Towards a taxonomy of data heterogeneity

Jan Roeder, Jan Muntermann, Thomas Kneib

Angaben zur Veröffentlichung / Publication details:

Roeder, Jan, Jan Muntermann, and Thomas Kneib. 2020. "Towards a taxonomy of data heterogeneity." In *WI2020 - Entwicklungen, Chancen und Herausforderungen der Digitalisierung, Band 1: Proceedings der 15. Internationalen Tagung Wirtschaftsinformatik 2020*, edited by N. Gronau, M. Heine, H. Krasnova, and K. Pousttchi, 293–308. Berlin: GITO. https://doi.org/10.30844/wi_2020_c6-roeder.

Towards a Taxonomy of Data Heterogeneity

Jan Roeder[1], Jan Muntermann[1], and Thomas Kneib[2]

[1] University of Goettingen, Chair of Electronic Finance and Digital Markets, Goettingen, Germany
jan.roeder@uni-goettingen.de
muntermann@wiwi.uni-goettingen.de
[2] University of Goettingen, Chair of Statistics, Goettingen, Germany
tkneib@uni-goettingen.de

Abstract. The increasing diversity of data available today poses a multitude of challenges to researchers and practitioners. Data understanding, i.e., describing, exploring, and verifying a data set at hand, becomes a critical process during which it is examined if data complies with the actual user needs. With an increasing complexity of the data universe accessible by organizations and decision-makers, this task has become even more important and challenging. Building on insights from information systems research, computer science, and statistics, we develop and evaluate a taxonomy of data heterogeneity for addressing this challenge. The proposed taxonomy provides a foundation for exploring the properties of data sets. Thereby, it is relevant for both researchers and practitioners as it provides a useful tool for describing and ultimately understanding data sets. We illustrate the effectiveness of our taxonomy by applying it to data sets available to the research community and industry.

Keywords: data science, data heterogeneity, data understanding, taxonomy, information value chain

15th International Conference on Wirtschaftsinformatik,
March 08-11, 2020, Potsdam, Germany

https://doi.org/10.30844/wi_2020_c6-roeder

1 Introduction

Due to an increasing digitization, a multitude of semi-structured and unstructured data are gaining in importance, along with the presence of traditional tabular data stored in spreadsheet files or relational databases. Today's relevance of data is underlined by the statement “data is the new oil”, which goes back to the mathematician Clive Humby [1]. This assessment is quoted in many publications today and is often controversially discussed. Regardless of whether this metaphor is completely apt, i.e. the multistage process of processing oil and data is comparable, today virtually all globally active organizations have realized that relevant and potentially important insights can be generated through the collection, processing, storage, and analysis of data. While traditional structured data has been stored and used on a large scale in relational databases and enterprise resource planning (ERP) systems for decades, less structured data such as images, videos, text, social media contributions or audio signals pose a greater challenge for an automated processing. For exploiting such data, it is equally relevant for practitioners and researchers to be able to understand the individual properties of data sets.

In order to structure these diverse manifestations of data, we develop a taxonomy of data heterogeneity. In particular, the question arises which dimensions and corresponding characteristics of data are useful for classifying data in terms of their heterogeneity. Such a classification scheme contributes to data understanding, which is an important prerequisite for an effective use and analysis of data. This applies to various actors who work with the data, i.e., store, clean or analyze it. An exemplary use case is duplicate detection. This can be almost trivial with well-structured data of small volume. Large amounts of video or audio data quickly present challenges that require very different and more sophisticated approaches. To this end, we build on knowledge from the fields of information systems, computer science, and statistics. We adapt a data-focused view in which the focus is on whether and how the data can be stored and processed. An appropriate systematization of the area of data heterogeneity can serve as a helpful foundation for identifying similarities between data sets or to identify their unique properties. Thus, we formulate the following research question:

RQ1: What are the theoretically grounded and empirically validated dimensions and characteristics to describe and classify heterogeneity of data sets?

The paper is structured as follows. First, we introduce the theoretical background. This is followed by a general description of the procedure for taxonomy development. Then, we provide a description of the individual iterations of the development process of the taxonomy. Afterwards, the applicability of the taxonomy is demonstrated. Finally, implications of the findings for research and practice are discussed.

https://doi.org/10.30844/wi_2020_c6-roeder

2 Theoretical Background

2.1 Information Value Chain and Data Heterogeneity

In an organizational context, the path from recording and storing data to factually generating knowledge based on which decisions can be made and actions can be taken involves multiple sub-steps for which different processes and employees with diverse competencies are required [2]. The information value chain in Figure 1 aims to capture the general steps of transforming and using data to support decision making and the subsequent execution of actions. The availability of large, diverse, and increasingly unstructured data sets transforms the traditional information value chain and involves new sets of people, processes and technologies [2, 3]. This does come with various challenges such as a potentially (too) large volume of data or a lack of veracity. With regard to the information value chain shown in Figure 1, we focus particularly on the first step, namely *data*.

