big data and cognitive computing

MDPI

Article

What Is (Not) Big Data Based on Its 7Vs Challenges: A Survey

Cristian González García [1,*] and Eva Álvarez-Fernández [2]

[1] Department of Computer Science, Faculty of Sciences, University of Oviedo, C/Federico García Lorca s/n, 33007 Oviedo, Asturias, Spain
[2] Independent Researcher, 33007 Oviedo, Asturias, Spain
* Correspondence: gonzalezcristian@uniovi.es

Abstract: Big Data has changed how enterprises and people manage knowledge and make decisions. However, when talking about Big Data, so many times there are different definitions about what it is and what it is used for, as there are many interpretations and disagreements. For these reasons, we have reviewed the literature to compile and provide a possible solution to the existing discrepancies between the terms Data Analysis, Data Mining, Knowledge Discovery in Databases, and Big Data. In addition, we have gathered the patterns used in Data Mining, the different phases of Knowledge Discovery in Databases, and some definitions of Big Data according to some important companies and organisations. Moreover, Big Data has challenges that sometimes are the same as its own characteristics. These characteristics are known as the Vs. Nonetheless, depending on the author, these Vs can be more or less, from 3 to 5, or even 7. Furthermore, the 4Vs or 5Vs are not the same every time. Therefore, in this survey, we reviewed the literature to explain how many Vs have been detected and explained according to different existing problems. In addition, we detected 7Vs, three of which had subtypes.

Keywords: Big Data; survey; challenges; Data Mining; KDD; Vs

Citation: González García, C.; Álvarez-Fernández, E. What Is (Not) Big Data Based on Its 7Vs Challenges: A Survey. *Big Data Cogn. Comput.* **2022**, *6*, 158. https://doi.org/10.3390/bdcc6040158

Academic Editor: Carson K. Leung

Received: 28 October 2022
Accepted: 6 December 2022
Published: 14 December 2022

1. Introduction

Over the last few years, certain things have changed in knowledge management. Now, more than ever, there are more readable, obtainable and useful data for people and companies. Companies are collecting more data than they know how to deal with [1]. Therefore, new knowledge and new technologies are needed to collect, store, process and display such a huge amount of data and take advantage of them [1]. In other words, thanks to these large amounts of data, companies can measure and know more about their businesses and thus use this knowledge in decision-making and in improving the performance of the company, its productivity, its competitiveness, its processes and its innovation [1–4].

One of the problems of most modern XXI century businesses is that they ignore relevant data because the vast majority of these businesses can electronically access piles of data, such as transactions made, the state of the assembly line, or customer data [5]. Thus, the accessibility and abundance of this information has made Data Mining a matter of importance and necessity [6].

In 1992, W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus estimated that the amount of data stored in existing databases would double every twenty months, or even faster, considering that both the size and number of databases would increase [7]. However, the current situation has seen a flood of data from different media and formats, which has managed to overcome the capacity for processing, analysing, storing and understanding these datasets [8].

Thus, one question arises: Is the current amount of data bigger before? The world is currently experiencing a data revolution, or, in other words, a flood of data, as a large amount of data flows every minute through the network from different sources using

different channels because we are in the digital age [9]. There is so much data in the world that the situation is overwhelming, as the amount of data is constantly growing and there is no end in sight [10]. In 2012, humans created around 2.5 quintillion bytes of data every day, meaning that 90% of existing data had been created in the last two years [11]. In addition, it is estimated that in 2024, we will generate 149 zettabytes every day [12]. This means that new tools are needed in knowledge management. Tools that are able to work with many different types of data in a fast, simple and effective way.

On the one hand, companies are trying to analyse all these data in search of trends that help them figure out what the customer thinks about the company or what they need or want. Examples of this are those shown by Hadoop on its official website [13] because in it we can find Facebook with its system of suggestions of friends and ads or LinkedIn with its system of suggestions of contacts and companies. Other companies dedicated to ads, such as Google [14] or Double Click, analyse all the cookies received to see which sites the customer frequents and thus show ads based on their browsing habits. Similar cases include those involving audience measurement services, such as Nielsen and Gartner or Strands and Amazon recommendation systems. Other times, as the McKinsey report rightly points out [15], if Big Data were used by different entities to produce money, as is the case with the medical services of the United States of America, they could obtain 300 billion dollars. In the European Union, 250 billion dollars of its public administration, retailers increase by 60% of the margin of operations or even produce a surplus of 600 billion dollars annually from the geolocation of people.

Thus, an important issue is to improve the ability to manage, understand and act on this huge amount of data in order to increase the understanding of the technologies that are needed to manipulate and mine such amounts of information and thus apply this knowledge to other fields, such as health, education, energy and defence [16]. Decades ago, companies had already used information from related databases to make predictions using data analysis, Data Mining and Knowledge Discovery in Databases (KDD) are often referred to as traditional techniques that provided rigorous decision-making techniques [1]. Other times, it may only be necessary to perform data analysis, or maybe we want to look for patterns within these to make predictions using Data Mining. However, many times, what each of them is and when they have to be applied, or even when they are named Big Data, can be confused. In addition, within KDD, there are different phases to perform, but not all authors perform the same. Therefore, what these techniques are, their differences, how they work, for which they are applied, their applications and a possible sequence of phases for KDD are explained in Section 2.

Due to the huge amount of growing information appearing and the properties of the datasets, this data is no longer manageable by traditional methodologies, techniques and tools [8,17]. Thus, Big Data tools and techniques make it possible to manage different mass datasets and apply traditional techniques to this data. These data differ from traditional ones in that they have some characteristics that also coincide with some of their challenges. These properties are the 'Vs', although the number and the ones included may vary according to the author. The study of the 'Vs' is presented in Section 3, which contains the methodology, the articles detected by the database and the type of article and the results and discussion about these.

After this study, we detected a total of 16 different 'Vs', with 10 'Vs' being the maximum used by one author. However, after the study, we proposed a total of 7 important 'Vs' according to those most used by the different authors. Notwithstanding, some of them even have subdivisions. These are volume, velocity, variety, veracity, variability, value and visualisation. Section 4 explains these '7Vs', considering different authors and their characteristics and challenges.

Finally, this survey concludes in Section 5 with conclusions and possible future work.

2. Data Mining versus KDD versus Big Data

There is some controversy when it refers to Data Mining and Big Data. In addition, when looking back in time, there are other terms that are included when talking about all these, such as Knowledge Discovery in Databases (KDD) and data analysis.

On the one hand, we have the problem of Data Mining being misunderstood with KDD, as some authors consider it the same [18–20], while others clearly show, based on references from different authors, that KDD is a complete process and Data Mining a step within the KDD process [6,21,22], but without explaining the differences between the two and giving only vague nuances.

On the other hand, we have the authors for whom Big Data and Data Mining are the same, or to be even more confused, they add the term data analysis. However, both terms are different, and despite having things in common and sometimes using the same tools, their purposes are different. One of the first examples of confusion here is that between data analysis and Data Mining [18].

Thus, even if it is said that both are the same or that Data Mining is currently known as Big Data due to the large amount of data that must be handled quickly [23], they are half-truths. Others, however, say that Big Data is scalable Data Mining [24]. Nevertheless, this requires the use of specific tools that hold and support the processing of large volumes of data. Notwithstanding, the datasets that Big Data works with have a number of characteristics that are known as the '7Vs' of Big Data.

In the following subsections, it will be defined exactly what data analysis and Data Mining are, as well as their differences, some details in data mining patterns and explanations about the classification of their methods. The next section will explain what KDD is and what phases it is made up of. Finally, in the last subsection, the term Big Data, with different definitions of both companies and academics, will be discussed so that an exact definition of what Big Data is exactly will be provided.

2.1. Data Mining

The search for patterns within data is known by many different names, such as knowledge extraction, information discovery, information gathering, data processing and Data Mining [21,25]. The last term is the most commonly used by statisticians, database researchers, information system managers and in business [21], as well as in informatics.

This term was coined by statisticians in the 1960s to refer to the bad practice, in their opinion and with some negative connotation, of analysing data without an initial hypothesis [25]. This coincides when computers were introduced to obtain data patterns from a sample of a total population [21,26]. They named this process "data dredging" or "data fishing" [21,23,25,26]. Meanwhile, the term currently used, Data Mining, appeared in the 1990s in the database community [25].

Data Mining is a subfield of computer science that touches many parts of computing, such as database management, artificial intelligence, machine learning, pattern recognition and data visualisation [19].

Data Mining seeks to create an automated analysis of large and complex datasets [19]. These sets can range from databases to sky maps, weather information, satellite information, industrial control processes, and many more [19].

Data mining is based on the extraction of previously unknown data, but which can be potentially useful due to the information they can provide to try to obtain patterns, also known as models [21,23], and relationships that may be generalised to make future predictions or make important decisions [6,10,19,27]. However, many of these patterns obtained will be banal and uninteresting; others are false or possess accidental coincidences within a dataset. In the event that these patterns are meaningful, they may allow for non-trivial predictions of new data [10]. Therefore, the purpose of Data Mining is to search for and analyse large amounts of data to discover useful patterns or relationships that serve to predict the future [18].

The goals of using Data Mining algorithms are, according to Lynn Greiner [18], the following:

- Apply clustering to group data into groups of similar elements.
- Search for an explanatory or predictive pattern for a target attribute in terms of other attributes.
- Search for frequent patterns and sub-patterns.
- Search for trends, deviations and interesting correlations between attributes.

Examples of the use of Data Mining can be found, for instance, in applications that try to predict things, such as weather or climate change, based on data describing past events to create a model that serves to predict possible future events that have the same pattern. Other examples of its application are in economics [10], or to obtain genetic information and anticipate existing diseases and treatments [28]. All Data Mining applications focus on obtaining interpretations of the data because they are looking for trends and correlations rather than testing a hypothesis [18].

2.1.1. Data Analysis versus Data Mining

Sometimes, there is confusion between the terms "data analysis" and "Data Mining" [18]. Data analysis analyses existing data and applies statistical and visualisation methods to test hypotheses about the data to discover exceptions. Meanwhile, Data Mining looks for trends within the data that can be used for further analysis. Therefore, Data Mining is able to provide new knowledge, which are the patterns, totally independent of preconceived ideas, that is, hypotheses.

According to Greiner [18], data analysis can be seen as the collection of methods for drawing inferences from data, thus extracting global information that is generally the property of the analysed data.

On the other hand, Data Mining can be defined according to the definition given by Lynn Greiner [18], and which was used by Microsoft when describing its Big Data architecture for the National Institute of Standards and Technology (NIST) survey of the United States of America [29]. Thus, according to them, the definition is: "Data Mining is the process of extracting data and its subsequent analysis from many dimensions or perspectives to produce a summary of the information in a useful way that identifies relationships within the data. There are two types of data mining: the descriptive one, which gives us information about the existing data, and the predictive one, which makes forecasts based on the data".

2.1.2. Patterns in Data Mining

Patterns, also known as models, that are obtained by applying Data Mining are achieved by using different machine learning techniques [5,10,23], statistics and artificial intelligence applied on sample datasets to reduce these sets to small understandable patterns [5].

Decision trees and association rules are often the basis for Data Mining [18]. However, other machine learning methods, such as artificial neural networks (ANN), are also used. Even though ANNs have been successfully applied, they often take a long time to generate patterns that are sometimes not understandable, and that is why they are not used very frequently [18]. Other types of algorithms that are often used are very important, as scanning techniques in data analysis are clustering algorithms, because at first there is no knowledge about the possible distributions existing in the data [18].

The patterns obtained can be of two types. These can be incomprehensible black boxes or transparent boxes that show the structure of the mentioned pattern, known as structural patterns. However, both possibilities are equally good. The difference in structural patterns is that these extracted patterns are represented in a structure that can be examined, reasoned and used to inform future decisions. Hence, its name, since they allow the data to be explained based on the pattern structure [10].

2.1.3. Classification in the Methods of Data Mining

According to Lior Rokach and Oded Maimom, Data Mining methods can be classified according to the taxonomy shown in [6], which shows that there are two possible types. The first is *verification*, which is used to verify a hypothesis. For this reason, it is less associated with Data Mining, as Data Mining focuses on the selection of a hypothesis.

The second type is *discovery*, which is used to obtain new pattern rules from a dataset. Within the methods of discovery, there are two branches [6,18], the first one being *descriptive* methods that focus on the comprehension of how data operate to create reports. The second branch is *predictive* methods, which contain those methods that aim to construct a behaviour pattern to obtain new samples to predict the values of one or more related variables in the sample. *Predictive* methods sometimes also help to understand data [30]. The latter can be differentiated into *regression* and *classification* methods.

