Braunschweig, Wiesbaden.

Dong, G., & Liu, H. (2018). *Feature engineering for machine learning and data analytics*. CRC Press.

Dumitrescu, R., Gausemeier, J., Kühn, A., Luckey, M., Plass, C., Schneider, M, & Westermann, T. (2015), Auf dem Weg zur Industrie 4.0: Erfolgsfaktor Referenzarchitektur. *It's OWL Clustermanagement*.

Edler, A. (2001). Nutzung von Felddaten in der qualitätsgetriebenen Produktentwicklung und im Service.

Gausemeier, J., Plass, C., & Wenzelmann, C. (2009). *Zukunftsorientierte Unternehmensgestaltung - Strategien, Geschäftsprozesse und IT Systeme für die Produktion von morgen*. Munich/Vienna: Carl Hanser Verlag.

Hildebrand, K., Gebauer, M., Hinrichs, H., & Mielke, M. (Eds.). (2015). *Daten- und Informationsqualität: Auf dem Weg zur Information Excellence, 3* (erweiterte). Springer Vieweg.

Hou, L., & Jiao, R. J. (2020). Data-informed inverse design by product usage information: A review, framework and outlook. *Journal of Intelligent Manufacturing, 31*(3), 529–552.

International, D. (2017). *DAMA-DMBOK: Data management body of knowledge* (2nd ed.). Technics Publications.

International Organization for Standardization. (1993). *ISO/IEC 2382–1:1993 Information technology — Vocabulary — Part 1: fundamental terms*.

Kammerl, D., Novak, G., Hollauer, C., & Mörtl, M. (2016). Integrating usage data into the planning of product-service systems. In *2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)* (pp. 375–379).

Kassner, L., Gröger, C., Mitschang, B., & Westkämper, E. (2015). Product life cycle analytics – Next generation data analytics on structured and unstructured data. *Procedia CIRP, 33*, 35–40.

Kayser, L., Mueller, R., & Kronsbein, T. (2019). Data collection map: A canvas for shared data awareness in data-driven innovation projects.

Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A. A., & Salehian, S. (2018). The 10 Vs, issues and challenges of big data. In *Proceedings of the 2018 International Conference on Big Data and Education* (pp. 52–56).

Kietz, J., Serban, F., Bernstein, A.. & Fischer, S. (2010). Data mining workflow templates for intelligent discovery assistance and auto-experimentation.

Kitchin, R., & McArdle, G. (2016). What makes big data, big data? Exploring the ontological characteristics of 26 datasets. *Big Data & Society, 3*(1), 2053951716631130.

Koohang, A., Harman, K., & Britz, J. (2008). *Knowledge management: Theoretical foundations, knowledge management / Alex Koohang*. Informing Science Press.

Kreutzer, R. (2019). *Methodik zur Bestimmung der Nutzenpotenziale von Felddaten cyber-physischer Systeme*. Dissertation, RWTH Aachen; IIF - Institut für Industriekommunikation und Fachmedien GmbH.

Kurbel, K. (2005). *Produktionsplanung und-steuerung im enterprise resource planning und supply chain management*, Oldenbourg Verlag.

Li, J., Tao, F., Cheng, Y., & Zhao, L. (2015). Big data in product lifecycle management. *The International Journal of Advanced Manufacturing Technology, 81*(1), 667–684.

Liu, B. (2011). *Web data mining: Exploring hyperlinks, contents, and usage data. SpringerLink Bücher*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Menon, R., Tong, L. H., & Sathiyakeerthi, S. (2005). Analyzing textual databases using data mining to enable fast product development processes. *Reliability Engineering & System Safety, 88*(2), 171–180.

Menzies, T., & Zimmermann, T. (2013). Software analytics: So what? *Software, IEEE, 30*, 31–37.