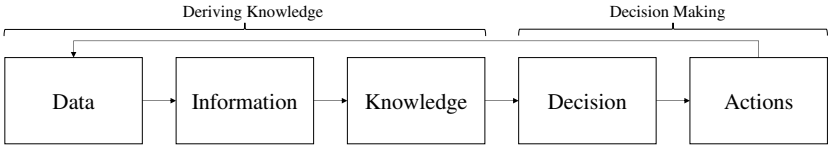

Figure 1. Information Value Chain [2]

In the context of data mining, one standard process is the Cross-Industry Standard Process for Data Mining (CRISP-DM) consisting of: 1) Business Understanding, 2) Data Understanding, 3) Data Preparation, 4) Model Building, 5) Testing, and 6) Deployment [4]. This process is iterative in nature and at the same time each subsequent step builds on the prior ones [5]. Thus, extensive data understanding is conducive and necessary for steps like data preparation and model building, e.g., by considering hierarchical or network relationships. This underlines the importance of this contribution, since a taxonomy of data heterogeneity could help researchers and practitioners alike to situate the data they are confronted with.

Heterogeneity typically refers to something consisting of dissimilar or diverse constituents [6]. It is relevant to a wide variety of areas. In the field of *statistical physics* and *economics*, the ability to quantify heterogeneity plays a prominent role. Two well-known measures to quantify the statistical heterogeneity are Shannon's entropy and Gini's index. Shannon's entropy is used to quantify the randomness of a probability law and is commonly utilized in statistical physics [7]. In comparison, Gini's index finds use in economics research and measures the evenness of a probability law [7]. *Sociology* defines heterogeneity as "differences in many or all of the characteristics of a group." [8]. For *statistics*, the term heterogeneity is commonly used in two contexts. First, meta-studies refer to heterogeneity to describe differences in the inferred treatment effect [9]. Second, in the case of panel data, fixed and random effects models are used to account for the unobserved heterogeneity between individuals [10]. In *database research*, a sub-field of computer science, heterogeneity

https://doi.org/10.30844/wi_2020_c6-roeder

of data often plays an important role in the context of schema integration, which aims to create a mapping between two or more database schemas [11]. The explanations above show that the understanding of heterogeneity may vary significantly between the domains. However, on the basis of the overview we have compiled, we define the term for this paper as follows: *"Data heterogeneity can be understood as a concept covering the qualitative differences within or between data sets that may involve different configurations along the information value chain."* The positioning of this paper is rather conceptual. We are concerned with distinct dimensions that help to characterize different aspects of data heterogeneity. However, the focus is not on the subsequent analysis steps (knowledge generation) or the value of the (strategic) business value of data for a specific task. While these are important aspects by themselves, they are out of scope for this analysis.

Wu, Zhu, Wu and Ding [12] investigate heterogeneity of *big data* and name large volume, autonomous sources with decentralized control, and the complex relationship among data as important aspects of *big data*. Regarding existing work on classification frameworks somehow related to data heterogeneity, Ranjan [13] analyzes the 10V model of big data and the importance of the dimensions in the context of different industry sectors. However, some identified dimensions do not seem relevant to data heterogeneity like *visualization*, which is a downstream task, or *value*, which is hard to quantify objectively. Kitchin and McArdle [14] apply a previously developed taxonomy of big data traits to 26 data sets. The comparison between survey, administrative, and big data hints at a rather statistical background of the paper. They identify relevant traits like exhaustivity or relationality. A related research project aims to develop a classification framework for big data against the background of IT project success [15]. However, the suggested model is limited due to the assumption that each dimension must have exactly three characteristics. Another article proposes a taxonomy for "dirty data", i.e. data that is wrong or has non-standard representations [16]. Because the model only considers structured data (numbers and strings), it is not suitable to account for the heterogeneity of today's data.

Thematically more distant work develops a taxonomy of data collaboratives. Relevant dimensions are type of data (e.g. natural phenomena) and content of data (e.g. words or locations) [17]. An article on the heterogeneity of IT landscapes describes a practicable measure to determine the diversity of relevant elements, such as vendor or product [18]. Lafky [19] investigates data heterogeneity in research networks and finds that the degree of data heterogeneity is, for example, determined by how much the data model deviates from standard design elements [19]. Based on the provided overview of related literature we identify a research gap with respect to a classification scheme that illuminates along which dimensions data heterogeneity can be differentiated.

2.2 Classification Schemes and Taxonomies

The grouping of similar elements based on certain characteristics is a fundamental and at the same time important task of research. This is not exclusive to information

https://doi.org/10.30844/wi_2020_c6-roeder

systems research but also applies to biology, archaeology, and many other disciplines [20]. Grouping objects into classes or categories is particularly important because it helps to structure a subject area and thus advances research by helping to uncover the interrelations between the various elements [21]. An inherently difficult and at the same time useful task is to identify dimensions and characteristics that not only enable the successful classification of objects but also are interesting and useful [22]. The term *classification* can refer both to the actual process of performing a classification but also to the resulting classification [20]. In the past a taxonomy was seen as being of empirical nature, while a typology was characterized as emerging from a conceptual approach [22]. In this paper, we adapt an integrative perspective which combines inductive and deductive taxonomy building [20]. Further, we follow the position of Gregor [23] after which taxonomies provide an analysis and description of the phenomena of interest (theory for analyzing).