In *regression*, we learn a function that classifies an element according to continuous variables to obtain a prediction or forecast [26,31]. Examples of *regression* are studies to determine the smoking mortality rate, the amount of biomass present in a forest, the probability of survival of a patient after a series of diagnostic tests or the prediction of demand for a product.

Meanwhile, *classification* methods learn a function that classifies an element into several defined classes [32]. A very relevant use of such methods is in image classification. This method is where many types of artificial intelligence algorithms are found, such as Artificial Neural Networks, Bayesian networks, Decision Trees, and Support Vector Machine (SVM) [33].

2.1.4. Data Mining Applications

Big Data needs machine learning techniques to deal correctly with the flooding of existing data and thus analyse these data to understand them, but it also provides what is necessary to evolve and improve the current Artificial Intelligence. This was already announced in 1990 [7] and is reflected today in the use of different techniques [30]. In recent years, many companies such as Yahoo! and Google have made great advances in AI thanks to the use of Big Data. AI requires large amounts of data to be trained, thus it can learn, so that they can create good quality models, either to create artificial intelligences, such as Cortana, Google Now, or Siri, translation text engines or linguistic models for natural language processors. This is why we could use in Big Data any of the algorithms or uses explained here; it depends on the quantity of data.

Data Mining has different types of algorithms specialising in different types of works. First, we have classification in which we can use different algorithms, such as Bayesian Networks, Decision Trees, Super Vector Machines, and ANN.

In the literature, we find some examples of Bayesian Networks for doing classification. The first uses are to detect anomalies and faults. Wang et al. [34] developed a Feature Weight Mixed Hidden Naïve Bayesian Network to improve performance and effectiveness in monitoring anomalies based on continuous variables of large-scale processes in comparison with a Mixed Hidden Naïve Bayesian Network in the practical case of Zhoushan thermal power plant in China. He et al. [35] used Naïve Bayes classifier to obtain a fault diagnosis to detect and classify analogue circuit faults better than previously published works. Zhen Xue et al. [36] has proposed a Naïve Bayes classifier to use on Field-Programmable Gate Array and obtain a better real-time efficiency than other Bayesian classifiers and Convolutional Neural Networks (CNN) accelerators.

Other times are used to classify text, as in Sanchís et al. [37], where the authors have used Naïve Bayes to classify the confidence estimation in Speech Recognition and then detect words that may have been misrecognised. Shirakawa et al. [38] use Naïve Bayes to detect noisy text, which means text with meaningless or misleading terms from their main topic. Kustanto et al. [39] use a Naïve Bayes classifier to detect the positive or negative

sentiment of the population about the healthcare application of the Indonesian government.

Other examples of Bayesian Networks are in the food industry to create a Naïve Bayes classifier to discriminate cacao nibs from six varieties of cacao clones [40], or in the combat against COVID-19 [41,42].

Other algorithms used to classify are Decision Trees. One of the possible uses is in healthcare to predict the admission in the Intensive Care Unit (ICU) of hospital of patients with COVID-19 at the point of admission to the hospital, trying to have a very good use of resources in those places [43]. Ghane et al. [44] use them to diagnose Parkinson's disease, Elhazmi et al. [45] for predicting mortality in critically ill adults with COVID-19 in the ICU, Hiranuma et al. [46] for assessing the effectiveness of treatment in intra-abdominal infections using cefmetazole, and in disease diagnosis, such as genomics or cancer, as explained in [47]. Other uses are in security [48] or in high efficiency video coding [49].

In the case of Super Vector Machines, we can find examples in the literature of use for classification and regression [50,51]. Astuti et al. [52] use SVM and Random Forest separately to classify raw chicken with E. coli and without it using gas sensors. Another case is using One-Class Super Vector Machines (OCSVM) to detect industrial anomalies in an effective way and obtain better anomaly detection performance [53]. In healthcare, SVM is used on medical diagnostics due to its high classification accuracy and to provide online diagnostics [50]. For instance, Bernardini et al. [54] use Sparse Balanced Support Vector Machines (SB-SVM) for diagnosing type 2 diabetes.

Other researchers try to improve the use of SVM by comparing it with other novel algorithms, such as in [51,55], because the complexity of SVM is very large to work with large datasets [51,55].

ANN have different uses. One of them is to use it to classify images. Azgomi et al. [56] use ANN to detect fruit diseases by using their pictures, saving time to farmers in the eye diagnosis part, and classifying them into bitter rot, black rot, scab and healthy fruits. Zhu et al. [57] use them to match images in the context of remote sensing. Qin et al. [58] use ANN to reconstruct high quality cardiac images taken using magnetic resonance. Wu et al. [59] apply ANN to detect breast cancer. On the other hand, other researchers have used ANN to distinguish between a liver with cancer and non-cancer lesions using CT images [60].

Ulloa-Cazarez et al. [61] use different types of ANN and compared them all to obtain the better one to predict student performance and give new strategies for making decisions. Ibragimov et al. [62] use ANN to predict the post-treatment of Stereotactic Body Radiation Therapy in the liver and identify critical-to-spare liver regions using 3D images, obtaining better results than SVM and Random Forest. Other uses of ANN are security to detect a good or bad connection [63].

Other uses of these algorithms are when they are mixed. For instance, the next paper shows how they combined Naïve Bayes and SVM to propose a better algorithm to detect network intrusions and improve security [64,65].

In other cases, we can find that some researchers have mixed Bayesian Networks and Decision Trees. In [66] used them to predict the delay transferring a patient from the ambulance to the hospital in Canada and trying to be proactive to avoid the Ambulance Offload Delay.

On the other hand, regression is a technique used to obtain a prediction or forecast. It is used in a variety of different studies. They used it to predict high-voltage switchgears based on the contact temperature and forecast the long-term temperature [67]. In addition, it is used to improve Speech Emotion Recognition according to real applications [68]. Sometimes, it is necessary to merge it with other algorithms, such as K-Nearest Neighbour (KNN), to improve the results. In this case [69], the authors extracted the features using linear regression and used KNN to detect latent defects and process variations in

transistors. Another merge is between regression, Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN) to forecast wind speed in short-terms [70].

To obtain descriptions, one of the methods is clustering. As with the previous techniques, they have different algorithms.

One of these is K-means. Abbas et al. [71] used K-means and K-medoids in the medical field using data from a hospital about births. Rong et al. [72] presents a novel algorithm based on K-means to have a better performance and improve the text clustering of different types of news.

Other researchers have proposed new algorithms. For instance, Jeong et al. [73] proposed a new algorithm to perform clustering and apply it to detect complex contagion in networks. [74] proposed the Z-Clust algorithm and applied it to improve the discovery of new drugs.

More research with clustering is needed to create groups with data. Tian et al. [75] use this technique to obtain separate cells on groups under different biological conditions. Krishnaveni et al. [76] used it to classify aerosols that can be found in a rural area, using Fuzzy C-means and obtaining three clusters. In [77], apply clustering to categorise packets from unknown Internet traffic and improve security and save resources.

Other times, it is used to traffic. For instance, to cluster maritime traffic to discover high risk or density in different traffic scenarios and improve maritime traffic surveillance, designing new designs and new strategies in an easier way [78]. Another example is to improve the communication systems of High-Speed Railway, where clustering is used to obtain the specific characteristics of this case [79].

As a previous example, this algorithm can be mixed, such as Feigin et al. [80], have done mixing Generative Adversarial Networks (Gans) with Clustering and offering a framework and testing it with three datasets of images of animals, vehicles and human faces.

2.2. Knowledge Discovery in Databases

The term Knowledge Discovery in Databases (KDD) was coined by Gregory Piatetsky-Shapiro in a 1989 workshop [81], as explained in [26] and stated by the author himself in [25]. In addition, it was confirmed after having carried out a check by performing several studies in previous years in different search engines of scientific journals, where the previous existence of this term was not found. The purpose of this term was to emphasise that knowledge is the final product of data-driven discovery [26].

Knowledge Discovery in Databases or KDD is the automatic identification process valid, novel and potentially usable of pattern understanding in large databases [26,30]. This process must be automatic due to the large amount of data that can be collected and that humans can already barely digest but that machines can facilitate. This is the purpose of KDD: trying to address the problem of data overload [26].

As what happens with Data Mining, KDD is the interaction of different fields, including databases, machine learning, pattern recognition, statistics, artificial intelligence, expert systems, data visualisation and high-performance computing [21,26].

KDD has been used in a wide variety of applications in different types of databases [7,26,82]. These are, for example, in medicine for the discovery of side effects of drugs or genetic sequence analysis, in finance for the prediction of bankruptcies and stock, in agriculture for the classification of diseases, in the social field for voting prediction, in marketing for the identification of subgroups and purchasing patterns, in insurance companies for fraud detection, in engineering to create expert automotive diagnostic systems, to estimate work, at the Hubble telescope, in performing a sky mapping, in the search for volcanoes on Venus, in biosequence databases, for earth geophysics in the inference of earthquakes and in the analysis of atmospheric data, among many other examples.

2.2.1. The Term KDD

As mentioned above, there is confusion between what KDD is and what Data Mining is, as some people affirm they are the same. Therefore, different definitions are collected below to state exactly what KDD is and its differences with Data Mining.

According to Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, KDD is described as [21,83]: "The non-trivial process of identifying patterns in valid, new, potentially useful, and ultimately understandable data". In another article by these same authors, they clearly specify that KDD refers to the entire process of discovering knowledge from data, while Data Mining is a step in this process, exactly the step in which algorithms are specified to extract patterns from the data [26].

On the other hand, W. J. Frawley, G. Piatetsky-Shapiro and C. J. Matheus add that KDD is the non-trivial extraction of potentially useful and previously unknown information from data [7].

Therefore, care must be taken not to confuse it with Data Mining, as Data Mining is one of the central steps of KDD; it is exactly the application of algorithms to obtain patterns or models, while KDD is the whole process of discovery [6,21,84].

Then, it can be said that Knowledge Discovery in Databases or KDD is the whole process of non-trivial and automatic obtaining of information and patterns of a dataset, which includes all the necessary phases for it, and which has as its cornerstone Data Mining in order to find valid, new, useful and ultimately understandable patterns. The phases of KDD are the analysis of the problem, data processing, Data Mining application to find patterns, pattern evaluation and pattern unfolding.

2.2.2. Phases of KDD

KDD consists of several phases or methods, depending on the author. These phases are what differentiate KDD from Data Mining, as they are both pre- and post-application phases of Data Mining. These phases serve to ensure that knowledge is derived from data, as the patterns discovered in Data Mining are sometimes banal because they are meaningless or invalid [26].

L. Rokach and O. Maimom [6] presented a hybrid method of KDD, based on other authors and consisting of 8 phases. Basically, what the authors of this method did was merge phases 3 and 4 of other methods to make all the data pre-processing in a single phase as opposed to 9-phase methods [26]. Here, we added more information and clarifications to the phases according to the different articles and used as a base the presented by L. Rokach and O. Maimom [6]. The KDD phases are as follows:

1. Development of the **understanding of the application domain**, relevant background knowledge and end-user objectives [6]. It must be relevant and provide a financial benefit [7].
2. **Selection of a dataset** or subset of data on which the study is to be carried out [6]. This set must have a desirable number of samples and the number of incomplete data or data with noise must be very small [7].
3. **Data pre-processing:** It has three steps [6].
    a. **Reduction** of the dimension through the selection of functions and the taking of useful samples for the intended purpose, which offers a reduction in the number of variables to be considered [26].
    b. **Cleaning** of data to eliminate noise generated by different data types, extreme values, and missing values due to default or non-compulsory values [7].
    c. **Transformation** of data to extract its attributes by discretisation. Discretisation occurs when specific data types are required for an algorithm to work properly [85]. What is done is the division of a continuous (numerical) type attribute into at least two subsets [86].
4. **Choose the right task or Data Mining method**. They can be classification, regression, clustering or summarisation [6].

5. **Choose the Data Mining algorithm** by selecting the specific method to be used for pattern searching [6]. A data mining algorithm is nothing more than a set of heuristic calculations and rules that allow a model to be created from data [31]. For example, artificial neural networks, support vector machines, Bayesian networks, decision trees or different clustering or regression algorithms. This phase is difficult as different algorithms can be used to perform the same work but each will give a different output [31].
6. **Use the chosen Data Mining algorithm** [6].
7. **Evaluation and interpretation of the extracted patterns**. This may mean having to iterate again between the previous phases. In addition, this pattern may involve viewing the extracted patterns or data [26].
8. **Display the pattern found in another dataset** for use and testing, and/or documentation of the pattern [6].