Meyer, M., Panzner, M., Koldewey, C., & Dumitrescu, R. (2022). 17 use cases for analyzing use phase data in product planning of manufacturing companies, in *Procedia CIRP.*

Meyer, M., Wiederkehr, I., Koldewey, C., & Dumitrescu, R. (2021). Understanding usage data-driven product planning: A systematic literature review. *Proceedings of the Design Society, 1*, 3289–3298.

Morgenstern, B. (1997). Definitionen und Begriffe der Informationsverarbeitung nach DIN 41859 und DIN 44300. in Morgenstern, B. (Ed.), *Elektronik: Für Elektrotechniker ab 1. Semester*, *Studium Technik*, 2, überarb. Aufl, Vieweg, Braunschweig, pp. 1–2.

Nalchigar, S., & Yu, E. (2018). Business-driven data analytics: A conceptual modeling framework. *Data & Knowledge Engineering, 117*.

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems, 22*(3), 336–359.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems, 24*(3), 45–77.

Puschel, L. C., Roglinger, M., & Brandt, R. (2020). Unblackboxing smart things—A multilayer taxonomy and clusters of nontechnical smart thing characteristics. *IEEE Transactions on Engineering Management*, 1–15.

Pyle, D. (1999). *Data preparation for data minin.* morgan kaufmann.

Raffeiner, M. (2019). Erkunden Sie Ihre Datenlandschaft. Datentreiber. https://www.datentreiber.de/blog/erkunden-sie-ihre-datenlandschaft/

Reinhart, F., Kühn, A., & Dumitrescu, R. (2017). Schichtenmodell für die Entwicklung von Data Science Anwendungen im Maschinen- und Anlagenbau. In *Wissenschaftsforum Intelligente Technische Systeme (WInTeSys)*, Heinz Nixdorf MuseumsForum, 321–334.

Runkler, T. A. (2020). *Data analytics: Models and algorithms for intelligent data analysis* (3rd ed.). Wiesbaden: Springer Vieweg.

Schäfer, A., Knapp, M., May, M., Voß, A., & für Intelligente Analyse und Informationssysteme IAIS, Fraunhofer Institut. (2012). Big Data – Vorsprung durch Wissen – Innovationspotenzialanalyse.

Shabestari, S. S., Herzog, M., & Bender, B. (2019). A survey on the applications of machine learning in the early phases of product development. *Proceedings of the Design Society: International Conference on Engineering Design*, *1*, 2437-2446.

Spitta, T., & Bick, M. (2008). *Informationswirtschaft: Eine Einführung*. Springer-Verlag.

Sternkopf, H., & Mueller, R. M. (2018). Doing good with data: Development of a maturity model for data literacy in non-governmental organizations. In *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Tan, P. N., Steinbach, M., & Kumar, V. (2016). *Introduction to data mining.* Pearson Education India.

Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., & Sui, F. (2018a). Digital twin-driven product design, manufacturing and service with big data. *The International Journal of Advanced Manufacturing Technology, 94*(9), 3563–3576.

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018b). Data-driven smart manufacturing. *Journal of Manufacturing Systems, 48*, 157–169.

Wilberg, J., Triep, I., Hollauer, C., & Omer, M. (2017). Big data in product development: Need for a data strategy. In *2017 Portland International Conference on Management of Engineering and Technology (PICMET)* (pp. 1–10).

Wu, J. (2012). Cluster analysis and K-means clustering: An introduction. In Wu, J. (Ed.). *Advances in K-means clustering: A data mining thinking*, Zugl: Tsinghua Univ Diss, 2010, *Springer Theses*, Springer, Heidelberg, pp. 1–16.

Zhang, L. (2016). *Big data analytics for fault detection and its application in maintenance*. Luleå University of Technology, 2016.

Ziegenbein, A., Stanula, P., Metternich, J., & Abele, E. (2018). Machine learning algorithms in machining: A guideline for efficient algorithm selection. In *Congress of the German Academic Association for Production Technology* (pp. 288–299). Springer, Cham.