3 Research Method

For developing the taxonomy of data heterogeneity, we adopt the approach to taxonomy building described by Nickerson, Varshney and Muntermann [20]. A taxonomy T is defined as a set of n dimensions D_i (i=1, …, n). Each of these dimensions consists of k_i ($k_i \geq 2$) mutually exclusive and collectively exhaustive characteristics C_{ij} (j = 1, ..., k_i). Consequently, for each object that is classified into the taxonomy, exactly one characteristic is assigned for each dimension D_i. Alternatively, this can be expressed as follows:

$$T = \{D_i, i = 1, \ldots, n | D_i = \{C_{ij}, j = 1, \ldots, k_i, k_i \geq 2\}\} \quad (1)$$

Figure 2 illustrates the individual process steps of the taxonomy development [20, 22].

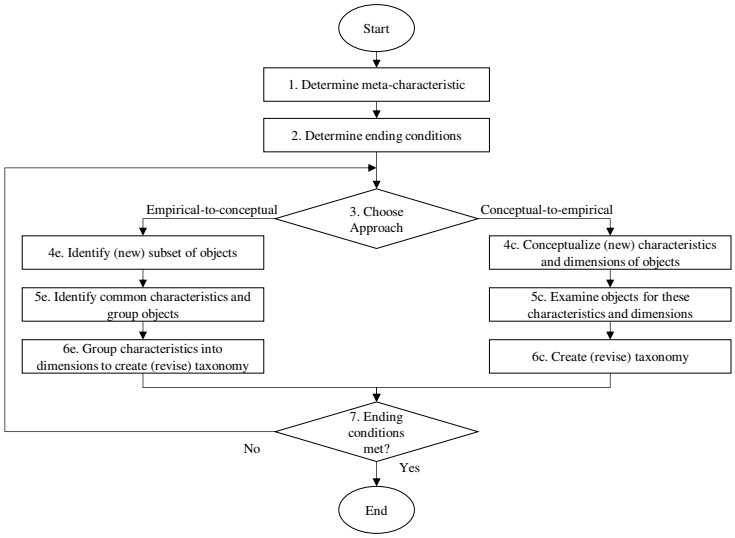

Figure 2. Taxonomy development method [20]

https://doi.org/10.30844/wi_2020_c6-roeder

The initial step is to define a meta-characteristic (step 1). This forms the basis for the selection of all following characteristics, as it influences the course of the taxonomy development. Since the taxonomy development process is inherently iterative, it is also necessary to define objective and subjective ending conditions (step 2). While the decision whether an objective ending condition is met is unambiguous in many cases, it is more difficult to decide in the case of subjective conditions. Both objective and subjective ending conditions are mentioned in Table 1.

The iterative nature of the taxonomy development process becomes apparent in steps three and seven. A choice must be made between the two approaches "empirical-to-conceptual" and "conceptual-to-empirical" (step 3). If comparatively limited data is available but well-founded insights exist, then the conceptual-to-empirical approach is recommended. New characteristics or dimensions are deduced based on the available theoretical foundation (step 4c), objects are examined for the identified aspects (step 5c) and the taxonomy is updated if necessary (step 6c). The researcher iterates over the objects, analyzes them with respect to the conceptual insights, and updates the taxonomy with respect to the changes in dimensions and characteristics.

If there is only little prior knowledge to build on, but sufficient data available, then choosing the empirical-to-conceptual approach is sensible. In this case, the researcher chooses and analyzes a subset of the data (step 4e) in order to identify suitable characteristics that can be grouped via dimensions (step 5e). This provides the basis for an update to the existent taxonomy (step 6e).

At the end of each iteration the objective and subjective ending conditions are checked. (step 7). If not all defined conditions are met, the researcher proceeds with step 3. Otherwise, the development process is completed.

4 Research Process

4.1 Taxonomy Development

The purpose of the taxonomy is a central determinant for choosing an appropriate meta-characteristic [20]. The overarching goal of our taxonomy is to capture different aspects of data heterogeneity in order to enhance the understanding of different kinds of data sets. Consequently, we specify *properties of data heterogeneity in diverse data sets* as our meta-characteristic. Further, we apply both objective and subjective ending conditions as proposed by Nickerson, Varshney and Muntermann [20] (Step 2, see Table 1). For our taxonomy development, we iteratively followed both conceptual-to-empirical (1st, 2nd, and 3rd iteration) and empirical-to-conceptual (4th iteration) approaches (Step 3) to determine and refine dimensions and characteristics of our taxonomy (Steps 4, 5, and 6). After each iteration, we checked if ending conditions were met (Step 7) and arrived at our final taxonomy after completing the 4th iteration. The following Table 1 provides an overview of our taxonomy development process.

https://doi.org/10.30844/wi_2020_c6-roeder

Table 1. Overview of iterations and objective and subjective ending conditions [20]

Iteration				*Ending Condition*
1	*2*	*3*	4	*Objective condition*
			●	Mutually exclusive: All objects have no more than one characteristic per dimension
			●	Collectively exhaustive: For all objects, a characteristic can be assigned to each dimension
			●	All relevant objects were analyzed
			●	No merge or split of object
			●	Each characteristic was assigned at least once
			●	No new dimension or characteristic was added
●	●	●	●	No dimension was merged or split
●	●	●	●	Every dimension is unique
●	●	●	●	Every characteristic per dimension is unique
●	●	●	●	No duplicate combinations of characteristics
				Subjective condition
●	●	●	●	Concise: Dimensions and characteristics are limited
			●	Robust: Sufficient number of dimensions and characteristics
			●	Comprehensive: Identification of all (relevant) dimensions of an object
●	●	●	●	Extendable: Possibility to easily add dimensions and characteristics in the future
			●	Explanatory: Dimensions and characteristics sufficiently explain the object

As a first step, we decided to build upon the existing literature that originates from the domains of information systems, statistics, and computer science and therefore decided to follow a conceptual-to-empirical approach. We conducted a literature review using the bibliographic databases AIS Electronic Library, Business Source Premier via EBSCOhost, JSTOR, ScienceDirect, SpringerLink and WISO. The search terms "data typology", "data taxonomy", "data heterogeneity", "taxonomy of heterogeneity", "taxonomy data", "big data taxonomy", "big data framework", and "data classification" were joined with the Boolean *OR* operator. Applying a filter for peer-reviewed articles resulted in 1030 search results. 40 articles were found to be potentially relevant based on the title and abstract, while 14 articles were found to be relevant after a full analysis of the content. Based on this analysis, we integrated the three strands we identified in the literature: The rather traditional perspective on (i) big data, (ii) deliberations about big data in statistics, and (iii) the perspective on data quality.

In our *first iteration* of taxonomy development, we built upon the existing literature on big data. Consequently, we integrated dimensions typically used to characterize *big data*, namely: *volume*, *velocity*, *variety* and in addition *veracity* and *value* in the more recent literature [24-26]. First, the *volume* (D_1) of the data is one of the most fundamental and intuitive dimensions to which we assign the characteristics $C_{1,j}$ =

https://doi.org/10.30844/wi_2020_c6-roeder

{fits into RAM, fits onto hard disk or must be stored in a distributed manner}. Since these characteristics are closely related to the computing capacities available, the organizational context needs to be considered when applying the taxonomy. Second, we integrated *velocity* (D_2), i.e., both the speed at which data is generated and the speed with which it is processed [24, 25]. Here, a distinction can be made between $C_{2,j}$ = {static data, data updated at defined intervals, data updated at irregular intervals, and continuous stream of data} [26]. As data originates from a variety of sources, the aspect of *variety* (D_3) plays an increasingly important role [27]. We adopt the common differentiation between $C_{3,j}$ = {structured, semi-structured or unstructured data}. The dimension *veracity* is not included in this taxonomy because the focus in the present analysis is not on the downstream use of the data and veracity is tightly coupled with steps later in the information value chain. Equivalently, the *value* dimension is not included for the same reasons.

In our *second* conceptual-to-empirical *iteration* of our taxonomy development, we built upon the literature with a more statistical background. Kitchin [28] mentions different dimensions that can be useful for highlighting the differences between "small" and "big" data for official statistics [29]. Here, the three established dimensions (as delineated in the 1st Iteration) are complemented with the dimensions *exhaustivity, resolution*, *relationality*, and *flexibility*. The aspect of *exhaustivity*, i.e., the extent to which the available data represents the entire (statistical) population [29], seems to be a crucial aspect of data heterogeneity [14]. Nevertheless, it is difficult to assess this issue exclusively at the specified *data* level of the information value chain. Hence, we do not include this dimension in the taxonomy. The dimension *resolution* can alternatively be referred to as *granularity* (D_4) [30]. It expresses what scale data points in a data set refer to [30]. This dimension is integrated since the resolution or granularity is a key characteristic of a data set and fundamentally determines whether it may be suitable for certain purposes (e.g., yearly vs. intraday sales numbers). We include it with the characteristics $C_{4,j}$ = {fine-grained, medium-grained, coarse-grained}. *Relationality* (D_5) expresses whether data can be joined with complementary data using unique identifiers [29]. This aspect plays a crucial role, especially in the context of relational databases. For example, knowing the ticker of a stock instantly unlocks a wide range of additional data that could be joined and integrated. It is included with the characteristics $C_{5,j}$ = {self-contained, intersecting}. The last mentioned dimension, *flexibility*, which asserts that in small data, i.e., data collection in the field, the data model is rigid and rather flexible in big data, is not added to the taxonomy dimensions, as we do not consider this to be a central aspect of data heterogeneity. Furthermore, we include the dimension *scale* (D_6), which captures an intrinsic hierarchy or represents the existence of a multi-level structure [31]. It is important to be aware of the hierarchical structures contained in the data, not only for data storage, but also for downstream steps like data cleaning and analysis. A distinctly hierarchical structure is a key property of a data set. Therefore, these characteristics are defined to be the constituents: $C_{6,j}$ = {single scale, multiple scales}.

In our *third* conceptual-to-empirical *iteration* we built upon additional insights from research on data quality and statistics. Here, we consider the *representational consistency* (D_7) highly relevant to characterize data heterogeneity [32]. The

https://doi.org/10.30844/wi_2020_c6-roeder

representational consistency describes whether data is represented in a consistent format and whether new data from the same source can be added easily. In contrast, if there is a lack of consistency, complex transformations and integration steps may be required to add new data to the established data set, making reliable integration more difficult. The following characteristics are added to this dimension $C_{7,j}$ = {consistent, inconsistent}. Additionally, the dimension *accuracy* (D_8) is relevant for understanding data heterogeneity [32]. Accurate data tends to be error-free and reliable. Data sets can vary substantially in terms of data accuracy. Therefore, it is even more important to be aware of the shortcomings of each data set. This kind of data quality is intrinsic, as it is not strongly bound to the context in which the data is processed or analyzed [32]. The included characteristics are defined as $C_{8,j}$ = {reliable, sporadic errors, error-prone}.

In our *fourth iteration* of the taxonomy development, we proceed with the empirical-to-conceptual approach. Here, the current version of the taxonomy is applied to empirical data to potentially adjust the dimensions and characteristics identified before. This serves the purpose to prune excess characteristics or include important characteristics missing so far. Ten competitions published by Kaggle, a platform aimed at data scientists and machine learning practitioners [33], serve as a data basis[1]. We used the most recent competitions and chose to undersample image classification tasks as they tend to be overrepresented on Kaggle. The data types of the analyzed data include tabular data, images, audio, and natural language texts. The constituent data sets of each competition are classified with respect to each dimension of the taxonomy resulting from the third iteration. An aspect that is inherent to the classification of such empirical data is the fact that it is often fuzzy and not always distinct to which characteristic a subject belongs. If no more detailed information is provided, we chose to adhere closely to the information contained in the description of the competition. As an example, in case of the *Recursion Cellular Image Classification* challenge, the provided data amounts to 46 Gigabytes of compressed image data, which would fit a HDD rather easily. However, the authors of the competition explicitly highlight that this is only a small subset of the actually relevant data, which amounts to multiple Petabytes. Thus, the actual use-case requires clearly exceeds the capabilities of even high-end workstations. A similar issue was apparent in case of the *Two Sigma: Using News to Predict Stock Movements* competition. For the competition, a relatively self-contained data set was provided by the creators. However, financial data that includes unique identifiers of financial instruments is highly intersecting and typically brought together with a multitude of supplementary measures and variables. Another aspect to consider regarding D_1 *volume* is the fundamental distinction between the storage requirements of the data itself and complex models that may be used in subsequent steps of the information value chain.

[1] Used competitions: https://www.kaggle.com/c/{placeholder}. [ieee-fraud-detection, aptos2019-blindness-detection, severstal-steel-defect-detection, kuzushiji-recognition, two-sigma-financial-news, jigsaw-unintended-bias-in-toxicity-classification, freesound-audio-tagging-2019, recursion-cellular-image-classification, understanding_cloud_organization, data-science-for-good-city-of-los-angeles]

https://doi.org/10.30844/wi_2020_c6-roeder

Especially in the case of text and image data, state of the art models may be composed of hundreds of millions of parameters, imposing a high demand on the graphical processing unit. We explicitly do not consider this aspect in this taxonomy, as it is situated in downstream steps of the information value chain. Regarding D_3 *variety* we note that unstructured and semi-structured data are represented in a large share of the competitions. In four out of ten competitions, we were able to identify elements that were classified as intersecting and that went beyond the mere matching of unstructured data with the associated labels (D_5). In terms of *accuracy* (D_8), we can observe that the data range from reliable to error-prone. Overall, we consider all subjective ending conditions to be fulfilled, as shown in Table 1. However, the subsequent evaluation could call this finding into question again in the event of significant inconsistencies.

4.2 The Final Taxonomy

The final version of the taxonomy is shown in Figure 3. A detailed definition of all mentioned characteristics is in the Appendix. Overall, core aspects of the big data concept have found their way into this taxonomy. However, dimensions such as granularity or scale also play an important role in measuring heterogeneity.

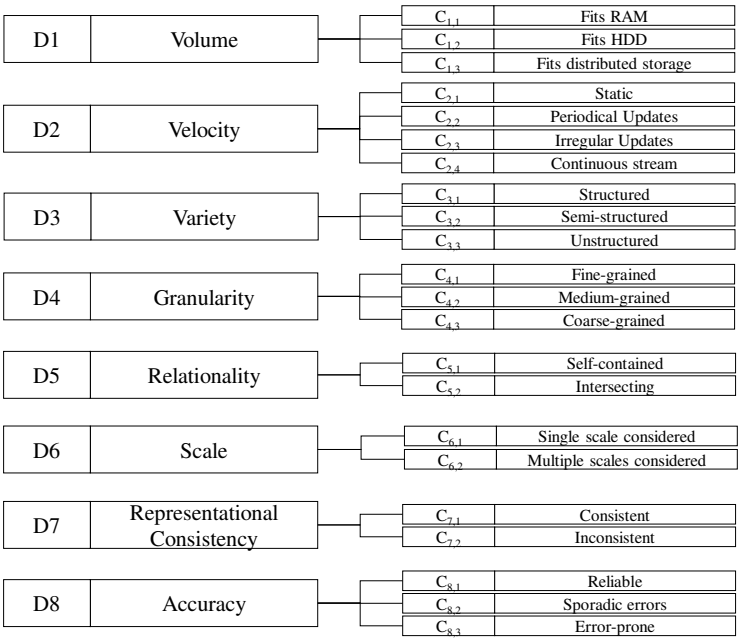

Figure 3. Final Taxonomy of Data Heterogeneity

https://doi.org/10.30844/wi_2020_c6-roeder

5 Evaluation

For evaluation purposes, we classify a selection of diverse data sets relevant to fields such as finance or politics. We choose to evaluate the data in the narrower sense (i.e., potential external data is not considered as strongly for example regarding relationality). Table 2 shows the results of the classification according to the proposed taxonomy. In the following, the dimensions for which classification was particularly interesting or difficult will be discussed.

Table 2: Data set classification based on the taxonomy

	D1			D2				D3			D4			D5		D6		D7		D8		
	Fits RAM	Fits HDD	Fits distr. storage	Static	Periodical updates	Irregular updates	Continuous stream	Structured	Semi-structured	Unstructured	Fine-grained	Medium-grained	Coarse-grained	Self-contained	Intersecting	Single scale	Multiple scales	Consistent	Inconsistent	Reliable	Sporadic errors	Error-prone
Raven-pack News Analytics [1]		●					●	●				●			●		●	●			●	
SEC EDGAR[2]			●		●				●			●		●		●		●		●		
Consumer Price Ind.[3]	●				●			●					●	●			●	●		●		
Geodata Hospitals[4]	●					●			●		●			●			●	●		●		
GAB hate speech corpus[5]		●					●		●		●				●	●		●			●	

[1]Ravenpack is a financial database that contains insights like sentiment or novelty for financial news.
[2]The primary database for U.S. companies to submit filings. [3]Published by the Federal Statistical Office of Germany.
[4]The data was obtained from geoportal.nrw. [5]GAB is a social media forum frequented by people characterized as "alt-right." [34]. The corpus was acquired from pushshift.io.

Since the full *Ravenpack* database has a volume of upwards of 100 GB, it does not fit the RAM of common PCs. *Granularity* can be viewed from two angles. With respect to time, the data are fine-grained, since they can be matched down to the millisecond. At the same time, as only the sentiment and novelty of the message is provided, the contents can no longer be traced granularly. Due to this goal conflict, we assign the characteristic medium-grained. The data is *intersecting*, as the company identifier and name of the news source can be joined with supplementary information. Since both a company and time level exist, we classify Ravenpack as having *multiple scales*. The

https://doi.org/10.30844/wi_2020_c6-roeder

relevant data has a consistent representation and may contain sporadic errors, like stale data points.

EDGAR is a database of company and individual filings operated by the United States Securities and Exchange Commission. Since the file size of all copies of just one specific type of filing already in 2012 was more than 200 GB, the characteristic *fits distributed storage* is assigned [35]. This has implications for the subsequent steps in the information value chain, both for retrieval which impacts the subsequent analysis. The filings can be retrieved as HTML, which are composed of different sections that contain natural language or tables. Hence, we consider the data to be semi-structured. Even though we consider the data to be *semi-structured*, the representation is rather consistent with a well-defined set of sections that are required for different filings.

The *consumer price index* data is provided by the Federal Statistical Office of Germany. The data is periodically updated to include new time periods. It is rather coarse grained as the values are provided on a monthly basis and aggregated to the federal level. We consider the provided data comparatively reliable. Naturally, the Statistical Office must aggregate different sub-measurements to create the provided data set, which may contain irregularities to some degree. At the same time, issues like missing values or false data values do not occur.

The retrieved *geo data* for locations of hospitals in Düsseldorf hospitals fits easily into RAM. It has a semi-structured format. The data is self-contained as no further or external data sets must be integrated to use the data. The reliability of the data is high.

The *GAB corpus* at hand measures about 60 GB but is limited to a time period of two years. However, modern HDDs provide the capacity to easily store data for an even larger time period. We consider the data to be *fine-grained*, as each comment is available accompanied by important meta information like the number of likes or the id of the parent comment. The *relationality* is high because each comment can be related to other comments, responses or users.

In summary, we note that the taxonomy developed conceptually and by analysis of Kaggle data sets can be applied to various data sets from different domains. The above explanation is also intended to demonstrate that the classification of new records is not always free of controversy. At the same time, we believe that the set of chosen dimensions can help to characterize differences in heterogeneity that can give important clues for the subsequent steps in the information value chain. Thus, we see the assessment that the subjective ending conditions are fulfilled as confirmed.

6 Discussion and Conclusion

In this paper, we built and presented a taxonomy of data heterogeneity, using both inductive and deductive steps for taxonomy building. The used data sources include the Kaggle platform and various other data providers. The value of the developed taxonomy is based on the idea that the identified dimensions delineate different aspects of data heterogeneity. Overall, we showed that the proposed dimensions and characteristics can help capturing different aspects of heterogeneity. Enabling researchers and practitioners to describe and comprehend different aspects of data

https://doi.org/10.30844/wi_2020_c6-roeder

heterogeneity is particularly relevant, as this fundamentally supports data understanding. Data sets that are similar with respect to their heterogeneity may require related techniques for cleaning, transforming, and analyzing the data. The mapping to processing and analysis steps further down the information value chain poses potential for subsequent research. One concrete example is statistical modeling situated in the information phase. Here it can be essential to consider existing data hierarchies in order to incorporate them in the model design. Industry practitioners, such as statisticians, data scientists or other professionals, can also benefit from a fundamental classification of data heterogeneity. In this way, it can be determined at an early stage how extensive steps such as data cleansing or pre-processing will be.

While the data of the analyzed Kaggle competitions is distinctly more heterogeneous than comparable data that is analyzed in typical scientific journals from business and economics, it is still subject to an inherent selection basis. On the one hand, in Kaggle Competitions data already has been cleaned, filtered and integrated in many cases. On the other hand, much of the data used in the business context today remains structured. In comparison, Kaggle data is much more heterogeneous. A final answer to this question is beyond the scope of this paper but offers potential for subsequent studies. Additionally, misclassifications can occur because it is not always possible to assign the characteristics with a high degree of separation. For example, a time variable that is provided granularly can be grouped using different levels (years, months, etc.), which can make it difficult to arrive at a sensible assessment.

Further analyses in areas where heterogenous data sources are prevalent provide ample opportunities for investigation. Future research could apply the developed taxonomy in a large-scale fashion to a larger number of data sets. New dimensions or characteristics could emerge or the (lack of) variation in a dimension could be used as an indicator to prune excess dimensions that are not essential. Thinking one step further, the analysis of the interdependencies of the elements along the entire information value chain is also promising and holds great potential for research and practice.

References

1. Humby, C., ANA Senior marketer's summit (2006)
2. Abbasi, A., Sarker, S., Chiang, R.H.L.: Big Data Research in Information Systems: Toward an Inclusive Research Agenda. Journal of the Assoc. for Inf. Sys. 17, i-xxxii (2016)
3. Chen, H.C., Chiang, R.H.L., Storey, V.C.: Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly 36, 1165-1188 (2012)
4. Wirth, R., Hipp, J.: CRISP-DM: Towards a standard process model for data mining. In: Proceedings of the 4th Int. Conf. on the Pract. Appl. of Knowledge Discovery and Data Mining, pp. 29-39. Citeseer, (2000)
5. Sharda, R., Delen, D., Turban, E., Aronson, J., Liang, T.P.: Business Intelligence and Analytics: Systems for Decision Support. Prentice Hall, Essex (2014)

https://doi.org/10.30844/wi_2020_c6-roeder

6. Merriam-Webster, https://www.merriam-webster.com/dictionary/heterogeneous (Accessed: 10.08.2019)
7. Eliazar, I., Sokolov, I.M.: Maximization of statistical heterogeneity: From Shannon's entropy to Gini's index. Physica A: Statistical Mechanics and its Applications 389, 3023-3038 (2010)
8. Lawson, T., Garrod, J.: Dictionary of Sociology. Fitzroy Dearborn, Chicago (2001)
9. Dominici, F., Parmigiani, G., Wolpert, R.L., Hasselblad, V.: Meta-Analysis of Migraine Headache Treatments: Combining Information from Heterogeneous Designs. Journal of the American Statistical Association 94, 16-28 (1999)
10. Wooldridge, J.M.: Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models. The Rev. of Econ. and Stat. 87, 385-390 (2005)
11. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal 10, 334-350 (2001)
12. Wu, X., Zhu, X., Wu, G., Ding, W.: Data mining with big data. IEEE Transactions on Knowledge and Data Engineering 26, 97-107 (2014)
13. Ranjan, J.: The 10 Vs of Big Data framework in the Context of 5 Industry Verticals. Productivity 59, 324-342 (2019)
14. Kitchin, R., McArdle, G.: What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data & Society 3, 1-10 (2016)
15. Volk, M., Hart, S., Bosse, S., Turowski, K.: How much is Big Data? A Classification Framework for IT Projects and Technologies. In: Proc. of the 22nd Americas Conf. on Inf. Sys. AIS, San Diego, CA (2016)
16. Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., Lee, D.: A Taxonomy of Dirty Data. Data Mining and Knowledge Discovery 7, 81-99 (2003)
17. Susha, I., Janssen, M., Verhulst, S.: Data Collaboratives as a New Frontier of Cross-Sector Partnerships in the Age of Open Data: Taxonomy Development. In: Proc. of the 50th Hawaii Int. Conf. on Sys. Sciences. AIS, Waikōloa Beach, Hawaii (2017)
18. Widjaja, T., Kaiser, J., Tepel, D., Buxmann, P.: Heterogeneity in IT landscapes and Monopoly Power of Firms: a Model to Quantify Heterogeneity. In: Proc. of the 33rd Int. Conf. on Inf. Sys. AIS, Orlando, FL (2012)
19. Lafky, D.B.: Heterogeneous Data in Federated Networks: A Framework for Solution Development. In: Proc. of the Americas Conf. on Inf. Sys. AIS, New York, NY (2004)
20. Nickerson, R.C., Varshney, U., Muntermann, J.: A method for taxonomy development and its application in inf. sys. European Journal of Information Systems 22, 336-359 (2013)
21. Glass, R.L., Vessey, I.: Contemporary application-domain taxonomies. IEEE Software 12, 63-76 (1995)
22. Bailey, K.D.: Typologies and taxonomies. Sage, Thousand Oaks, CA (1994)
23. Gregor, S.: The Nature of Theory in Inf. Sys. MIS Quarterly 30, 611-642 (2006)
24. Laney, D.: 3D Data Management: Controlling Data Volume, Velocity, and Variety. META Group (2001)
25. Lycett, M.: 'Datafication': making sense of (big) data in a complex world. European Journal of Information Systems 22, 381-386 (2013)
26. Goes, P.B.: Big Data and IS Research. MIS Quarterly 38, iii-viii (2014)
27. Elmasri, R., Navathe, S.: Fund. of Database Systems. Addison-Wesley, Boston, MA (2011)
28. Kitchin, R.: The opportunities, challenges and risks of big data for official statistics. Statistical Journal of the IAOS 31, 471-481 (2015)

https://doi.org/10.30844/wi_2020_c6-roeder

29. Kitchin, R.: The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. SAGE Publications Ltd, London (2014)
30. Monk, A., Prins, M., Rook, D.: Rethinking Alternative Data in Institutional Investment. The Journal of Financial Data Science Winter 2019, 14-31 (2019)
31. Gelman, A., Hill, J.: Data analysis using regression and multilevel / hierarchical models. Cambridge University Press, New York, NY, USA (2006)
32. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. J Manage Inform Syst 12, 5-33 (1996)
33. Kaggle, https://www.kaggle.com/ (Accessed: 14.08.2019)
34. Ellis, E.G., https://www.wired.com/2016/09/gab-alt-rights-twitter-ultimate-filter-bubble/ (Accessed: 20.11.2019)
35. García, D., Norli, Ø.: Crawling EDGAR. The Spanish Review of Financial Economics 10, 1-10 (2012)
36. Müller, O., Junglas, I., Brocke, J.v., Debortoli, S.: Utilizing big data analytics for information systems research: challenges, promises and guidelines. European Journal of Inf. Sys. 25, 289-302 (2016)

Appendix: Description of Characteristics

	D_1 Volume
Fits RAM	The data fits into memory of a business pc or workstation.
Fits HDD	The data does not fit memory, out-of-core algorithm required.
Fits distributed storage	The data is too large, distributed storage becomes necessary.
	D_2 Velocity
Static	The data set does not change over time [26].
Periodical updates	The data set is extended and updated at regular points in time [26].
Irregular updates	The data set is updated and extended at varying points in time.
Continuous stream	Quasi-flood of continuous data stream, often with low latency [36].
	D_3 Variety
Structured	The stored data is represented in a strict format (e.g. database structure) [27].
Semi-structured	Schema information is mixed with the data. Conformity to the defined format is less strict compared to structured data [27].
Unstructured	Barely any type definitions are included in the loosely structured data [27].
	D_4 Granularity
Fine-grained	Data is fine-grained compared to the phenomenon of interest [30].
Medium-grained	An intermediate scale between fine-grained and coarse-grained.
Coarse-grained	The data points are coarse-grained compared to the phenomenon of interest, i.e. each data point covers a lot of time, space, … [30].
	D_5 Relationality
Self-contained	Data contains no or few fields that can be related to complementary data sets [29].
Intersecting	Data contains common fields that can be matched with complementary data sets [29].

https://doi.org/10.30844/wi_2020_c6-roeder

D_6 Scale	
Single scale considered	The data is measured at the same scale [31].
Multiple scales considered	The data is measured at multiple scales. (Dis-)aggregation may be necessary to relate parts of the data to each other [31].
D_7 Representational Consistency	
Consistent	Data is represented in the same format and compatible with previous data from the same source [32].
Inconsistent	Data is represented in the different formats. Integrating new data with previous data is cumbersome and error prone [32].
D_8 Accuracy	
Reliable	The data is correct, reliable, and precise [32].
Sporadic errors	Data may contain some errors or missing values.
Error-prone	Frequently the data is not correct, reliable, and precise [32].

https://doi.org/10.30844/wi_2020_c6-roeder