2.2.3. KDD Applications

The main part of KDD is the use of Data Mining algorithms, but according to the different phases of KDD to check that the whole process has been done correctly according to this process. Then, we could use it for any of the applications explained in Section 2.1.4. Here, we show some explicit uses of KDD. In addition, in this work, they use different Data Mining algorithms, but apply different phases of KDD. However, sometimes they do not explain which phases.

For instance, Vučetić et al. [87] proposed a novel data mining algorithm to mine knowledge from real-world datasets and detect relationships between the data. Oliveira et al. [88] use the KDD process to make decisions to reduce harmonic distortions. Chen et al. [89] create a tutorial on KDD for movie recommendation, analysing the rating of millions of movies, obtaining that senior people are censored less than young people. Molina-Coronado et al. [90] can be found a survey about the use of KDD in Intrusion Detection Methods in information systems.

Sanchez et al. [91] apply this methodology to design and implement a tool for mood detection about the suggestions of new songs according to the music playlist and genre. They used ANN. Kam et al. [92] suggest the use of KDD in the Internet of Things (IoT) [93] and Industry 4.0, also known as Industrial Internet of Things, to integrate and apply data mining to the heterogeneity and unstructured data that they produced. Rosa et al. [94] apply it to analyse data on the satisfaction of students with the quality of the different services that are offered in a private high school using Binary Logistic Regression, Linear Programming mathematical model, and Fisher's Discriminant Linear Function (FDLF).

Table 1 shows a comparison of the KDD phases used in these articles. In all of them, we can suppose that they did the first one because they are research articles. In these cases, we use the mark '?'. We can find that Vučetić et al. [87] and Chen et al. [89] do not explain them explicitly, and Rosa et al. [94] explains briefly two phases used in the article.

On the other hand, Oliveira et al. [88] use almost all steps, but the Data Mining steps are done just as one and stepping over the method phase, which can be deduced from the algorithm used, and do not try with other different datasets, just with other types of algorithms.

In [90] uses five phases. However, they called the second step Data Collection, separated the pre-processing (cleaning and other features) and data reduction, put together under Data Mining (techniques, algorithms and use), and finalises with the interpretation and evaluation. Additionally, they used different datasets, but they did not include them in their phases.

In [91] explains it in five phases, but use more. However, they separated pre-processing (just for cleaning) from transformation and reduction (called data normalisation). In Data Mining they group 2 phases (Method and Algorithm). After that, they use the pattern and evaluate and interpret it. Notwithstanding, they do not test or explain the use of the pattern well with other datasets and do not include this in their phases.

Kam et al. [92] uses five phases without giving many details about the pre-processing or data mining part or the use of other datasets.

Table 1. Comparison between the KDD phases used in the articles.

	1	2	3 Pre-Processing			Data Mining			7	8
	Domain	**Selection/Data Collection**	**Data Reduction**	**Cleaning**	**Transformation**	**4 Method**	**5 Algorithm**	**6 Use**	**Interpretation and Evaluation**	**Another Dataset**
[87]	?									
[88]	?	X	X	X	X		X	X	X	
[89]	?									
[90]	?	X	X	X	X	X	X	X	X	?
[91]	?	X	X		X			X	X	
[92]	?	X	X		X			X	X	
[94]	?	X						X		

2.3. Big Data

The term Big Data first emerged in 1997 in a series of unpublished non-academic slides from Silicon Graphics (SGI) by John Mashey entitled "Big Data and the Next Wave of InfraStress" [95]. However, Big Data had already been discussed in 1984, but not in the context as it is known today. On the other hand, in the academic and publishable field, the first book in which this term was named was a 1998 data mining book [96]. All of this is according to the history of the term investigated by Francis Diebold in [97]. Sometimes Big Data is also called Data-Intensive Computing [14,98], or Big Data is referred to as the fourth paradigm [98,99].

Big Data is a term that is increasingly used to describe the process of seriously applying computational power to massively large, highly complex, heterogeneous and growing datasets that have multiple sources [8,100,101], which were therefore not manageable with the methodologies and tools used for Data Mining [8], which in 2009 already gave enough problems for data capture [98]. The analysis of this data with Big Data tools can lead to much stronger conclusions for Data Mining applications, although there have been many difficulties in its use [96]. Therefore, the purpose of Big Data is to answer different questions to get early answers or predictions about the future [102].

One of the problems of Big Data is the lack of consensus in its definition or the different definitions or opinions given by people [103]. This has been collected in an article that gathers the opinions of 40 relevant people [104], where it explains that many times, to define Big Data references are made to the '3Vs', while others refer to the tools used. Other times, there is confusion with Data Mining, because there are people, among which companies and researchers stand out, who affirm it is the same.

As can be seen, the term Big Data is very corrupt due to the lack of an exact definition and the different ambiguous definitions that different groups have given it, among which the academy and industry stand out [105]. Therefore, in the following subchapters, we will define the meaning of "big", and we will see the different definitions given by companies and the scientific community about which Big Data is.

2.3.1. Definitions of Companies and Academics

Many companies and researchers got into Big Data, and thus they created different articles signed by that company to establish what it is Big Data for them from their point of view and to talk about it.

One of these companies was Oracle in 2013 [2]. According to Oracle, Big Data encompasses traditional data from relational databases, data generated by machines such as sensors and logs, and social data, including social networks and different existing blogs. In addition, it considers that there are '4Vs' that Big Data must comply with, which are volume, velocity, variety and value. Thus, Oracle defines Big Data as a diverse dataset that is

obtained and analysed thanks to the use of specific tools and that helps to improve different aspects of companies, ranging from internal processes to the prediction of relevant data. As can be seen, they used the word "Big" to define the size of the data, the importance, and the influence of these.

Intel also gave its own definition, and for them, Big Data is the quantities of large, complex or unstructured data, which can be collected and analysed using emerging technologies such as Hadoop and MapReduce, in order to obtain significant knowledge [106]. For Intel, it only refers to the size of the data.

Gartner was another company that did its own article on Big Data [107,108]. In this article, they explain that Big Data is a term used to support the exponential growth, availability and use of current information to lead company decisions. However, these can be bad decisions that interfere with the company's architecture. In addition, they note that some managers focus only on the volume of data when it turns out that they must also look at velocity and variety, that is, the '3Vs'. For this, according to Gartner, "innovative forms of information processing are needed that allow for improved vision, decision making and process automation". For Gartner, the real challenge is to obtain patterns that can help companies make better decisions using innovative technologies. As seen, Gartner's definition speaks of '3Vs' and explains how the purpose of Big Data is to obtain patterns that help companies make decisions.

A special case was that of Microsoft, where they wrote an article about it, taking the perspective of different important workers within the Redmond company. According to the perspective of different workers or staff related to Microsoft [101], it could be summarised in that Big Data is the process of managing and working with massive amounts of digitised data that require high computing capacity to obtain useful information in order to make decisions quickly and efficiently using algorithms of artificial intelligence. Here, 'Big' refers to the massive amount of data and the importance of these.

IDC, a market analysis firm, dedicated a report to Big Data [109]. This report highlights the importance of analysing data to generate value by correctly extracting information from the digital universe. The problem is that data are growing exponentially and encapsulated in many different types of files. To this end, they argue that this is now possible thanks to the convergence of technologies. In addition, they also highlight that Big Data is not something new and that it is not a 'thing', but a dynamic or activity that crosses many computer borders. Its definition is as follows: 'Big Data technologies describe a new generation of technologies and architectures designed to economically extract value from very large and widely varied data volumes, allowing a high capacity for capture, discovery and/or analysis'. As seen in this definition and the rest of the article, they take into account '3V', volume, variety and velocity. In addition, they focus mainly on new technologies that allow compliance with these '3V'. However, it refers only to data analysis and not to what would be mining, which is to discover trends or, rather, prediction.

On the other hand, we have the NIST, which gives several definitions, all of which depend on the point of view from which it was taken. The first one has to do with the architecture and scalability of these to deal with data: 'Big Data consists of large data sets, with the characteristics of volume, variety, velocity and/or variability that require a scalable architecture for efficient storage, manipulation and analysis' [102]. According to this definition, Big Data refers to the inability of traditional data architectures to work with new datasets that require meeting the "4V" characteristics, which according to NIST are volume, variety, velocity and variability [110]. On the other hand, the other definition focuses on scalability and places Big Data as the heart of horizontal scaling, where everything is distributed among several computers and expanded with the addition of new equipment, against the vertical, where the internal capacity of a computer to improve performance is increased [102]: "The Big Data paradigm consists of the distribution of data systems using a horizontal, resource-independent scaling to achieve the scalability necessary for efficient processing of large data sets". Thus, in these definitions, they treat the word "Big" as large in volume and in need of computation.

Thus, NIST argues that Big Data is a term used to describe the large volumes of data in the connected, digitised, sensor-loaded and information-driven world [102] having this data overwhelmed by the traditional analysis approaches, known as Data Mining, and maintaining a data growth so fast that they manage to leave behind scientific and technological advances in this area. Big Data allows managing and treating these heterogeneous volumes, which do not necessarily have to be large or larger than formerly, but may be due to the required processing speed or efficiency required when processing a volume of data and which is not provided by traditional tools and methods. However, it must be borne in mind, as they well maintain with their definitions, that this term has been used to describe many concepts because there are different aspects interacting with each other and depending on the point of view in which it is looked at and where it is used.

Another definition was given by Jonathan Stuart Ward and Adam Barker [105], who conducted a study on the different definitions existing in the network that were given by different companies. The definition is as follows: "Big data is a term that describes the storage and analysis of large or complex datasets using a number of techniques that include, but are not limited to, Not Only SQL databases (NoSQL), MapReduce and Machine Learning".

The definition given by Wei Fan and Albert Bifet about Big Data is based on the ability to extract useful information from large datasets that, due to their volume, variety and velocity, whether static or streaming, it was not possible to be extracted before [8]. As can be seen, they consider "3V" and not only cling to the large amount of data, but to the variety and velocity, that is, efficiency. In addition, they consider it, similar to others, the evolution of ancient techniques, where KDD and Data Mining were encompassed.

On the other hand, the definition of Emmanuel Letouzé states that Big Data are the tools and methodologies that aim to transform massive amounts of raw data, both structured and unstructured, which are difficult to process with traditional software techniques in data for analytical purposes [9].

Other authors state that Data Mining is currently known as Big Data due to the large amount of data that must be handled quickly and that in order to deal with applications of this type, a new type of applications have been developed, which offer parallelism to work with these large amounts of data in a cluster of computers connected by Ethernet cables or switches, thus avoiding the use of supercomputers [23]. As can be seen, in this definition, they make it clear that Big Data is the evolution, as it is the architecture used.

On the other hand, the authors of [103] also studied this lack of definition and contrasted the different definitions found. Thus, in broad terms, they understand by Big Data the datasets that cannot be perceived, acquired, managed, and processed by traditional software and hardware computer tools within an acceptable time. Thus, this definition refers to traditional tools, something that is true but does not clarify the function or purpose that Big Data has, something that is quite important.

Instead, in this other survey [22], they define Big Data as "the ability to obtain private information about consumer preferences and products by crossing it with information from tweets, blogs, product evaluations and social media data opening a wide range of possibilities for organisations to understand their customers and predict what they want and demand, and optimise the use of resource". Following this, this definition only reflects the part of companies looking for information from their customers based on a few sites, when it turns out that Big Data is much more and allows not only to analyse information, but also trends of any kind, as it is applicable to everything, including astronomy, physics, defence and security, and any kind of data, such as images and video.

Based on everything seen so far, a possible definition of Big Data based on these definitions and according to what it does and the expected result of its use could be: Big Data is the useful search for information in datasets of any kind that is hardly feasible using traditional methodologies, techniques and tools due to the characteristics of the data to be analysed and that are known as "V", which are volume, velocity, variety, veracity, variability, value and visualisation, in order to obtain results that allow to analyse data, take

current trends, optimise the use of resources and/or predict the future, in a fast and efficient manner.

2.3.2. Big Data Applications

As shown before, Big Data needs new methodologies, techniques, and tools due to the characteristics of the data needed for a specific work. In this subsection, we show different Big Data works and the corresponding match with the 'Vs' presented in this survey.

Mohammadi et al. [111] explains the use of Machine Learning to obtain information and analyse data from the IoT and obtains learning from this domain. In addition, the heterogeneity of objects in these networks creates a large heterogeneity of data. Then, due to the quantity of devices that exist in the IoT, the IoT produces big data and/or has a fast creation of content, deriving this in close relationships with Big Data. In addition, they talked about the high noise in the data because of possible errors in the acquisition and transmission of it. Exactly, this explanation matches at least the Volume, Velocity, Variety, Veracity and Value 'Vs'.

Lin et al. [112] express in their survey the importance and necessity of gathering, analysing and visualising the medical big data that has appeared in recent years. They focus on chronic diseases and health monitoring. In this case, this matches with the Volume, Velocity and Visualisation 'Vs' at least.

Nti et al. [12] present a review of the applications and challenges of Machine Learning in Big Data. In it, they explain the techniques that are used to analyse these huge amounts of data in real time: 15% use Deep Neural Networks, 15% SVM, 14% ANN, 12% Decision Trees, and 11% ensemble learning techniques. This case matches at least the Volume and Velocity 'Vs'. Manley et al. [113] review the use of Machine Learning and/or Big Data in the ecosystem service field, using different algorithms to classify, apply regression, and clustering.

Nguyen et al. [114] surveyed and remarked on the use of Big Data in the oil and gas Industry 4.0. For instance, Big Data is important for managing and processing the data and making decisions to mitigate risks. However, the adaption is slow due to different problems, and the integrations with the existing systems according to have good cybersecurity. This one match at least with the Volume, Velocity and Value 'Vs'.

Rawat et al. [115] reviewed the evolution and use of Big Data in cybersecurity, where it is used as a security tool to improve protection. For example, to combat hacking, malware, social attacks, human errors, and advanced persistent threats. Sometimes, they used specific tools or other more generals (Hadoop, Spark, Cassandra, MongoDB), different machine learning or data mining techniques and algorithms or tools to visualise the information using tools or graphs. Then, in this field, the 'Vs' that match are Volume, Velocity, Variety, Value and Visualisation at least.

Ma et al. [116] propose the use of Big Data to create a sustainable digital twin in the IoT and improve manufacturing in a smart way. Here, with Big Data, they can gather data, make predictions, and mining in real time under complex conditions, saving energy and reducing the cost of production. In this case, Volume, Velocity and Value are the matches in the 'Vs', at least.

Jaber et al. [117] apply Big Data to predict climate factors according to information about natural disasters. To achieve this, they must process and manage a huge volume of data from different sources with interoperability problems. In addition, their tool helps monitor catastrophes. In this case, the 'Vs' match at least with Volume, Velocity, Variety, Value and Visualisation.

3. Literature Review

In this subsection, we explain the methodology used to find the articles that are focused on the 'Vs' of Big Data, the articles found in the electronic database, and the results and discussion about them.

3.1. Methodology

In addition, we have searched in three different web databases the most related articles that explain, highlight, or are focused on the 'Vs' of Big Data. As databases, we have used:

- IEEE: IEEE Xplore (https://ieeexplore.ieee.org/Xplore/home.jsp, assessed on 29 November 2022).
- SD: ScienceDirect–Elsevier (http://www.elsevier.com, assessed on 29 November 2022).
- Wiley (https://onlinelibrary.wiley.com/, assessed on 29 November 2022).

To search for articles, we used the search term 'Big Data AND "3V"' OR 'Big Data AND "3 V"', changing the number for the corresponding number of 'V' in all the databases. We use these terms because we would like to find all articles in which they are talking about Big Data and the 'Vs'. In addition, in this case, the 'Vs' can appear with the number together or separated, such as '3V' or '3 V'. We found between 3 and 9 'Vs' because the maximum number found was 9 'Vs' in IEEExplorer.

The scope has been Title, Abstract and Keywords because innovations and the main theme or important parts of articles are usually represented in some of these three parts.

The date was from the first appearance until 29 November 2022.

In these searches, we have excluded editorials, special issues, and articles that are not related to Big Data or are about Big Data but do not specify the importance of the 'Vs' in the title, Abstract or Keywords.

In this search, we found conferences, magazines, research articles (called journals in IEEE), books, chapters, and review articles. The first is from 2013 in IEEExplore, 2016 in ScienceDirect and 2017 in Wiley. The last one in each one is from 2022.

3.2. Articles by Database and Type

Next, we present the articles that were found in each of the databases and discuss how many 'Vs' they use. In addition, we show the years of the articles, the number and type of articles in that search (excluding exclusions), and the number of exclusions and their reasons. Finally, we found a total of 105 articles, 87 from IEEExplorer, 12 from ScienceDirect, and 6 from Wiley. Table 2 shows a summary of this information.

IEEExplore has 87 articles that match our criteria:

- 3V: 2013 to 2022: 27 Conferences, 3 Magazines, 1 Journal.
  - 2 conferences have been removed because they are about other fields.
- 4V: 2014 to 2022: 25 Conferences, 1 Magazine, 1 Book, 2 Journals.
  - 3 conferences have been removed because they are about other fields, and 1 more for being an editorial.
- 5V: 2014 to 2022: 18 Conferences, 3 Journals.
  - 3 conferences, 2 journals, and 1 magazine have been removed because they are about other fields.
- 6V: 2020: 1 Magazine.
  - 1 conference talks about 10 'Vs'.
- 7V: 2014 to 2021: 3 Conferences.
- 8V: 2019: 1 Conference.
  - 3 have been removed because they are about other fields.
- 9V: 0.
- 10V: 2021: 1 Conference.
  - It has appeared in the 6 'Vs' search.

In ScienceDirect, a total of 12 articles were found:

- 3V: 2018 to 2019: 3 Research articles, 1 Book chapter.
- 4V: 2016 to 2019: 1 Review article, 5 Research articles.
- 5V: 2016 to 2022: 1 Review article, 1 Research article.
- 6V, 7V, 8V, and 9V: 0.

Wiley has 6 articles according to our criteria:

- 3V: 2017 to 2022: 2 Books, and 1 Journal.
- 4V: 2019: 1 Journal.
  - 1 has been removed because it is about other fields.
- 5V: 2017: 1 Journal.
  - 3 have been removed: 1 is about Big Data but not about the 'Vs', and 2 are about other fields.
- 6V: 2022: 1 Book.
  - 1 has been removed because it is about Big Data but not about the 'Vs'.
- 7V: 0.
  - 2 have been removed because they are Issues and not about Big Data.
- 8V: 0.
  - 1 has been removed because they are Issues and not about Big Data.
- 9V: 0.
  - 1 has been removed because it is about Big Data but not about the 'Vs'.

Table 2. Number of articles about Big Data and the 'Vs' in different databases.

Database\Vs	3Vs	4Vs	5Vs	6Vs	7Vs	8Vs	9Vs	10Vs	Total
IEEE	31	29	21	1	3	1	0	1	87
SD	4	6	2	0	0	0	0	0	12
Wiley	3	1	1	1	0	0	0	0	6
Total	38	36	24	2	3	1	0	0	105

The majority of all the articles have been published in conferences, with a total of 75 in IEEExplorer. The other types of articles have been 5 Magazines in IEEExplorer, 20 articles in journals among all the databases and without distinguishing between review articles or research articles, and 4 books (1 in SD and 3 in Wiley). Table 3 shows this information in more detail.

Table 3. Type of article in each database according to the Vs they use.

Type\Vs	3Vs	4Vs	5Vs	6Vs	7Vs	8Vs	9Vs	10Vs	Total
				IEEE					
Conferences	27	25	18		3	1		1	75
Magazines	3	1		1					5
Articles	1	2	3						6
Books		1							1
				SD					
Articles	3	6	2						11
Books	1								1
				Wiley					
Articles	1	1	1						3
Books	2			1					3
Total	38	36	24	2	3	1	0	1	105

3.3. Results and Discussion

The difference between the data analysis that was traditionally done and the term Big Data is that the latter must meet certain conditions that were not 'so important' before. Big Data must comply with the '3Vs' [1,107,118]. These 'Vs' are the volume and amount of data to manage, the velocity of creation and use required to process this data and the variety of different types of data sources. These '3Vs' were first defined in [119], according to [8], or even since 1997 according to [120].

There are many authors who have defended the '3Vs' [108,109], however, sometimes some authors point to the '4Vs' because of the problems that explains [121,122], being this the veracity of these data, while others advocate for the '4Vs', including in them the value [2], and others by the same '4Vs' but adding variability [102], although some authors also speak of veracity [123]. On the other hand, other authors maintain that two extras must be included in those of the first author, which is different from those who advocate the '4Vs' [105,124].

One of the authors of the '5Vs' is [8], thus adding variability and value, while other authors who speak of '5Vs' add veracity and value [22,125,126].

Others speak of '6Vs', leaving out variability [127], and naming the new one Visibility without explaining it. Other authors add vulnerability instead of variability [128], which is very important in our lives, but Big Data depends on the type of application and how people share and use that data.

Later, several authors added a seventh V, that is, visualisation [129]. With 7 'Vs' too, Khan et al. [130] and Gupta et al. [131] propose the use of volume, velocity, variety, veracity, value, validity and volatility. Respecting our proposal, validity is a part of veracity, and the volatility of Khan is a part of variability. The volatility of Gupta depends on the problem; it could be inside velocity because it has to be quickly stored and/or processed, or inside the veracity if it can change and be invalid. Another case with 7 'Vs' proposes the classic 3 'Vs', and adds variability, veracity, visualisation, and value [131].

Fatima et al. [132] introduce in their 8 'Vs' two new ones: Viscosity, and virality. According to our 'Vs', viscosity is inside the variety, and virality is discarded because it can happen with just one tweet or video. However, it is true that sometimes virality can happen with a large dataset of passwords, private data, or any other interesting thing for people.

Other authors add a new 6 'Vs' (Vincularity, Validity, Value, Volatility, Valence, and Vitality) to the classic 4 'Vs' (Volume, Variety, Velocity, and Veracity), having a total of 10. In this case, we propose 5 new ones because Value appears in other articles. However, they just test the 4 'Vs', postponing the explanations and testing of the new 6 'Vs' for future work. In our case, Validity is included in a subtype of Veracity, and according to other articles and its meaning, Volatility would fit inside Variability.

However, as analysed in [133], some Vs are very little used by the authors, since the volume is the most common in 39.64% of the articles they investigated and the variability is only present in 1.8%.

Big Data has continued to evolve since the '3Vs' were first established and new features have been added, which already existed but have not been necessary until more research was done on Big Data. For the same reason, under the study of the different authors and the current literature, and based on what Big Data needs, it arises the combination of them that grants a total of '7Vs' that allow defining both features and needs, and some of the most important challenges of Big Data, as well as its subtypes. The latter case is that of velocity, veracity, and variability, which have different subtypes.

Table 4 summaries the different 'Vs' found in the literature according to this study but does not include all the found articles. A total of 16 'Vs'.

Table 4. Vs literature review.

Authors\Vs	Volume	Velocity	Variety	Veracity	Variability	Value	Visibility	Vulnerability	Visualisation	Validity	Volatility	Viscosity	Virality	Vincularity	Valence	Vitality
[1,107–109,118,119]	X	X	X													
[123]	X	X	X	X												
[2]	X	X	X			X										
[102]	X	X	X		X											
[8]	X	X	X		X	X										
[22,125,126]	X	X	X													
[127]	X	X	X	X		X	X									
[128]	X	X	X	X		X		X								
[129,134]	X	X	X	X	X	X			X							
[130,131]	X	X	X	X		X				X	X					
[128]	X	X	X	X	X	X			X							
[132]	X	X	X	X		X			X			X	X			
[135]	X	X	X	X		X				X	X			X	X	X

4. Challenges According to the '7Vs' of Big Data

In this section, we describe '7Vs' that we have found and are the most used and more related to general data used in Big Data, including an explanation and challenges that exist according to them. Some of them have different subtypes according to the differences detected in the literature and working with them. Some authors create a different V for similar purposes, and here, we have mixed them into one due to the similarities.

4.1. Volume

The volume or quantity of data is enormous and gigantic. This feature describes exactly that: the large amount of data that is coming, the large amount that is working daily ranging from gigabytes to exabytes and more. Storage capacity has been doubling approximately every 3 years, yet in many fields this has been insufficient, as has been the case in medical and financial data [99].

Initially, these data came from databases. However, after the 'boom' in improved technologies, more sites are now taken into account, such as the Internet, smart objects and sensors and actuators [136], applications, and even photos and videos, or the Internet of Things (IoT) [93], or Online Social Networks. In addition, some researchers should deal with datasets thousands of times larger than those generated only a decade ago, such as satellite procedures, telescopes, high-performance instruments, sensor networks, particle accelerators and supercomputers [137]. This also makes us human beings into walking data generators [1], since we are the ones who continuously generate many of these data with the use of mobile devices, among others.

Many companies currently try thousands of terabytes (10^{12}) daily to obtain useful information for their businesses and about their users. The amount of data is growing because every day there is more information and more existing users, which implies, as some already estimated, that it has exponential growth. It is estimated that there will be 500 times more data in 2020 than in 2011 [102].

As seen in Table 5, there is a lot of data and the way to analyse them is using Big Data tools to manage this knowledge in real or almost real time. These data demonstrate the rise of Big Data applications, where data collection has grown enormously and is beyond the capacity of the software tools traditionally used to capture, manage and process these data within a "tolerable time" [100].

Table 5. Big Data statistics.

Type	Quantity	Commentary/Year
Earth Satellites [7]	1 terabyte (10^{12})	In 1 day in 1990
Websites indexed by Google [8]	1 million	1998
Websites indexed by Google [8]	1000 million	2000
Computing [138]	320 terabytes	2 h of human genome study in 2008
Websites indexed by Google [8]	1 trillion (10^{18})	2008
Astronomical or physical particle experiment [137]	1 petabyte	In 1 year in 2009
Facebook [15]	30 billions of content shared each month	2010
Photos per second on Facebook [139]	1 million	2010
Photos stored on Facebook [139]	260 billions = 20 petabytes	2010
Photos uploaded per week on Facebook [139]	1 billion = 60 terabytes	2010
Bookstore of the United States of America Congress [15]	235 terabytes of data collected	April 2011
Hadron Collider at the discovery of the Higgs Boson [97]	1 petabyte (10^{15})	Per second in 2012
Human race [11]	2.5 quintillions (10^{30}) of data bytes	Every day in 2012
Walmart user information every day [1]	2.5 petabytes (10^{15})	2012
Multi-Media Messages (MMS) [140]	28,000 per second	2012
New data [1]	2.5 exabytes	New data every day since 2012 and doubling every 40 months
Electronic data [16]	1.2 zettabytes	Every year in 2012
Twitter [141]	10,300,000 tweets in 1 h 30 m	Presidential debate in 2012
GitHub [142]	550,000 repositories	Q2 2012
Creators of social content [143]	600 million (10^{6})	33% of Internet users in 2013
Other periodical publications [143]	10,000	Newspapers and others in 2013
Blogs [143]	70 million (10^{6})	2013
Google queries per day [8]	More than 1000 million	2013
Tweets per day [8]	+250 million	2013
Facebook updates per day [8]	+800 million	2013
YouTube views per day [8]	+4000 million	2013
Jet engine [2]	10 terabytes	30 min in 2013
Internet [143]	20 exabytes (10^{18}) of information	2013
Internet [144]	40.7% of the population used it in 2014 = 2.954 million	7,259,691,769 people in 2014
Web pages [143]	1.5 trillion (10^{12})	2013
Twitter [145]	310 million active monthly users	2013
Twitter [145]	500 million tweets per day	August 2013
Tweets [143]	20 thousand million (10^{9})	50 million of users/2013
Tweets [145]	143,199 per second	3 August 2013
GitHub [142]	1,300,000 repositories	Q4 2013
GitHub [142]	2,200,000 repositories	Q4 2014
Sequencing of human gene [103]	600 Gb	2014
Flickr [146]	Almost 70 million public photos uploaded monthly	2015
YouTube [147]	More than 1000 million users (10^{9}) = 1/3 Internet users	2015

YouTube [147]	+100 million hours of video views daily	2015
Hospital data [103]	167 Tb to 665 Tb	2015
Emails [148]	204 million	In 1 min in 2016
Pandora: hours of music heard [148]	61,000 h	In 1 min in 2016
Flickr [148]	3 million uploads	In 1 min in 2016
Flickr [148]	20 million photos viewed	In 1 min in 2016
Google [148]	2 million searches	In 1 min in 2016
Google Photos [149]	200 million users	In its first year in 2016
Google Photos [149]	1.6 billion (10^9)	In its first year in 2016
Google Photos [149]	2 trillion (10^{18}) tags	In its first year in 2016
Google Photos [149]	24 billion (10^9) selfies	In its first year in 2016
Facebook [150]	1650 million (10^6) users	31 March 2016
Annual Internet traffic [151]	1 zettabyte (10^{18})	2016
Facebook [133]	+500 terabytes of data per day	2017
GitHub [152]	100,000,000 repositories	2018
ELMo [153]	94 million of parameters	2018
BERT-Large	340 million of parameters	2018
GPT [154]	110 million of parameters	2018
GPT-2 [153]	1.5 billion of parameters	2019
Megatron-LM [153]	8.3 billion of parameters	2019
T5 [153]	11 billion of parameters	2019
Annual Internet traffic [151]	2.3 zettabytes (10^{18})	2020
Square Kilometre Array [155]	524 terabytes per second (estimated)	Will be produced in 2020 (postponed to 2027)
Turing-NLG [153]	17.2 billion of parameters	2020
GTP-3 [153]	175 billion of parameters	2020
Daily generated data [12]	56 zettabytes	16 December 2020
Megatron-Turing [153]	15 datasets of a total of 339 billion tokens	2021
Megatron-Turing [153]	530 billion of parameters	2021
Daily generated data [12]	Estimated 149 zettabytes	2024

Often this knowledge must be efficient as it needs to be real time due to problems with the space to store it [100]. An example of this is the Square Kilometre Array (SKA) radio telescope [155], which will become the largest radio telescope in the world and aims to discover the beginning and end of the universe. Researchers estimate that SKA will produce approximately 4.1 pebibits (2^{50}) per second or 562 terabytes (10^{12}) per second. As can be seen, there will be a future with even more data, with huge amounts to analyse. Some estimates indicate that the amount of new information will double every three years [18].

Another example of the amount of data generated occurred on 4 October 2012, when the presidential debate between the President of the United States of America, Barack Obama, and Governor Mitt Romney had Twitter users post more than ten million tweets in an hour and a half [141]. Studying this Twitter data about the US presidency debate [141], it can be observed that people tweeted more when talking about the health insurance of older people. Thus, based on this, it can be determined what mattered to people at that time.

Flickr, a social network of photos, is also widely used in investigations because of the information that can be obtained from its images. About 70 million new public photos are currently uploaded monthly [146]. Thus, thanks to the analysis of these photos, they could be used to study human society, social events or even disasters [100].

On the other hand, the scientific advisor to the President of the United States of America, John Paul Holdren, said in 2011 that Big Data was an important issue because every year about 1.2 zettabytes (10^{21}) of electronic data are created. The equivalent in terabytes is 1,200,000,000, ranging from scientific experiments to telescope data and tweets [16]. This is certified by other estimates made in 2012 [1], where they predicted the creation of 2.5 exabytes (10^{18}) each day, equivalent to 2,500,000 terabytes, but that this creation capacity would double every 40 months, approximately. Thus, this prediction is quite similar to that made by John Paul Holdren.

However, this amount of data is not currently being created because, years ago, there were already certain applications that generated large amounts of data. In 1990, satellites orbiting Earth generated one terabyte (10^{15}) of information per day. This meant that if a photo were taken every second and a person was asked to analyse all these photos, assuming he worked at night and on weekends, it would take him many years to analyse all the photos of a day [7]. This is useful to emphasise that now, 26 years later and with improved technologies, both hardware and software, we can make faster and automatic analysis, also on images with better resolution and more data.

Another relevant project, both in terms of importance and data size, is the project 'The Genome 1000 Project', which deals with the human genome. In this project, two hours of Solexa execution created 320 terabytes of information. This made it impossible to save and compute in 2008 [138].

In relation to this evolution, it can be seen the prediction of Eron Kelly, who predicted that in the next five years, we will generate more data than all those generated by humanity in the last 5000 years [101]. Meanwhile, the NIST expects data generation to double every two years, reaching 40,000 exabytes by 2020, of which one-third is expected to be useful if analysed. This is something that highlights the evolution of the data humans generate and its exponential growth throughout history, as well as what is expected to happen.

To handle all this information, some years ago the different systems have started to be migrated to the Cloud and perform what is known as Cloud Computing, whether in a private, public or hybrid system. This resulted in a saving of money and a new way of processing data, which facilitated the use of Big Data in companies thanks to the different technologies provided by companies and the scaling of these tools [103].

However, it should be kept in mind that not always because of having larger datasets will one get better predictions; this can be classified as arrogance, because Big Data is not the substitute for data collection and Data Mining [156]. Therefore, despite having these large sets, one should not forget the basic techniques of measurement and construction of reliable and valid data and the dependencies between these data [157].

4.2. Velocity: Reading and Processing

Velocity describes how fast data is processed, because in Big Data, the creation of this data is continuous; it never ceases. On the Internet, whether through an Online Social Network such as Twitter, or on Facebook, or by different services such as blogs or video services, there are people at all times writing information, uploading a video, sending emails or accessing web pages. This happens all the time, thousands of datasets every minute. This makes for a great velocity of content creation or reading.

The vast majority of these services require data from their users, how they use their service or their preferences in order to adapt their content or ads to their users. This creates a great need for data processing velocity. This velocity can be of four types: batch or interval, near time or nearly time, which is almost real time, real-time, and streaming.

As can be seen, this "V" has two types of velocities, the one of reading or content creation and the one of processing, which can be independent or dependent, according to the requirements of the application.

In addition, there are applications that need more processing velocity than data volume and thus allow a company to be much more agile than its competitors by offering

real-time or near-real-time applications [1,102]. It is important, at this point, to bear in mind that this does not refer to bandwidth or protocol issues, but to the velocity of content creation and the manageability of these, which is divided into their storage, analysis and visualisation, hoping that with the use of Big Data this time will be minimum [102,119].

An example of this "V" is the competition that took place at the SC08 International Conference for High Performance Computing congress in 2008 [158]. Participants had to consult with the Sloan Digital Sky Survey (SDSS). The SDSS is a space research project in which three-dimensional images of space are taken in order to map it [159]. The winner took 12 min to make a query in a parallel cluster, while the same query without using parallelism took 13 days [137]. This highlights the importance of processing velocity when computing large amounts of data.

4.3. Variety

Variety describes the organisation of data, whether structured, semi-structured, unstructured or mixed. Currently, there are countless possible sources of data given by the wide variety of existing sources for collecting them, being in many cases unstructured, difficult to handle and very noisy data [1]. We are talking about millions of sensors around the world, tens of social networks in which millions of messages are published daily and hundreds of different formats ranging from plain text to images.

There are websites where users write their opinions: Twitter, Facebook, LinkedIn, Instagram, YouTube, Tumblr, Flickr and other social networks, which are considered the most valuable resources [17]. Another very important variety of data is that offered by different governments when they provide different open data, as these data are offered in different formats such as Excel, CSV, Word, PDF, JSON, RDF-N3, RDF-Turtle, RDF-XML, XML, XHTML, plain text, HTML, RSS, KML and GeoRSS. Other types of data that are interesting are videos, in different formats such as FLV, WMV, AVI, MKV or MOV, which are also often accompanied by comments; in this case, we find services such as YouTube and Vimeo, among others. A similar case is that of audio services or radios, which have different formats, such as MP3, OGG, FLAC, MIDI or WAV.

The latest technologies, such as different mobile devices, televisions, cars, tablets, smartphones or any other Smart Object [136], offer many data through the different means they have, either through their sensors or their GPS systems. Mentioning sensors, they are all available in the market and they are opening up more and more thanks to their ease of use through an Arduino or a Raspberry Pi. To this must be added other IoT-supporting devices [8], which are one of the cornerstones as a source of information, both structured and semi-structured or unstructured, from Big Data [103].

It should not be forgotten other very important data to measure and used to watch user interaction, such as mouse clicks, keystrokes, page scrolling, reading time in an article, shared content, or interactions with content, as many social networks do, such as Facebook, Twitter, and Weibo. All this is often accompanied by photos, which further expand the multitude of formats and treatments with JPG, PNG, TIFF, or BMP.

Moreover, it is sometimes necessary to deal with legacy data in different databases, whether SQL or NoSQL, documents, emails, telephone conversations or scanned documents [106]. Other times, it may be information from web pages that are in HTML or well-structured XMLs or PDFs that do not have a structured way of displaying data. At other times, there are different data groups of different types in compressed files in RAR, RAR4, ZIP, 7Z, or TAR, which must first be tried to decompress and analyse its contents.

As can be seen, there are many formats, and here there are only a few examples of the most used ones. Each of these files also needs special treatment, even if they are of the same type. For example, in the case of images, not all formats have the same properties and small differences. In addition, every year, we have new devices or services that provide new useful data or the same data, but in other formats, or modifying how that information was shown, which implies modifying certain parts of the data-reading software. It must also be borne in mind that, in addition to the formats, all this also depends on the

application, since it is not the same to analyse telescope or satellite images as social images or to analyse user data with time data.

Thus, this heterogeneity of data, incompatible formats and inconsistent data is one of the greatest barriers to effective data management [119]. In addition, these data are constantly changing and new systems are being added that either require or modify the way in which existing information is provided to the end user, contrary to the way it was done in the past, where there was only one structured database with a well-defined and slowly changing schema [2]. Some authors call it viscosity due to the fact that we have to use data from different sources, and sometimes it requires a transformation to use it [132]. This one is another of the problems; not every data we need has the same structure or format file if we require it for different sources, such as different governments, enterprises or portals. Then, we have to create different parsers and translate all the information into one common format. Other times, some of this information can be in a hard format to use, such as PDF or HTML, and the transformation is required to work easier with it.

4.4. Veracity: Origin, Veracity and Validity

Veracity is given because not everything that exists is true and false or erroneous data may be being collected. Therefore, when working, one should be careful with data sources and check that they are true data, thus trying to obtain accurate and complete information. This, in turn, can be divided into three sub-sections, namely, origin, veracity and validity.

Data origin is important to maintain quality, protect security and maintain privacy policies. It should be borne in mind that the data in Big Data move from the individual limits to those of groups or communities of interest and that these range from a regional or national limit to the international one. For this reason, the source helps to know where these data come from and what their original source is, for instance, by using metadata. This is very important because, knowing its origin, you can maintain the privacy of this data and thus be able to make some important decisions, such as the right to be forgotten. Other data to be inserted could be supply chain-related, such as calibration, errors, or missing data (timestamps, location, equipment serial number, transaction number, and authority).

Veracity includes the guarantee of the information of the means used to collect the information. A clear example is if sensors were used, which should include traceability, calibration, version, sampling time and device configuration. The reason for this can be a malfunctioned or uncalibrated device [122]. On the other hand, Online Social Networks give to us the opinion of the users, but maybe we cannot trust it [130].

Another example of a lack of or problems with veracity is the one mentioned in [160]. In this article, it is mentioned that, as a preliminary study to verify a certain assumption in 2006, the author used Google Trends to find out if the president who had been elected to the Real Madrid Football Club was the president with most queries on Google search engines. As the author points out, the surname of this president was Calderón. The problem, as he stated, is that on the same day, a president with the same surname was elected in Mexico. Thus, the problem was that Google Trends did not differentiate what Calderón was meant by each query; thus, it merged them. In other words, there was a lack of complementary data, a lack of veracity.

Validity refers to the accuracy and precision of the data, or rather the quality of the data. Examples can be given if, in continuous and discreet data on the gender of people and being male = 1 and female = 2, a 1.5 is received because this does not mean a new gender, but an error. Another type of error is the fraud of clicks on pages, clicks made by robots, hidden ads, etc.

An example of veracity is what the author of this article commented [121]; in it, he gave an example of several problems that had occurred in the United States of America. The first one is that politicians always like to talk about more data, but ultimately these are never among their priorities, as this has to compete with other things that offer a more immediate impact, which makes the money insufficient to keep the data accessible and

digestible. The second problem is that data are institutionalised when they should be isolated from politicians, for example, with the alleged creation of the Environmental Statistics Agency (BES), which, in addition to collecting data, would analyse them. The third and final problem arises when chemical companies refuse to present their pollution data based on else; they reveal much useful data for their competitors. As stated in the article by David Goldton, it can be deduced that these data may not be easily accessible, outdated, incorrect, biased, or difficult to understand.

Another example, although some add it as part of the Variety and noise type, is the case of the Fukushima nuclear power plant, when people began to write about it and to share data from poorly calibrated sensors and/or that were not exactly in the said area [133]. In this case, it is a problem of Veracity due to the origin because of the lack of calibration and error, of veracity for being incorrect data and of validity for not being precise.

Clearly, if a large dataset is available, for instance, to see a tendency and there are few "bad" data, these will be lost and automatically ignored, thanks to the fact that they will be hidden among "good" data. However, if it is something casual, maybe this "bad" data can spoil the experiment, which makes the veracity of these data extremely important [102].

4.5. Variability: Structure, Time Access and Format

Variability is due to changes that the data have over time.

Maybe the data structure and how users want to interpret that data can change with time, modifying its structure, which would modify the way of parsing the data tree, perhaps by modifying the model in XML.

Other times, the time to access this data is different because they take more time to create the data or to update it, not always updating the information constantly. For instance, they do not update the information of that file more than once, as happens with some files from some Open Data portals. Another example is when they can delete the data, such as some companies or governments, when they are not obliged to keep the data for more than one year [130].

Another possible problem is when they change the format in which they are offered by migrating from one format, such as XML, to another, such as JSON, or the composition of these, adding or removing internal elements of their structure.

Because of these problems, one has to be aware of these changes whenever one wants to update them, but the original data should also be kept unchanged, that is, to know how data were changing over time. This has the drawback of data redundancy and the necessary storage space. In addition, of course, the necessary system for monitoring changes and comparing them with existing data already stored in our system.

This point is very important because, having tools that process this data, it will be necessary to adapt them over time, as well as they should be able to detect new changes in the file; thus, a human being makes decisions about those changes: to add them, modify the programme, to avoid them, etc. Or maybe we have to change the source of data if they stop updating it or just delete it.

4.6. Value

The value of the data lies in what they can bring to the company and whether they can provide an advantage, as they help to make decisions based on questions that can now be answered thanks to the analysis of this data or in the discovery of trends. However, this value is variable because it depends on the data available, the hidden information they contain and whether it is found and extracted. Therefore, the challenge is to identify this value and extract it to perform an analysis of the data provided to us and to find this value. According to the survey analysed in [3], which corresponds to that carried out by MIT and IBM to 3000 executives, analysts and managers of more than 30 companies in 100 countries, one in five said they were concerned about data quality or ineffectiveness of government data as a major concern.

Thus, if these data with which one works have no value or have insufficient value for the company, project, or research, they will create a monetary loss because of storage, processing and human resources, as well as time. This is why, probably, it is the most important V, according to [130].

4.7. Visualisation

This V focuses on representing the data obtained in something that is readable, according to [133]. When dealing with so much data, many companies have had to hire people who are dedicated only to these visualisations; thus, they offer added and visual value to their employees. In addition, they created new tools for viewing these that worked in a correct and fast way. Moreover, due to the large size of the data to work, and sometimes at velocity, it becomes very difficult to create visualisation because the current tools have poor performance in terms of functions, scalability, and response time [8,22,99]. This is especially complicated when making real-time applications. Other times, with so much data available, the visualisation can happen to be difficult to understand.

An example of visualisation is the creation of Hive by Facebook [161], although now it is the one from the Apache Foundation, and it allows SQL-style queries to be performed using a command line. On the other hand, we have Hue [162] that offers a graphical interface that gives Hive support as well as to other tools of the Hadoop ecosystem, in addition to providing graphics, monitoring and management.

Another case is that of eBay, where its employees can see the relevant data of users of the platform and thus be able to perform sentiment analysis on this data [133].

Then, we see that this V is important because it represents the challenges of visualising useful information for a user or company in a clear and fast way that allows the visualisation or decision making of the processed data.

5. Conclusions and Future Work

In this article, we have introduced the importance of Big Data from different points of view and studied the literature to show the differences and similarities between Data Mining, KDD and Big Data, given a better perspective about what Big Data is and is not, and their properties and challenges.

First, we have discussed the differences between Data Analysis, Data Mining, Knowledge Discovery in Databases, and Big Data. With this discussion, we have researched in the literature the use, meaning, definitions and novel applications of each one to show the purpose of them and avoid misunderstandings in the literature. According to this study, Data Analysis and Data Mining are different things. Moreover, Data Mining can be used as the main part of KDD. Regarding KDD, in this survey, we studied the phases of KDD and improved the information of the 8-phase method. On the other hand, Big Data can be used to improve and do these three processes because of the tools and methodologies it provides, showing the challenges that affect some current applications.

Second, we have reviewed the literature about the 'Vs' in three electronic databases. This literature review shows that the most used database to describe or discuss the characteristics or explain articles that are focused on them is IEEExplore, usually in conferences. In addition, they have proposed a total of 16 different 'Vs' but some of them can be merged and others are not explained. According to this study's challenges, we detected a total of 7Vs and their subproperties: volume, velocity (lecture and processing), variety, veracity (origin, veracity, and validity), variability (structure, time access, and format), value and visualisation. These 7Vs are properties of Big Data, but they are challenges too that people who work with Big Data and/or research in Big Data have to know and take into account, as we show in the Big Data Applications section.

In this survey, we reviewed relevant articles about these technologies along history and some novel applications. From this point, more surveys and Systematic Literature Review (SLR) can be useful. For instance, about the use of the different terms along history to know if they are mixed in some specific case, the type of applications in Big Data that

use Data Mining or KDD and the most used algorithms in them. According to the 7Vs, a possible future work would be an SLR about other challenges, including security, ethics, privacy, standards, energy consumption, architectures, abstraction and used algorithms and patterns. Moreover, an SLR about the Big Data application and the 'Vs' used in them based on this survey will be useful to better detect the most relevant ones and the reasons why, and an SLR about the different phases and uses of KDD too.

Author Contributions: Conceptualisation, C.G.G.; methodology, C.G.G.; validation, C.G.G. and E.Á.-F.; investigation, C.G.G.; writing—original draft preparation, C.G.G. and E.Á.-F.; writing—review and editing, C.G.G. and E.Á.-F.; visualisation, C.G.G. and E.Á.-F.; supervision, C.G.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McAfee, A.; Brynjolfsson, E. Big data: The Management Revolution. *Harv. Bus. Rev.* **2012**, *90*, 60–68.
2. Dijcks, J.-P. *Oracle: Big Data for the Enterprise*; Oracle: Redwood, CA, USA, 2013.
3. Lavalle, S.; Lesser, E.; Shockley, R.; Hopkins, M.S.; Kruschwitz, N. Big Data, Analytics and the Path from Insights to Value. *Winter* **2011**, *52*, 21–31.
4. Chen, H.; Chiang, R.H.L.; Storey, V.C. Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Q.* **2012**, *36*, 1165–1188.
5. Menzies, T.; Hu, Y. Data mining for very busy people. *Computer* **2003**, *36*, 22–29.
6. Rokach, L.; Maimom, O. *Data Mining with Decision Trees: Theory and Applications*; World Scientific Publishing Co. Pte Ltd: Danvers, MA, USA, 2007; ISBN 9789812771711.
7. Frawley, W.J.; Piatetsky-Shapiro, G.; Matheus, C.J. Knowledge Discovery in Databases: An Overview. *AI Mag.* **1992**, *13*, 57–70.
8. Fan, W.; Bifet, A. Mining Big Data: Current Status, and Forecast to the Future. *ACM SIGKDD Explor. Newsl.* **2013**, *14*, 1.
9. Letouzé, E. Big Data for Development: Challenges & Opportunities. 2012. Available online: https://unstats.un.org/unsd/trade/events/2014/beijing/documents/globalpulse/Big%20Data%20for%20Development%20-%20UN%20Global%20Pulse%20-%20June2012.pdf (accessed on 27 October 2022).
10. Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques*; 3rd ed.; Morgan Kaufmann: Burlington, MA, USA, 2007; ISBN 9780123748560.
11. Cloud Security Alliance. Top Ten Big Data Security and Privacy Challenges. 2012. Available online: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Top_Ten_v1.pdf (accessed on 27 October 2022).
12. Nti, I.K.; Quarcoo, J.A.; Aning, J.; Fosu, G.K. A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. *Big Data Min. Anal.* **2022**, *5*, 81–97.
13. The Apache Software Foundation. Apache™ Hadoop®. Available online: http://hadoop.apache.org/ (accessed on 27 October 2022).
14. Ahrens, J.; Hendrickson, B.; Long, G.; Miller, S.; Ross, R.; Williams, D. Data-Intensive Science in the US DOE: Case Studies and Future Challenges. *Comput. Sci. Eng.* **2011**, *13*, 14–24.
15. Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C.; Byers, A.H. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*; McKinsey Global Institute: Washington, DC, USA, 2011.
16. Mervis, J. Agencies Rally to Tackle Big Data. *Science* **2012**, *336*, 22.
17. Bello-Orgaz, G.; Jung, J.J.; Camacho, D. Social big data: Recent achievements and new challenges. *Inf. Fusion* **2016**, *28*, 45–59.
18. Greiner, L. What is Data Analysis and Data Mining? Available online: https://www.dbta.com/Editorial/Trends-and-Applications/What-is-Data-Analysis-and-Data-Mining-73503.aspx (accessed on 27 October 2022).
19. Friedman, J.H. Data Mining and Statistics: What's the connection? *Comput. Sci. Stat.* **1998**, *29*, 3–9.
20. Manaris, B. Natural Language Processing: A Human-Computer Interaction Perspective. In *Advances in Computers*; Elsevier: Amsterdam, The Netherlands, 1998; Volume 47, pp. 1–66, ISBN 9780120121472.
21. Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. The KDD process for extracting useful knowledge from volumes of data. *Commun. ACM* **1996**, *39*, 27–34.
22. Assunção, M.D.; Calheiros, R.N.; Bianchi, S.; Netto, M.A.S.; Buyya, R. Big Data computing and clouds: Trends and future directions. *J. Parallel Distrib. Comput.* **2015**, *79–80*, 3–15.

23. Leskovec, J.; Rajaraman, A.; Ullman, J.D. *Mining of Massive Datasets*; Cambridge University Press: Cambridge, UK, 2014; ISBN 9781139058452.
24. Labrinidis, A.; Jagadish, H.V. Challenges and opportunities with big data. *Proc. VLDB Endow.* **2012**, *5*, 2032–2033.
25. Piatetsky-Shapiro, G. From Data Mining to Big Data and Beyond. Available online: https://www.kdnuggets.com/2012/04/from-data-mining-to-big-data-and-beyond.html (accessed on 27 October 2022).
26. Fayyd, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. From Data Mining to Knowledge Discovery in Databases. *AI Mag.* **1996**, *17*, 37–54.
27. Ha, S.H.; Park, S.C. Application of data mining tools to hotel data mart on the Intranet for database marketing. *Expert Syst. Appl.* **1998**, *15*, 1–31.
28. Buxton, B.; Hayward, V.; Pearson, I.; Kärkkäinen, L.; Greiner, H.; Dyson, E.; Ito, J.; Chung, A.; Kelly, K.; Schillace, S. Big data: The next Google. *Nature* **2008**, *455*, 8–9.
29. NIST Big Data Public Working Group: Reference Architecture Subgroup. *NIST Big Data Interoperability Framework: Volume 5, Architectures White Paper Survey*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2015; Volume 5.
30. Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining*; The MIT Press: Cambridge, MA, USA, 1996; ISBN 9780262560979.
31. Data Mining Algorithms (Analysis Services—Data Mining). Available online: https://msdn.microsoft.com/en-us/library/ms175595.aspx (accessed on 27 October 2022).
32. Hand, D.J. *Discrimination and Classification*; John Wiley and Sons Inc.: New York, NY, USA, 1981; Volume 1, ISBN 9780471280484.
33. González García, C.; Núñez-Valdez, E.R.; García-Díaz, V.; Pelayo G-Bustelo, C.; Cueva Lovelle, J.M. A Review of Artificial Intelligence in the Internet of Things. *Int. J. Interact. Multimed. Artif. Intell.* **2019**, *5*, 9–20.
34. Wang, M.; Sheng, L.; Zhou, D.; Chen, M. A Feature Weighted Mixed Naive Bayes Model for Monitoring Anomalies in the Fan System of a Thermal Power Plant. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 719–727.
35. He, W.; He, Y.; Li, B.; Zhang, C. A Naive-Bayes-Based Fault Diagnosis Approach for Analog Circuit by Using Image-Oriented Feature Extraction and Selection Technique. *IEEE Access* **2020**, *8*, 5065–5079.
36. Xue, Z.; Wei, J.; Guo, W. A Real-Time Naive Bayes Classifier Accelerator on FPGA. *IEEE Access* **2020**, *8*, 40755–40766.
37. Sanchis, A.; Juan, A.; Vidal, E. A Word-Based Naïve Bayes Classifier for Confidence Estimation in Speech Recognition. *IEEE Trans. Audio. Speech. Lang. Process.* **2011**, *20*, 565–574.
38. Shirakawa, M.; Nakayama, K.; Hara, T.; Nishio, S. Wikipedia-Based Semantic Similarity Measurements for Noisy Short Texts Using Extended Naive Bayes. *IEEE Trans. Emerg. Top. Comput.* **2015**, *3*, 205–219.
39. Kustanto, N.S.; Nurma Yulita, I.; Sarathan, I. Sentiment Analysis of Indonesia's National Health Insurance Mobile Application using Naïve Bayes Algorithm. In Proceedings of the 2021 International Conference on Artificial Intelligence and Big Data Analytics, Bandung, Indonesia, 27–29 October 2021; pp. 38–42.
40. Castro, W.; De-la-Torre, M.; Avila-George, H.; Torres-Jimenez, J.; Guivin, A.; Acevedo-Juárez, B. Amazonian cacao-clone nibs discrimination using NIR spectroscopy coupled to naïve Bayes classifier and a new waveband selection approach. *Spectrochim. Acta—Part A Mol. Biomol. Spectrosc.* **2022**, *270*, 120815.
41. Yoshikawa, H. Can naive Bayes classifier predict infection in a close contact of COVID-19? A comparative test for predictability of the predictive model and healthcare workers in Japan. *J. Infect. Chemother.* **2022**, *28*, 774–779.
42. Bhatia, S.; Malhotra, J. Naïve Bayes Classifier for Predicting the Novel Coronavirus. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), Tirunelveli, India, 4–6 February 2021; pp. 880–883.
43. Shanbehzadeh, M.; Nopour, R.; Kazemi-Arpanahi, H. Using decision tree algorithms for estimating ICU admission of COVID-19 patients. *Inform. Med. Unlocked* **2022**, *30*, 100919.
44. Ghane, M.; Ang, M.C.; Nilashi, M.; Sorooshian, S. Enhanced decision tree induction using evolutionary techniques for Parkinson's disease classification. *Biocybern. Biomed. Eng.* **2022**, *42*, 902–920.
45. Elhazmi, A.; Al-Omari, A.; Sallam, H.; Mufti, H.N.; Rabie, A.A.; Alshahrani, M.; Mady, A.; Alghamdi, A.; Altalaq, A.; Azzam, M.H.; et al. Machine learning decision tree algorithm role for predicting mortality in critically ill adult COVID-19 patients admitted to the ICU. *J. Infect. Public Health* **2022**, *15*, 826–834.
46. Hiranuma, M.; Kobayashi, D.; Yokota, K.; Yamamoto, K. Chi-square automatic interaction detector decision tree analysis model: Predicting cefmetazole response in intra-abdominal infection. *J. Infect. Chemother.* **2023**, *29*, 7–14.
47. Alex, S.; Dhanaraj, K.J.; Deepthi, P.P. Private and Energy-Efficient Decision Tree-Based Disease Detection for Resource-Constrained Medical Users in Mobile Healthcare Network. *IEEE Access* **2022**, *10*, 17098–17112.
48. Wang, X.; Liu, F. Data-Driven Relay Selection for Physical-Layer Security: A Decision Tree Approach. *IEEE Access* **2020**, *8*, 12105–12116.
49. Kuang, W.; Chan, Y.-L.; Tsang, S.-H.; Siu, W.-C. Machine Learning-Based Fast Intra Mode Decision for HEVC Screen Content Coding via Decision Trees. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 1481–1496.
50. Chen, Y.; Mao, Q.; Wang, B.; Duan, P.; Zhang, B.; Hong, Z. Privacy-Preserving Multi-Class Support Vector Machine Model on Medical Diagnosis. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 3342–3353.
51. Lei, H.; Guoxing, Y.; Chao, H. A sparse algorithm for adaptive pruning least square support vector regression machine based on global representative point ranking. *J. Syst. Eng. Electron.* **2021**, *32*, 151–162.

52. Astuti, S.D.; Tamimi, M.H.; Pradhana, A.A.S.; Alamsyah, K.A.; Purnobasuki, H.; Khasanah, M.; Susilo, Y.; Triyana, K.; Kashif, M.; Syahrom, A. Gas sensor array to classify the chicken meat with E. coli contaminant by using random forest and support vector machine. *Biosens. Bioelectron. X* **2021**, *9*, 100083.
53. Pang, J.; Pu, X.; Li, C. A Hybrid Algorithm Incorporating Vector Quantization and One-Class Support Vector Machine for Industrial Anomaly Detection. *IEEE Trans. Ind. Inform.* **2022**, *18*, 8786–8796.
54. Bernardini, M.; Romeo, L.; Misericordia, P.; Frontoni, E. Discovering the Type 2 Diabetes in Electronic Health Records Using the Sparse Balanced Support Vector Machine. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 235–246.
55. Ali Hammouri, Z.A.; Delgado, M.F.; Cernadas, E.; Barro, S. Fast SVC for large-scale classification problems. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 1.
56. Azgomi, H.; Haredasht, F.R.; Safari Motlagh, M.R. Diagnosis of some apple fruit diseases by using image processing and artificial neural network. *Food Control* **2023**, *145*, 109484.
57. Zhu, H.; Jiao, L.; Ma, W.; Liu, F.; Zhao, W. A Novel Neural Network for Remote Sensing Image Matching. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2853–2865.
58. Qin, C.; Schlemper, J.; Caballero, J.; Price, A.N.; Hajnal, J.V.; Rueckert, D. Convolutional Recurrent Neural Networks for Dynamic MR Image Reconstruction. *IEEE Trans. Med. Imaging* **2019**, *38*, 280–290.
59. Wu, N.; Phang, J.; Park, J.; Shen, Y.; Huang, Z.; Zorin, M.; Jastrzebski, S.; Fevry, T.; Katsnelson, J.; Kim, E.; et al. Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening. *IEEE Trans. Med. Imaging* **2020**, *39*, 1184–1194.
60. Dong, X.; Zhou, Y.; Wang, L.; Peng, J.; Lou, Y.; Fan, Y. Liver Cancer Detection Using Hybridized Fully Convolutional Neural Network Based on Deep Learning Framework. *IEEE Access* **2020**, *8*, 129889–129898.
61. Ulloa-Cazarez, R.L.; Garcia-Diaz, N.; Soriano-Equigua, L. Multi-layer Adaptive Fuzzy Inference System for Predicting Student Performance in Online Higher Education. *IEEE Lat. Am. Trans.* **2021**, *19*, 98–106.
62. Ibragimov, B.; Toesca, D.A.S.; Yuan, Y.; Koong, A.C.; Chang, D.T.; Xing, L. Neural Networks for Deep Radiotherapy Dose Analysis and Prediction of Liver SBRT Outcomes. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 1821–1833.
63. Haghighat, M.H.; Li, J. Intrusion detection system using voting-based neural network. *Tsinghua Sci. Technol.* **2021**, *26*, 484–495.
64. Wisanwanichthan, T.; Thammawichai, M. A Double-Layered Hybrid Approach for Network Intrusion Detection System Using Combined Naive Bayes and SVM. *IEEE Access* **2021**, *9*, 138432–138450.
65. Gu, J.; Lu, S. An effective intrusion detection approach using SVM with naïve Bayes feature embedding. *Comput. Secur.* **2021**, *103*, 102158.
66. Li, M.; Vanberkel, P.; Zhong, X. Predicting ambulance offload delay using a hybrid decision tree model. *Socioecon. Plann. Sci.* **2022**, *80*, 101146.
67. Feng, X.; Zhou, Y.; Hua, T.; Zou, Y.; Xiao, J. Contact temperature prediction of high voltage switchgear based on multiple linear regression model. In Proceedings of the 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC), Hefei, China, 19–21 May 2017; pp. 277–280.
68. Li, S.; Song, P.; Zhang, W. Transferable discriminant linear regression for cross-corpus speech emotion recognition. *Appl. Acoust.* **2022**, *197*, 108919.
69. Huang, L.; Song, T.; Jiang, T. Linear regression combined KNN algorithm to identify latent defects for imbalance data of ICs. *Microelectron. J.* **2022**, *131*, 105641.
70. Duan, J.; Chang, M.; Chen, X.; Wang, W.; Zuo, H.; Bai, Y.; Chen, B. A combined short-term wind speed forecasting model based on CNN–RNN and linear regression optimization considering error. *Renew. Energy* **2022**, *200*, 788–808.
71. Abbas, S.A.; Aslam, A.; Rehman, A.U.; Abbasi, W.A.; Arif, S.; Kazmi, S.Z.H. K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir. *IEEE Access* **2020**, *8*, 151847–151855.
72. Rong, Y.; Liu, Y. Staged text clustering algorithm based on K-means and hierarchical agglomeration clustering. In Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 27–29 June 2020; pp. 124–127.
73. Jeong, W.; Yu, U. Effects of quadrilateral clustering on complex contagion. *Chaos Solitons Fractals* **2022**, *165*, 112784.
74. Bhagat, H.V.; Singh, M. DPCF: A framework for imputing missing values and clustering data in drug discovery process. *Chemom. Intell. Lab. Syst.* **2022**, *231*, 104686.
75. Tian, Y.; Zheng, R.; Liang, Z.; Li, S.; Wu, F.-X.; Li, M. A data-driven clustering recommendation method for single-cell RNA-sequencing data. *Tsinghua Sci. Technol.* **2021**, *26*, 772–789.
76. Krishnaveni, A.S.; Madhavan, B.L.; Ratnam, M.V. Aerosol classification using fuzzy clustering over a tropical rural site. *Atmos. Res.* **2022**, *282*, 106518.
77. Monshizadeh, M.; Khatri, V.; Kantola, R.; Yan, Z. A deep density based and self-determining clustering approach to label unknown traffic. *J. Netw. Comput. Appl.* **2022**, *207*, 103513.
78. Xin, X.; Liu, K.; Loughney, S.; Wang, J.; Yang, Z. Maritime traffic clustering to capture high-risk multi-ship encounters in complex waters. *Reliab. Eng. Syst. Saf.* **2023**, *230*, 108936.
79. Zhou, T.; Qiao, Y.; Salous, S.; Liu, L.; Tao, C. Machine Learning-Based Multipath Components Clustering and Cluster Characteristics Analysis in High-Speed Railway Scenarios. *IEEE Trans. Antennas Propag.* **2022**, *70*, 4027–4039.
80. Feigin, Y.; Spitzer, H.; Giryes, R. Cluster with GANs. *Comput. Vis. Image Underst.* **2022**, *225*, 103571.
81. Piatetsky-Shapiro, G. Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop. *AI Mag.* **1990**, *11*, 68–70.

82. Fayyad, U.; Haussler, D.; Stolorz, P. KDD for Science Data Analysis: Issues and Examples. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland Oregon, 2–4 August 1996; pp. 50–56.
83. Fayyd, U.M.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*; Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R., Eds.; Morgan Kaufmann: Menlo Park, CA, USA, 1996; pp. 1–34, ISBN 0-262-56097-6.
84. Microsoft. *Data Mining*; 2006. Available online: https://msdn.microsoft.com/en-us/library/aa227240(v=vs.60).aspx (accessed on 27 October 2022).
85. Microsoft. Discretization Methods (Data Mining). Available online: https://msdn.microsoft.com/en-us/library/ms174512.aspx (accessed on 27 October 2022).
86. Fayyad, U.M.; Irani, K.B. Multi-interval discretization of continuous-valued attributes for classification learning. In Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI-93), Chambèry, France, 28 August–3 September 1993; pp. 1022–1027.
87. Vučetić, M.; Hudec, M.; Božilović, B. Fuzzy functional dependencies and linguistic interpretations employed in knowledge discovery tasks from relational databases. *Eng. Appl. Artif. Intell.* **2020**, *88*, 103395.
88. de Oliveira, E.F.; de Lima Tostes, M.E.; de Freitas, C.A.O.; Leite, J.C. Voltage THD Analysis Using Knowledge Discovery in Databases with a Decision Tree Classifier. *IEEE Access* **2018**, *6*, 1177–1188.
89. Chen, Z.; Zhu, S.; Niu, Q.; Zuo, T. Knowledge Discovery and Recommendation with Linear Mixed Model. *IEEE Access* **2020**, *8*, 38304–38317.
90. Molina-Coronado, B.; Mori, U.; Mendiburu, A.; Miguel-Alonso, J. Survey of Network Intrusion Detection Methods from the Perspective of the Knowledge Discovery in Databases Process. *IEEE Trans. Netw. Serv. Manag.* **2020**, *17*, 2451–2479.
91. Sanchez Sanchez, P.A.; Cano Zuluaga, J.; Garcia Herazo, D.; Pinzon Baldion, A.F.; Rodriguez Mercado, G.; Garcia Gonzalez, J.R.; Perez Coronell, L.H. Knowledge Discovery in Musical Databases for Moods Detection. *IEEE Lat. Am. Trans.* **2019**, *17*, 2061–2068.
92. Kamm, S.; Jazdi, N.; Weyrich, M. Knowledge Discovery in Heterogeneous and Unstructured Data of Industry 4.0 Systems: Challenges and Approaches. *Procedia CIRP* **2021**, *104*, 975–980.
93. González García, C.; García-Bustelo, C.P.; Espada, J.P.; Cueva-Fernandez, G. Midgar: Generation of heterogeneous objects interconnecting applications. A Domain Specific Language proposal for Internet of Things scenarios. *Comput. Netw.* **2014**, *64*, 143–158.
94. Rosa, C.R.M.; Steiner, M.T.A.; Steiner Neto, P.J. Knowledge Discovery in Data Bases: A Case Study in a Private Institution of Higher Education. *IEEE Lat. Am. Trans.* **2018**, *16*, 2027–2032.
95. Mashey, J.R. Big Data and the next wave of infraStress. In *Computer Science Division Seminar*; University of California: Berkeley, CA, USA, 1997.
96. Weiss, S.M.; Indurkhya, N. *Predictive DATA Mining: A Practical Guide*, 1st ed.; Morgan Kaufmann: San Francisco, CA, USA, 1997; ISBN 978-1558604032.
97. Diebold, F. *On the Origin(s) and Development of the Term Big Data*; University of Pennsylvania: Philadelphia, PA, USA, 2012.
98. Hey, T.; Tansley, S.; Tolle, K. *The Fourth Paradigm: Data-Intensive Scientific Discovery*; Microsoft Research: Redmond, WA, USA, 2009; ISBN 9780982544204.
99. Philip Chen, C.L.; Zhang, C.Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf. Sci.* **2014**, *275*, 314–347.
100. Wu, X.; Zhu, X.; Wu, G.-Q.; Ding, W. Data mining with big data. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 97–107.
101. Howie, T. The Big Bang: How the Big Data Explosion Is Changing the World. Available online: https://blogs.msdn.microsoft.com/microsoftenterpriseinsight/2013/04/15/the-big-bang-how-the-big-data-explosion-is-changing-the-world/ (accessed on 27 October 2022).
102. NIST Big Data Public Working Group: Definitions and Taxonomies Subgroup. *NIST Big Data Interoperability Framework: Volume 1, Definitions*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2015; Volume 1.
103. Chen, M.; Mao, S.; Liu, Y. Big Data: A Survey. *Mob. Netw. Appl.* **2014**, *19*, 171–209.
104. Dutcher, J. What Is Big Data? Available online: https://datascience.berkeley.edu/what-is-big-data/ (accessed on 25 May 2016).
105. Ward, J.S.; Barker, A. Undefined By Data: A Survey of Big Data Definitions. *arXiv* **2013**, arXiv:1309.5821.
106. Intel IT Center. *Big Data Analytics. Intel's IT Manager Survey on How Organizations Are Using Big Data*; Intel Corporation: Santa Clara, CA, USA, 2012.
107. Pettey, C.; Goasduff, L. Gartner Says Solving "Big Data" Challenge Involves More Than Just Managing Volumes of Data. Available online: https://web.archive.org/web/20180924135856/https://www.gartner.com/newsroom/id/1731916 (accessed on 13 November 2018).
108. Gartner Inc. IT Glossary: Big Data. Available online: https://www.gartner.com/en/information-technology/glossary/big-data (accessed on 27 October 2022).
109. Gantz, B.J.; Reinsel, D. Extracting Value from Chaos. *IDC* **2011**, *1142*, 1–12.
110. NIST Big Data Public Working Group: Technology Roadmap Subgroup. *NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2015; Volume 7.
111. Mohammadi, M.; Al-Fuqaha, A.; Sorour, S.; Guizani, M. Deep Learning for IoT Big Data and Streaming Analytics: A Survey. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 2923–2960.

112. Lin, R.; Ye, Z.; Wang, H.; Wu, B. Chronic Diseases and Health Monitoring Big Data: A Survey. *IEEE Rev. Biomed. Eng.* **2018**, *11*, 275–288.
113. Manley, K.; Nyelele, C.; Egoh, B.N. A review of machine learning and big data applications in addressing ecosystem service research gaps. *Ecosyst. Serv.* **2022**, *57*, 101478.
114. Nguyen, T.; Gosine, R.G.; Warrian, P. A Systematic Review of Big Data Analytics for Oil and Gas Industry 4.0. *IEEE Access* **2020**, *8*, 61183–61201.
115. Rawat, D.B.; Doku, R.; Garuba, M. Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security. *IEEE Trans. Serv. Comput.* **2021**, *14*, 2055–2072.
116. Ma, S.; Ding, W.; Liu, Y.; Ren, S.; Yang, H. Digital twin and big data-driven sustainable smart manufacturing based on information management systems for energy-intensive industries. *Appl. Energy* **2022**, *326*, 119986.
117. Jaber, M.M.; Ali, M.H.; Abd, S.K.; Jassim, M.M.; Alkhayyat, A.; Aziz, H.W.; Alkhuwaylidee, A.R. Predicting climate factors based on big data analytics based agricultural disaster management. *Phys. Chem. Earth Parts A/B/C* **2022**, *128*, 103243.
118. Ang, K.L.-M.; Ge, F.L.; Seng, K.P. Big Educational Data & Analytics: Survey, Architecture and Challenges. *IEEE Access* **2020**, *8*, 116392–116414.
119. Laney, D. 3D Data Management: Controlling Data Volume, Velocity, and Variety. *META Gr. Res. Note* **2001**, *6*, 70.
120. Saggi, M.K.; Jain, S. A survey towards an integration of big data analytics to big insights for value-creation. *Inf. Process. Manag.* **2018**, *54*, 758–790.
121. Goldston, D. Big data: Data wrangling. *Nature* **2008**, *455*, 15.
122. Deepa, N.; Pham, Q.-V.; Nguyen, D.C.; Bhattacharya, S.; Prabadevi, B.; Gadekallu, T.R.; Maddikunta, P.K.R.; Fang, F.; Pathirana, P.N. A survey on blockchain for big data: Approaches, opportunities, and future directions. *Futur. Gener. Comput. Syst.* **2022**, *131*, 209–226.
123. NIST Big Data Public Working Group: Security and Privacy Subgroup. *NIST Big Data Interoperability Framework: Volume 4, Security and Privacy*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2015; Volume 4.
124. IBM. Big data at the speed of business. Available online: https://web.archive.org/web/20161121123223/http://www-01.ibm.com/software/data/bigdata/ (accessed on 13 November 2022).
125. Liu, Z.; Zhang, A. Sampling for Big Data Profiling: A Survey. *IEEE Access* **2020**, *8*, 72713–72726.
126. Tripathi, M.K.; Kumar, R.; Tripathi, R. Big-data driven approaches in materials science: A survey. *Mater. Today Proc.* **2020**, *26*, 1245–1249.
127. Syed, D.; Zainab, A.; Ghrayeb, A.; Refaat, S.S.; Abu-Rub, H.; Bouhali, O. Smart Grid Big Data Analytics: Survey of Technologies, Techniques, and Applications. *IEEE Access* **2021**, *9*, 59564–59585.
128. Terzi, R.; Sagiroglu, S.; Demirezen, M.U. Big Data Perspective for Driver/Driving Behavior. *IEEE Intell. Transp. Syst. Mag.* **2020**, *12*, 20–35.
129. Seddon, J.J.J.M.; Currie, W.L. A model for unpacking big data analytics in high-frequency trading. *J. Bus. Res.* **2017**, *70*, 300–307.
130. Khan, M.A.; Uddin, M.F.; Gupta, N. Seven V's of Big Data understanding Big Data to extract value. In Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, Bridgeport, CT, USA, 3–5 April 2014; pp. 1–5.
131. Gupta, Y.K.; Kumari, S. A Study of Big Data Analytics using Apache Spark with Python and Scala. In Proceedings of the 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 3–5 December 2020; pp. 471–478.
132. Fatima Ezzahra, M.; Nadia, A.; Imane, H. Big Data Dependability Opportunities & Challenges. In Proceedings of the 2019 1st International Conference on Smart Systems and Data Science (ICSSD), Rabat, Morocco, 3–4 October 2019; pp. 1–4.
133. Sivarajah, U.; Kamal, M.M.; Irani, Z.; Weerakkody, V. Critical analysis of Big Data challenges and analytical methods. *J. Bus. Res.* **2017**, *70*, 263–286.
134. Hattawi, W.; Shaban, S.; Al Shawabkah, A.; Alzu'bi, S. Recent Quality Models in BigData Applications. In Proceedings of the 2021 International Conference on Information Technology (ICIT), Amman, Jordan, 14–15 July 2021; pp. 811–815.
135. Bhardwaj, D.; Ormandjieva, O. Toward a Novel Measurement Framework for Big Data (MEGA). In Proceedings of the 2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 12–16 July 2021; pp. 1579–1586.
136. González García, C.; