Classifying Big Data Taxonomies: A Systematic Literature Review

Daniel Staegemann, Matthias Volk, Alexandra Grube, Johannes Hintsch, Sascha Bosse,
Robert Häusler, Abdulrahman Nahhas, Matthias Pohl and Klaus Turowski
Magdeburg Research and Competence Cluster Very Large Business Applications, Faculty of Computer Science,
Otto-von-Guericke University Magdeburg, Magdeburg, Germany
{daniel.staegemann, matthias.volk, alexandra.grube, johannes.hintsch, sascha.bosse, robert.haeusler,

Keywords: Big Data, Taxonomy, Literature Review, Classification, Categorisation, Systematization, Data Characteristics, Structured, Analysis.

Abstract: As big data is a rather young, but growing discipline, lots of confusion about the general nature of this term exists. Consequently, multiple research endeavours to discover unique characteristics, technologies, techniques and their interconnections were conducted, resulting in comprehensive classification approaches. For this purpose, various taxonomies on big data exist in literature. However, due to the multitude of approaches and partial contradictions, no real clarification is achieved. To overcome this issue, a systematic literature review was conducted, which identifies and analyses big data taxonomies. As a result, a classification of those taxonomies is proposed, which additionally tracks sub-domains that are not yet covered by the existing taxonomies so far. Eventually, the publication at hand serves as a starting point for further taxonomy related research endeavours in the big data domain.

1 INTRODUCTION

The term big data, as known today, was coined approximately 20 years ago (Diebold 2012). Whether it is in astronomy where telescopes produce hundreds of gigabyte of data every night (Kremer et al. 2017), in healthcare with organizations trying to find patterns in their data to improve services (Wang et al. 2018), in disaster warning were social media data is exploited (Wu and Cui 2018), or in the endeavour to improve urban transportation management (Fiore et al. 2019), big data turned into an omnipresent part of today's society. Furthermore, it is a key technology in industry 4.0 (Dobos et al. 2018; Wang et al. 2016) and the economic value of its application is scientifically substantiated (Müller et al. 2018; Brynjolfsson et al. 2011; Bughin 2016). Though there are many definitions of big data available, it is often characterized by *volume*, referring to the massive size of data to handle, *velocity*, which stands for the speed data is being produced or transmitted, *variety*, representing the spectrum of data formats and sources and *variability*, indicating the ongoing changes that occur in the data. Overall big data describes ways to acquire, store, process and analyse large-scale data for which traditional techniques are not suitable and therefore new system landscapes emerged (NIST 2019). Over time, many existing technologies, concepts and application areas have been scientifically analysed, and numerous new approaches, techniques, systems and research questions have been developed and described. As a result, the body of literature for big data analytics constantly increases and widens its range (Chen and Zhang 2014; Staegemann et al. 2019b). However, there are still many scientific gaps and one of them is a lack of clarity, necessitating means of supporting understanding and application of big data to facilitate its application (Volk et al. 2019). To bridge this gap, several taxonomies and classifications have been developed and presented in certain big data related topics and application areas, such as (Miller 2013; Hartmann et al. 2016; Kumari et al. 2018).

A taxonomy is a classification, which divides its subject into different categories, classes or families (Nickerson et al. 2013). If applicable, further hierarchical ramifications are listed. In that way, many different fields are systematized, analysed and overviewed. In the publication at hand, the abovementioned big data taxonomies are the main focus. Even though, those structures organize their respective areas of interest, they themselves are not

Staegemann, D., Volk, M., Grube, A., Hintsch, J., Bosse, S., Häusler, R., Nahhas, A., Pohl, M. and Turowski, K.
Classifying Big Data Taxonomies: A Systematic Literature Review.
DOI: 10.5220/0009390102670278
In *Proceedings of the 5th International Conference on Internet of Things, Big Data and Security (IoTBDS 2020)*, pages 267-278
ISBN: 978-989-758-426-8

classified with regards to each other. This however, would allow for a straightforward overview of the domain and therefore increase accessibility. Furthermore, potential contradictions among the numerous taxonomies could be revealed, setting the foundation for future work to increase the corresponding consistency. Therefore, the following research question is addressed in the course of this work:

RQ: How can the existing big data taxonomies be classified in relation to each other and which sub-domains constitute a research gap in that regard?

To find a suitable answer for this RQ, first, it is necessary to obtain an overview about existing taxonomies. After that, an investigation about their similarities and differences needs to be carried out. In doing so, open research areas, for which no relevant taxonomies have been found, may be discovered. Hence, to answer the RQ, the following sub-research questions (SRQ) will be answered:

SRQ1: Which big data taxonomies do currently exist in the literature and which subject do they cover?

SRQ2: How can the identified taxonomies be classified?

SRQ3: In which way are taxonomies from the same category complementary and state the same conclusion?

SRQ4: Which sub-domains are not covered by a taxonomy, but could benefit from the provided clarity?

While SRQ1 has the purpose to identify the according literature as a necessary foundation and SRQ2 brings them into a joint systematic, SRQ3 aims at investigating, to which extend there exists an academic consensus regarding a topic. Finally, SRQ4 provides subsequent scientists with potential avenues for further research, facilitating the advancement of the domain.

To approach those questions, a systematic literature review has been conducted (Webster and Watson 2002; Levy and Ellis 2006). For this purpose, a two-stepped refinement process was implemented. While the first step is the collection of publications whose titles comprise relevant keywords, in the second step a qualitative analysis was carried out. To ensure transparency, its process (Vom Brocke et al. 2009) as well as the results are thoroughly documented in the following sections. While this section constitutes the introduction and functions as a motivation, the second section describes the search for literature according to the topic as well as the review and selection approach in detail. This has not only been done to understand and comprehend the results of this paper, but also to indicate the limits of this literature search and define a clear border to research articles, not considered in this work. The third section presents and analyses the publications obtained through the literature review. It sets existing big data taxonomies and classifications in relation to each other, in which way a bigger picture of the existing literature arises and facilitates the opportunity to outline the topic's current state of research. Furthermore, each taxonomy will be described and the characteristics are highlighted. Additionally, the taxonomies are reviewed and eventual dissented views or gaps in the literature are considered. The fourth section highlights the overall findings and points out identified opportunities for further research. In the end, a conclusion is given and a possible way of expanding on the publication at hand addressed.

2 LITERATURE REVIEW

To provide a transparent and expedient overview of the current existing taxonomies in big data, this structured literature review has been written in the concept-focused style suggested by Jane Webster and Richard T. Watson (2002).

2.1 Review Protocol

For locating and accessing the relevant literature, an internet-based search has been conducted, using different search engines and databases. Here, dblp and IEEE Xplore were utilized to provide search results from specifically computer science related databases. In addition, to cover various resources and different databases, the widely used scientific search engine Google Scholar has been used as an all-around search solution. To round off the search range, the database searches from the scientific publishers Springer (access via SpringerLink) and Elsevier (access via ScienceDirect) were used as well. Utilizing five different search engines from three different contexts facilitates a representative view of currently existing and relevant literature and therefore ensures the quality of the search results.

To obtain the most relevant results, this literature review is based on two search terms, consisting of three keywords. The first search term is "big data taxonomy" and the second is "big data taxonomies". Big data defines the field that has been targeted, while taxonomy defines the actual objective that has been aimed to be investigated. To limit the initial amount

of found out papers and increase their relevancy, the keywords had to appear in the title of the literature item. This way a priority has been set to academic publications, whose central research questions deal with the development and presentation of taxonomies related to big data. However, no other search configuration such as time range, citations or similar has been amended or set. Furthermore, all items listed, including citation entries (appearing in Google Scholar) have been selected to be reviewed. Using this approach, an initial literature corpus has been obtained. Table 1 shows the results of the search attributed to each search engine.

Table 1: Search engine results overview.

Search Term	Search Engine	Results	Results without Duplicates
Big Data Taxonomy	Google Scholar	51	42
	dblp	16	16
	IEEE Xplore	7	7
	SpringerLink	2	2
	ScienceDirect	4	4
Big Data Taxonomies	Google Scholar	5	4
	dblp	1	1
	IEEE Xplore	0	0
	SpringerLink	0	0
	ScienceDirect	0	0
Total Unique Items		47	

While Google Scholar, due to its nature as a meta database, obtained the most results, it also contained nine duplicates. Therefore, in a first processing step, those duplicates had to be removed. Besides that, some of the publication were found in more than one search engine, necessitating further cleansing. Finally, considering those adjustments, 47 unique papers have been found as the output of the first step of the literature review.

2.2 Qualitative Analysis

Since it is highly likely, that not all of those 47 obtained literature items are relevant in the context of the publication at hand, they had to be further filtered by the means of a qualitative analysis. For this purpose, in the second step of the refinement process, each one was completely read and subsequently evaluated based on the inclusion and exclusion criteria depicted in Table 2. This means, for a publication to be included in the literature corpus, all the inclusion criteria had to be met, while at the same time, not a single one of the exclusion criteria applied.

Table 2: Inclusion and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Literature is published in a journal, conference, book, a white paper or report (latest edition only).	Publications refer or cite taxonomies, that have not been created for big data research purposes.
Literature provides at least one relevant big data specific taxonomy.	Literature that presents a taxonomy of which big data is part, but not the main research field.
If the publication is of an interdisciplinary nature, the discussed taxonomy provides relevant perspectives on big data.	Taxonomy/classification is part of the data-set that is examined in the publication itself or in a data mining context (for example attribute value taxonomy).
	Publication is written in a language other than English or German.
	Literature items, that name their research work a taxonomy to describe the synoptical and holistic nature of the work, but do not provide a specific taxonomy.

As a result of the filtering based on those criteria, a set of relevant literature has been determined, which comprises 28 items. Their breakdown based on the type of literature is shown in Table 3. The table also indicates the number of publications that were removed in the previously mentioned step.

Table 3: Breakdown of the obtained literature items.

Type	After Step 1	After Step 2
Journal Paper	24	19
Conference Paper	9	5
Book	5	1
Report/White Paper	5	3
Thesis	2	0
Unknown	2	0
Total included		28

Those 28 publications, as the findings of the described structured literature review, constitute the most appropriate and relevant works regarding the research questions. Therefore, they are also the foundation for the following considerations.

3 CLASSIFICATION OF THE BIG DATA TAXONOMIES

After obtaining the relevant literature, a thorough analysis is conducted. On the one hand this provides an overview of the existing taxonomies on an individual level and on the other hand, it allows for them to be transferred into a common structure, showing existing research gaps.

3.1 General View

The qualitative analysis of the 28 paper revealed two main classes, each reviewed taxonomy can be assigned to. Those are "Technological" and "Characteristics and Requirements", which describe the main nature of paper the taxonomy is embedded in. The classification in its entirety is depicted in Figure 1.

In this illustration, a classification tree is shown, which assigns each taxonomy provided in the literature to a research area of big data. The classifications are not based on a complete division of big data disciplines, however they show the thematic reference of the taxonomies reviewed.

The category *Technological* classifies taxonomies that either relate directly or have been created to lead to existing or emerging software, software architectures, tools, systems or algorithms. In most cases they are named in the classification itself or are subject of discussion, for which the taxonomy was used. Hence, it takes a clear technical point of view. *Characteristics and Requirements* however are taxonomies that are of a descriptive or distinguishing nature and focus on a management or character side of big data and its fields and applications. From the second level of the technological side, we have the "Synopsis" subclass, which can be further divided into "Platforms" and "Multidisciplinary". While for the naming of the latter category, contentwise, *Interdisciplinary* would have been a slightly better fit, this term was not used to avoid confusion with another category of that name, which is introduced later on. *Synopsis* refers to taxonomies that function as an overall overview or are intended to classify big data characteristics. If the taxonomy classifies big data specific platforms or architectures, it is categorised to *Platforms*. *Multidisciplinary* includes overview taxonomies that are related to big data either in a specific area of application or for a specific set of data. Another sub-class below technological is "Data". *Data* as the key subject for big data is further divided into "Data Acquisition", "Data storage" and "Data Analytics". This classification is rather rough and shall give a broad direction as the definition of those disciplines in literature is slightly divergent. In the course of this taxonomy, *Data Acquisition* is regarded as the actual process of collecting data and also comprises the different sources, data can be extracted from, while *Data Storage* refers to the actual process of storing data and the related software,

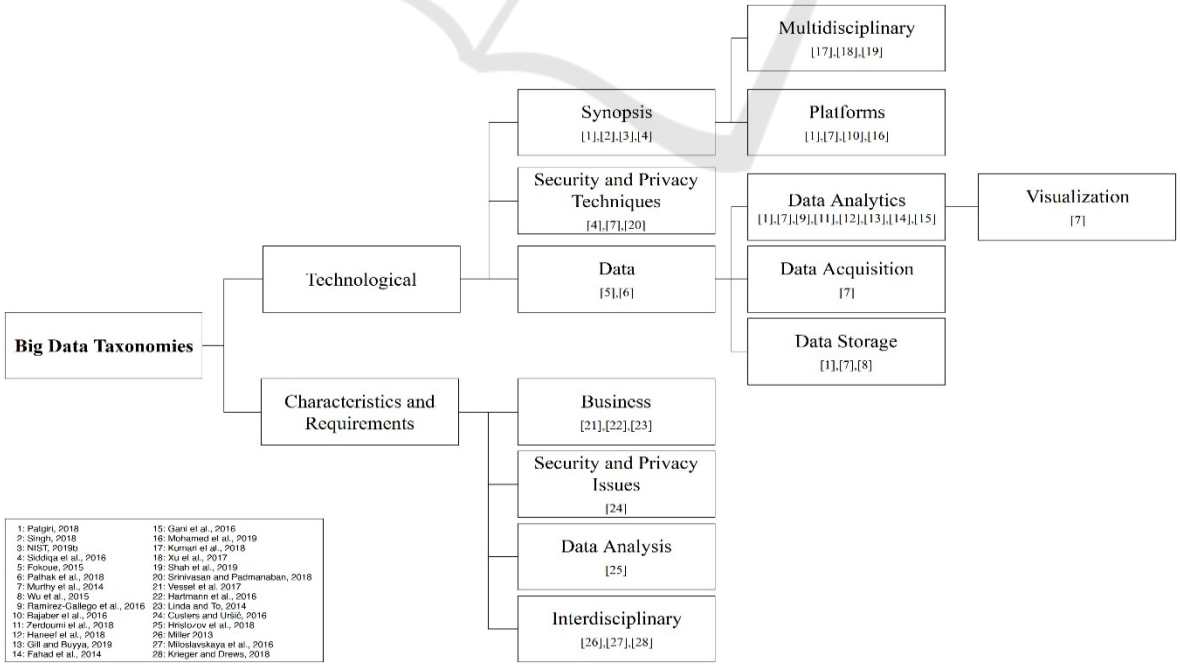

Figure 1: Categorisation of the taxonomies.

hardware or structures for this task. *Data Analytics* includes all kind of algorithms, techniques, software, methods and concepts to pre-process or process the data and gain information. This also includes "Visualization" techniques, however for a better overview, methods and ways to visualise data were defined as another sub-class. "Security and Privacy Techniques" as a sub-class of *Technological* comprises straightforward techniques and methods to secure the data and system landscape or increase privacy. While in "Security and Privacy Aspects" there is a similarly named sub-class below *Characteristics and Requirements*, that one is from a less technical position and more towards an analytical or descriptive point of view. Another sub-class for *Characteristics and Requirements* is "Business". This category highlights the economical side of big data, which is due to its potential, of very high interest for the public and private sector. The taxonomy further comprises the "Interdisciplinary" and "Data Analysis" sub-classes that again refer to big data used for specific applications or fields and for analysing large-scale data respectively.

3.2 Detailed View

Big data is a rather young and broad discipline with yet no universal consensus on definitions and dividing lines (Timmins et al. 2018). This is also recognisable in the taxonomies created in this field. They are very varied in size and details, and even though, sometimes taxonomies deal with the same subject, they are still not easily comparable and reflect a different approach on the topic. However, two main categories were identified, namely *Technological* as well as *Characteristics and Requirements*. Additionally, for each of those classes further sub-categories were identified at which the found out taxonomies are aligned. In the following subsections, each taxonomy will be outlined by naming the source and giving a brief description, to work out its context and characteristics. Along with the summaries, an evaluation of the existing concepts and the coverage of the subfield is given. Because some of the analysed publications are related to several sub-categories, they will also appear repeatedly in the following sections. However, the descriptions will differ in those cases, since, depending on the context, other parts of the publication are regarded.

3.2.1 Technological – Synopsis

In (Patgiri 2018), the authors present a classification of big data into six different fields, which serves as a synoptically overview that is further explained and elaborated in sub-taxonomies. One of those is a semantic depiction of big data, which focuses on the "V"s big data is often being described with and broadens them with further characteristics in the same manner.

Another taxonomy that provides a holistic overview about definition of big data, putting volume, velocity, and variety in a relation, is proposed by (Singh 2018).

In the second volume of the big data interoperability framework of the National Institute of Standards and Technology (NIST), a reference architecture taxonomy is presented, that indicates technologies, workflows and key roles of big data and puts those in relation to each other. The report is addressed to managers, procurement officers, marketers, technical community and rather records consensus on big data techniques and concepts than focusing on a specific research question.

The paper entitled "A survey of big data management: Taxonomy and state-of-the-art" (Siddiqa et al. 2016) introduces a synoptic taxonomy. This classifies big data into data storage, pre-processing and processing and formulates for each category three problems as well as a recommendation of a technique or algorithm that can be applied. Furthermore, the taxonomy matches six main big data challenges to each of the solutions provided.

As one may note, two of the taxonomies in this overview level are developed to define big data by focusing on the characteristic "V"s, hence they are of a complementary nature. The others position very diverse in terms of the range and detail. However, the division of big data itself into sub-categories is not consistent and not transferable. It implies that there is no agreement on a lucid big data overview, which clusters big data into different sub-disciplines or rather fields.

3.2.2 Technological – Data

Within big data literature, data are often being described using their technical properties, such as structure, format or attributes. Those information are not covered within the taxonomies found, however they partially refer to some of those aspects. A taxonomy that deals with the data characteristics, which are specific or relevant for big data has not been detected in this search. According to the taxonomies found, there is a measurement for data size that leads to a classification. This approach was introduced in (Fokoue 2015). In here, the concept of a taxonomy uses a rather unique classification based

on the ratio between the sample size of the data to be dealt with and the size of the dimensions. Within the described ratio and the sample size, six categories emerge and are being discussed.

Within the contribution "Construing the big data based on taxonomy, analytics and approaches" (Pathak et al. 2018), a cross-functional view is presented. By harnessing the idea of a taxonomy, an approach dealing with different big data topics like storage, analytics, state of the art and future trends in a holistic way is proposed. It classifies data "based on method of data collection, accessibility pattern, source of data generation, and statistical approach" (Pathak et al. 2018).

Another area, which was not ascertained and might profit of a corresponding taxonomy is data quality. Either referring to attributes that describe data quality from a big data point of view, or possible techniques, that respect or even increase data quality.

3.2.3 Technological – Data Acquisition

With the paper "Big Data Taxonomy" (Murthy et al. 2014) only one taxonomy was found that focuses on the acquisition of the data itself. Along with a classification of data structure, ahead of the taxonomy, it provides a division of different industry domains and subdomains in which (big) data is being generated.

The occurrence of only one particular contribution within this subcategory may have different reasons. Apart from the sole lack of research, in terms of systematizations and categorisations, also the general complexity of this particular domain could be causative. In any case, potential gaps that might be useful or possible to fill are a clear overview of not only data sources, but also the way they are collected and/or transferred. Further, the data integration in terms of merging large-scale data sets/databases together or integrating them into an existing system landscape could be useful to investigate further.

3.2.4 Technological – Data Storage

Further contributions have been found, which are particularly focusing on the storing and management of the data in different variations. Although there are different approaches or variations in the main emphasis, the academic assertion is very similar and supplementary. In "Survey of Large-Scale Data Management Systems for Big Data Applications" (Wu et al. 2015) a three stepped procedure is presented at which the management software used for big data applications is explored. For each step a separate taxonomy is developed. Firstly, the focus has been set to the data point of view, describing physical and conceptual levels. Secondly, a taxonomy of system architectures is provided and different approaches, as well as the related software solutions, classified. In step three, the consistency model is the point of attention. This last step deals with the challenge of scaling down data management systems without losing consistency. Again, different approaches and software solutions are classified.

An entirely different approach that deals with the existing storage possibilities was introduced by (Patgiri 2018). His taxonomy comprises four categories for different existing storage possibilities. It focuses on architecture, structure, implementation and usable devices. Furthermore, NoSQL, which can be used for the data storage, is regarded. To classify those techniques and software, four paradigm classes are used.

A specific taxonomy that classifies different database software products by the way data is being stored is discussed in (Murthy et al. 2014). Additionally to that, an overview table is presented, providing detailed information regarding the characteristics for each concept.

3.2.5 Technological – Data Analytics

There are various ways to prepare, process and analyse data, yet there is no taxonomy found, that synoptically creates an overview of existing methods and sets those in relation to each other. Currently, the existing classifications rather set the focus on specific techniques or algorithms. In the already referred work of (Patgiri 2018), one of the taxonomies provides categories and sub-categories for the intention or approach data is being analysed for. Another presented taxonomy classifies machine-learning algorithms into eleven different kinds.

Murthy et al. (2014) provides a brief overview of the most used machine learning algorithms that are used to analyse data. The algorithms are classified to *Supervised*, *Unsupervised* and *Semisupervised* as well as *Re-enforcement*.

In "Data discretization: taxonomy and big data challenge" (Ramírez-Gallego et al. 2016) a taxonomy, resulting out of the findings of a literature review, is described. This approach classifies the most important discretization methods into two main classes with various different subclasses.

The paper of (Zerdoumi et al. 2018) presents a classification of graph processing platforms, that are mostly specialized in large-scale data. It can be used where general-purpose systems might have performance issue. Furthermore, they also proposed a pattern recognition taxonomy specified for big data

use. Within the four suggested categories and their sub-categories, the different approaches of pattern recognition have been technically explained and comprehensively analysed.

By examining the current technologies of information retrieval, a comparative taxonomy is created in (Haneef et al. 2018). This classifies methods or systems according to types and parameters.

"Bio-Inspired Algorithms for Big Data Analytics: A Survey, Taxonomy, and Open Challenges" (Gill and Buyya 2019) — This taxonomy categorises the algorithms in swarm-based, ecological and evolutionary. Furthermore, it provides references for each algorithm with date and an example of application.

In the paper entitled "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis" (Fahad et al. 2014), the authors present five different clustering approaches and classify 24 algorithms accordingly.

Gani et al. (2016) classify and sub-classify current indexing techniques and algorithms based on their strategy. Moreover, big data characteristics ("V"s) have been applied to the taxonomy and mark especially big data related techniques.

Due to the fact, that most of the aforementioned taxonomies, related the data analytics, are very different, a comparison is very limited. However, the missing synoptical taxonomy would be beneficial to holistically sum up the existing information and to investigate, if relations among them can be found and also to find superordinate categories. Nonetheless, there are two taxonomies about machine learning. They both refer to existing algorithms. However, they differ in the number of algorithms sorted as well in categories named. This might be due to the time gap between the publication dates or due to the nature of each publication. Though there are already a number of taxonomies in this field, there might be some concepts missing in this list. For example data cleansing to detect corruptions or false data when acting on a large scale. Although the taxonomies listed above contain techniques and algorithms that are being used for prescriptive or predictive analyses, it might be beneficial to develop a categorisation that focuses this use of big data. Additionally, mathematical approaches used to analyse big data are to our knowledge, not yet classified.

3.2.6 Technological – Visualization

Visualization of data is not a very recent topic, however with the perspective of big data, there is only one classification mentioned. The already referenced work by (Murthy et al. 2014) also classifies the most common visualization software and algorithms based on the way they process data.

Hence, it might be worth developing other approaches to categorise visualization techniques that highlight current development or focus on other aspects like suitability for certain datasets or with regards to current big data architecture concepts. In addition, a very specific but arising topic in the domain are self-service tools to visualize the data exploration.

3.2.7 Technological – Platforms

Concrete taxonomies related to the category of *Platforms* were found in four publications. In (Patgiri 2018), different technologies are broken down into three classes and subsequently described.

In the taxonomy of Murthy et al. (2014), big data architectures are divided in the two main categories, *Batch* and *Streaming*, referring to the approaches the systems are based on.

The taxonomy presented in "Big Data 2.0 Processing Systems: Taxonomy and Open Challenges" (Bajaber et al. 2016) presents the state of the art for big data platforms in order to understand the recent developments. It points out four main groups, namely *General Purpose Systems*, *Big SQL Systems*, *Big Graph Processing Systems* and *Big Stream Processing Systems*.

In (Mohamed et al. 2019) a holistic taxonomy is presented. It originates out of the findings of a literature review and shows four stages of big data, starting from source and format of data to data processing, analytics and visualization. For each stage, it categorises techniques and software and creates sub-categories if needed.

The two compute infrastructure taxonomies are very alike and only differ in details, while the other two approaches are disparate. In comparison, the last taxonomies indicate that there is no clear line between big data platforms and platforms or techniques used for other purposes or applications, as the categorisations are very different in platforms and software listed.

3.2.8 Technological – Multidisciplinary

Some of the found out taxonomies are dealing with multiple disciplines. Kumari et al. (2018) present an interdisciplinary taxonomy that focuses on multimedia big data, which is being processed to be used for IoT applications. The taxonomy is split in big data architecture layers and technical functions and broken down into further detailed sub-taxonomies.

Along the taxonomy, further literature is referenced and technical aspects are explained and compared.

Another multidisciplinary approach was introduced in (Xu et al. 2017). This taxonomy research refers to fault diagnosis in industrial big data such as IoT or cloud computing, and combines it with fault diagnostic methods that were commonly applied in the time before big data became relevant. In this way, the taxonomy is classified in three categories with further sub-categories. Traditional methods are separated from the ones based on industrial big data.

Shah et al. (2019) dealt as well with the big data analytics and IoT. In particular, they defined a taxonomy for disaster management processes, comprising seven classes and their properties respectively the utilized techniques.

The existing taxonomies are very detailed and provide a specific point of view for their area of application. However, some important areas might be missing. For example, a taxonomy that refers to business intelligence and the integration of big data in the already existing architectures. Also big data used for recommender systems or the development of hardware particularly used for big data purposes might be possible task for future research on taxonomies or classifications.

3.2.9 Technological – Security and Privacy Techniques

Apart from general concepts which are related to the used analysis methods, tools, technologies and application areas, also security and privacy techniques were highlighted in some of the found out taxonomies. In the paper entitled "State-of-the-art Big Data Security Taxonomies" (Srinivasan and Padmanaban 2018), a taxonomy is introduced that focuses on Hadoop and Hadoop related systems. Apart from listing the security essentials and levels, it describes three categories in which the security challenges and vulnerabilities have been sorted in.

Siddiqa et al. (2016) propose four different categories for security in big data. For each of them, two problems or issues are formulated as well as possible solutions or approaches. Along with the taxonomy, further details for this classification have been given.

Within the very comprehensive work presented by (Murthy et al. 2014) again relevant information have been found, for this particular area. In here, most common challenges regarding the technical aspects of security and privacy within the big data architectures and management are provided, grouped in main categories and briefly described.

All three existing approaches regarding the *Security and Privacy Techniques* show different ways of dealing with this topic. While the taxonomy about security challenges focuses on Hadoop respectively Hadoop related systems only, the other two constitute an overall approach, but with a very different outcome. However, the last two mentioned taxonomies are not developed within a publication that focuses on security or privacy. Therefore, further research could be instructive.

3.2.10 Characteristics and Requirements

All taxonomies listed in *Characteristics and Requirements* focus on different views or applications of big data and are hardly comparable. They reveal specific information and provide conclusive details. Sometimes, specific information are not only given for a certain area or purpose, but also for a multidisciplinary context.

Regarding the *Business* context, three contributions have been found, that present promising taxonomies. The paper "IDC's Worldwide Big Data and Analytics Software Taxonomy" (Vesset et al. 2017) shows the big data and analytics software market by creating three segments from the conceptual architecture perspective. Although it describes big data itself as a subset throughout all three market segments, the taxonomy is more comprehensive by including big data analytics software and traditional software for data management, analysis and visualization.

In (Hartmann et al. 2016), a taxonomy is proposed that describes how companies gain value or monetize data by focusing on six dimensions, which are subsequently split into sub-clusters. Furthermore, the taxonomy has been practically used by statistically analysing a sample set of 100 start-up firms and applying clusters on this sample according to the taxonomy.

In the approach presented by (Linda and To 2014), big data characteristics from an organizational point of view have been examined. It describes the data sources and data structures that are important for big data management.

Security and privacy Aspects were of major interest in the contribution of (Custers and Uršič 2016). In a continuous text, which focuses on the taxonomy developed for privacy reasons in big data, a classification, taking in the view of a data controller and data subject, has been worked out. It formulates and defines different types of data reuse. *Characteristics and Requirements* regarding the *Data Analysis* were described in the paper "Analytical

Competences in Big Data Era: Taxonomy" (Hristozov et al. 2018). This taxonomy describes competences required for data analysis in the big data context. It formulates requirements and characteristics of skills that are essential for successfully analysing big data. Some of the found out contributions were also *Interdisciplinary*. Miller (2013) deals in his taxonomy with the risks that are associated with the utilization of big data in cloud computing.

Miloslavskaya et al. (2016) classify characteristics of information system threats, vulnerabilities, incidents as well as attacks against security operation centres in a big data context. The resulting taxonomy also defines relevant parameter and gives further descriptions.

In (Krieger and Drews 2018), a taxonomy is presented that focuses on big data with auditing purposes such as accounting and fraud detection. It defines dimensions throughout auditing, data management and analytics and presents their characteristics.

There are several other opportunities for research that apply to this category. To name a few, business intelligence and business decision making could be potential topics. In addition, an overall classification of application fields could advance research in big data.

4 FINDINGS

With regards to the research question raised at the beginning of this paper, the taxonomies resulting from this literature search have been thematically and hierarchically clustered, which provides a straight forward outline of the current body of literature. It already reveals the nature of each taxonomy as well as areas with less devised taxonomies. Within each subject area, the presentation of each taxonomy shows its main topic as well as further observed characteristics. It provides a more detailed overview and has led to the previously described observations. It is not always possible or expedient to compare taxonomies that are sorted in the same area. However, in some cases it reveals a lack of consensus while others seem rather confirmatory. A significant discrepancy seems to exist when dealing with the actual big data techniques and platforms, which indicates, that there is no distinct border among them or for technologies used in big data and other disciplines. The same observation is made for dividing big data into sub-disciplines or sub-categories on a basic level, or when dealing with security and privacy terms. On the contrary, data storage taxonomies are highly consistent and big data taxonomies dealing with the typical "V" characteristics are complementary. In terms of taxonomy research carried out, there are areas of big data with a higher coverage than others. When looking at data visualization, data acquisition or the characteristics and requirements of the data analysis, those topics seem to be less extensively analysed than other areas like for example, data storage, data analytics and big data platforms. Nevertheless, nearly in all areas, potential gaps for creating additional taxonomies have been detected. Additionally, besides the already mentioned specific topics, there was not a single taxonomy found in the course of the conducted literature research that deals with ways or methods of testing big data solutions or systems. Equally, the numerous potential causes for failures or quality reduction in big data analysis (Staegemann et al. 2019b) were not part of the obtained considerations. This emphasizes once more, that, while big data in general is popular, the quality assurance is neglected (Staegemann et al. 2019a). Furthermore, regarding *Characteristics and Requirements*, there are some other topics that might be beneficial. One example is the training for human resources working with big data or analysing big data platforms, taking into account current legal conditions or recent changes. Another one concerns the different possibilities for the visualization of the results of an analysis. While (Murthy et al. 2014) proposed an according taxonomy, it focusses on graphical depictions. Hence, other possible options, such as texts, tables or audio-based representations, are not regarded. Relevant for both categories, *Technological* as well as *Characteristics and Requirements* might be to investigate the current state of running big data solutions cost-effective and with fewer resources. This topic is important from both, an economical, but also an ecological point of view and comprises approaches like the optimisation of algorithms, the choice of used hardware or considerations regarding server consolidation and virtual machines placement, with the latter already being covered by (Nahhas et al. 2019). However, since the consolidation is not directly big data related, it is not part of the literature review's obtained results, despite being somewhat relevant in the grand scheme.

5 CONCLUSIONS

In the presented research, a structured literature review was conducted and thereupon used to find and

evaluate existing big data taxonomies, answering SRQ1. Following the proposition of (Vom Brocke et al. 2009), the search and review process is thoroughly described to enable subsequent scientists to retrace the results and build their own research upon them. Each taxonomy has been described and, according to SRQ2, categorised. Subsequently, the taxonomies have been compared with the other ones from the same category, providing the answer to SRQ3. Furthermore, corresponding to SRQ4, potential research gaps have been identified. While this list of determined research gaps does not claim to be exhaustive, it constitutes a starting point for readers with expertise or experience in big data to identify promising research topics. The formal creation of a meta taxonomy for big data however, might be the next step, expanding on the present work and providing even more clarity. In this course, it could also be beneficial to expand the scope of the regarded literature, allowing to incorporate also taxonomies that are relevant, but not directly aimed at big data.

REFERENCES

Bajaber, Fuad; Elshawi, Radwa; Batarfi, Omar; Altalhi, Abdulrahman; Barnawi, Ahmed; Sakr, Sherif (2016): Big Data 2.0 Processing Systems: Taxonomy and Open Challenges. In Journal of Grid Computing 14 (3), pp. 379–405.

Brynjolfsson, Erik; Hitt, Lorin M.; Kim, Heekyung Hellen (2011): Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance? In SSRN Electronic Journal.

Bughin, Jacques (2016): Big data, Big bang? In Journal of Big Data 3 (1), pp. 1–14.

Chen, C. L. Philip; Zhang, Chun-Yang (2014): Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. In Information Sciences 275, pp. 314–347.

Custers, Bart; Uršič, Helena (2016): Big data and data reuse: a taxonomy of data reuse for balancing big data benefits and personal data protection. In International Data Privacy Law 6 (1), 4-15.

Diebold, Francis X. (2012): On the Origin(s) and Development of the Term 'Big Data'. In SSRN Electronic Journal.

Dobos, Peter.; Tamás, Péter.; Illés, Béla.; Balogh, R. (2018): Application possibilities of the Big Data concept in Industry 4.0. In IOP Conference Series: Materials Science and Engineering 448, p. 12011.

Fahad, Adil; Alshatri, Najlaa; Tari, Zahir; Alamri, Abdullah; Khalil, Ibrahim; Zomaya, Albert Y. et al. (2014): A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. In IEEE Transactions on Emerging Topics in Computing 2 (3), pp. 267–279.

Fiore, Sandro; Elia, Donatello; Pires, Carlos Eduardo; Mestre, Demetrio Gomes; Cappiello, Cinzia; Vitali, Monica et al. (2019): An Integrated Big and Fast Data Analytics Platform for Smart Urban Transportation Management. In IEEE Access 7, pp. 117652–117677.

Fokoue, Ernest (2015): A Taxonomy of Big Data for Optimal Predictive Machine Learning and Data Mining.

Gani, Abdullah; Siddiqa, Aisha; Shamshirband, Shahaboddin; Hanum, Fariza (2016): A survey on indexing techniques for big data: taxonomy and performance evaluation. In Knowledge and Information Systems 46 (2), pp. 241–284.

Gill, Sukhpal Singh; Buyya, Rajkumar (2019): Bio-Inspired Algorithms for Big Data Analytics: A Survey, Taxonomy, and Open Challenges. In Big Data Analytics for Intelligent Healthcare Management: Elsevier, pp. 1–17.

Haneef, Israr; Munir, Ehsan Ullah; Qaiser, Ghazia; Umar, Hafiz Gulfam Ahmad (2018): Big Data Retrieval: Taxonomy, Techniques and Feature Analysis. In International Journal of Computer Science and Network Security 18 (11), pp. 55–59.

Hartmann, Philipp Max; Zaki, Mohamed; Feldmann, Niels; Neely, Andy (2016): Capturing value from big data – a taxonomy of data-driven business models used by start-up firms. In International Journal of Operations & Production Management 36 (10), pp. 1382–1406.

Hristozov, Dimitar; Toleva-Stoimenova, Stefka; Rasheva-Yordanova, Katia (2018): Analytical Competences in Big Data Era: Taxonomy. In Proceedings of the ICERI2018 Conference. 11th Annual International Conference of Education, Research and Innovation. Seville, Spain, 12.11.2018 - 14.11.2018, pp. 7182–7191.

Kremer, Jan; Stensbo-Smidt, Kristoffer; Gieseke, Fabian; Pedersen, Kim Steenstrup; Igel, Christian (2017): Big Universe, Big Data: Machine Learning and Image Analysis for Astronomy. In IEEE Intelligent Systems 32 (2), pp. 16–22.

Krieger, Felix; Drews, Paul (2018): Leveraging Big Data and Analytics for Auditing: Towards a Taxonomy. In Proceedings of the ICIS 2018. International Conference on Information Systems. San Francisco, USA, 13.12.2018-16.12.2018.

Kumari, Aparna; Tanwar, Sudeep; Tyagi, Sudhanshu; Kumar, Neeraj; Maasberg, Michele; Choo, Kim-Kwang Raymond (2018): Multimedia big data computing and Internet of Things applications: A taxonomy and process model. In Journal of Network and Computer Applications 124, pp. 169–195.

Levy, Yair; Ellis, Timothy J. (2006): A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. In Informing Science: The International Journal of an Emerging Transdiscipline 9, pp. 181–212.

Linda, Lai Sau Ling; To, Wai Ming (2014): Big Data: Trend, Taxonomy, and Challenges. In The 3rd International Conference on Network, Communication and Computing. Hong Kong, China, 26.12.2014-28.12.2014, pp. 1–6.

Miller, Holmes E. (2013): Big-data in cloud computing: a taxonomy of risks. In Information Research 18 (1).

Miloslavskaya, Natalia; Tolstoy, Alexander; Zapechnikov, Sergey (2016): Taxonomy for Unsecure Big Data Processing in Security Operations Centers. In IEEE 4th International Conference on Future Internet of Things and Cloud Workshops. Vienna, Austria, 22.08.2016 - 24.08.2016, pp. 154–159.

Mohamed, Azlinah; Najafabadi, Maryam Khanian; Wah, Yap Bee; Zaman, Ezzatul Akmal Kamaru; Maskat, Ruhaila (2019): The state of the art and taxonomy of big data analytics: view from new big data framework. In Artificial Intelligence Review, pp. 1–49.

Müller, Oliver; Fay, Maria; Vom Brocke, Jan (2018): The Effect of Big Data and Analytics on Firm Performance: An Econometric Analysis Considering Industry Characteristics. In Journal of Management Information Systems 35 (2), pp. 488–509.

Murthy, Praveen; Bharadwaj, Anurag; Subrahmanyam, P. A.; Roy, Arnab; Rajan, Sree (2014): Big Data Taxonomy. Edited by Cloud Security Alliance.

Nahhas, Abdulrahman; Bosse, Sascha; Staegemann, Daniel; Volk, Matthias; Turowski, Klaus (2019): A holistic view of the server consolidation and virtual machines placement problems. In Proceedings of the 15th International Conference on Signal Image Technology & Internet based Systems. Sorrento, 26.11.2019-29.11.2019.

Nickerson, Robert C.; Varshney, Upkar; Muntermann, Jan (2013): A method for taxonomy development and its application in information systems. In European Journal of Information Systems 22 (3), pp. 336–359.

NIST (2019): NIST Big Data Interoperability Framework: Volume 1, Definitions, Version 3. Gaithersburg, MD.

Patgiri, Ripon (2018): Taxonomy of Big Data: A Survey.

Pathak, Ajeet Ram; Pandey, Manjusha; Rautaray, Siddharth (2018): Construing the big data based on taxonomy, analytics and approaches. In Iran Journal of Computer Science 1 (4), pp. 237–259.

Ramírez-Gallego, Sergio; García, Salvador; Mouriño-Talín, Héctor; Martínez-Rego, David; Bolón-Canedo, Verónica; Alonso-Betanzos, Amparo et al. (2016): Data discretization: taxonomy and big data challenge. In Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 6 (1), pp. 5–21.

Shah, Syed Attique; Seker, Dursun Zafer; Hameed, Sufian; Draheim, Dirk (2019): The Rising Role of Big Data Analytics and IoT in Disaster Management: Recent Advances, Taxonomy and Prospects. In IEEE Access 7, pp. 54595–54614.

Siddiqa, Aisha; Hashem, Ibrahim Abaker Targio; Yaqoob, Ibrar; Marjani, Mohsen; Shamshirband, Shahabuddin; Gani, Abdullah; Nasaruddin, Fariza (2016): A survey of big data management: Taxonomy and state-of-the-art. In Journal of Network and Computer Applications 71, pp. 151–166.

Singh, Rakesh Kumar (2018): Taxonomy of Big Data Analytics: Methodology, Algorithms and Tools. In International Journal on Future Revolution in Computer Science & Communication Engineering 4 (12), pp. 101–104.

Srinivasan, Madhan Kumar; Padmanaban, Revathy (2018): State-of-the-art Big Data Security Taxonomies. In Y. Raghu Reddy, Vasudeva Varma, Jane Huang Cleland, Umesh Bellur, Shubashsis Sengupta, Naveen Sharma et al. (Eds.)Proceedings of the ISEC 2018. The 11th Innovations in Software Engineering Conference. Hyderabad, India, 09.02.2018 - 11.02.2018. New York, New York, USA: ACM Press, pp. 1–7.

Staegemann, Daniel; Volk, Matthias; Jamous, Naoum; Turowski, Klaus (2019a): Understanding Issues in Big Data Applications - A Multidimensional Endeavor. In Twenty-fifth Americas Conference on Information Systems. Cancun.

Staegemann, Daniel; Volk, Matthias; Nahhas, Abdulrahman; Abdallah, Mohammad; Turowski, Klaus (2019b): Exploring the Specificities and Challenges of Testing Big Data Systems. In Proceedings of the 15th International Conference on Signal Image Technology & Internet based Systems. Sorrento, 26.11.2019-29.11.2019.

Timmins, Kate A.; Green, Mark A.; Radley, Duncan; Morris, Michelle A.; Pearce, Jamie (2018): How has big data contributed to obesity research? A review of the literature. In International journal of obesity (2005) 42 (12), pp. 1951–1962.

Vesset, Dan; Gopal, Chandana; Schubmehl, David; Olofson, Carl W.; Bond, Steward; Fleming, Maureen et al. (2017): IDC's Worldwide Big Data and Analytics Software Taxonomy. Edited by International Data Corporation.

Volk, Matthias; Staegemann, Daniel; Pohl, Matthias; Turowski, Klaus (2019): Challenging Big Data Engineering: Positioning of Current and Future Development. In Proceedings of the IoTBDS 2019. 4th International Conference on Internet of Things, Big Data and Security. Heraklion, Crete, Greece, 02.05.2019 - 04.05.2019: SCITEPRESS - Science and Technology Publications, pp. 351–358.

Vom Brocke, Jan; Simons, Alexander; Niehaves, Björn; Reimer, Kai; Plattfaut, Ralf; Cleven, Anne (2009): Reconstructing the Giant. On the Importance of Rigour in Documenting the Literature Search Process. In Proceedings of the ECIS 2009. 17th European Conference on Information Systems. Verona, Italy, 08.06.2009-10.06.2009.

Wang, Shiyong; Wan, Jiafu; Zhang, Daqiang; Di Li; Zhang, Chunhua (2016): Towards smart factory for industry 4.0: a self-organized multi-agent system with big data based feedback and coordination. In Computer Networks 101, pp. 158–168.

Wang, Yichuan; Kung, LeeAnn; Byrd, Terry Anthony (2018): Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. In Technological Forecasting and Social Change 126, pp. 3–13.

Webster, Jane; Watson, Richard T. (2002): Analyzing the Past to Prepare for the Future: Writing a Literature Review. In MIS Quarterly 26 (2), pp. xiii–xxiii.

Wu, Desheng; Cui, Yiwen (2018): Disaster early warning and damage assessment analysis using social media data and geo-location information. In Decision Support Systems 111, pp. 48–59.

Wu, Lengdong; Yuan, Liyan; You, Jiahuai (2015): Survey of Large-Scale Data Management Systems for Big Data Applications. In Journal of Computer Science and Technology 30 (1), pp. 163–183.

Zerdoumi, Saber; Sabri, Aznul Qalid Md; Kamsin, Amirrudin; Hashem, Ibrahim Abaker Targio; Gani, Abdullah; Hakak, Saqib et al. (2018): Image pattern recognition in big data: taxonomy and open challenges: survey. In Multimedia Tools and Applications 77 (8), pp. 10091–10121.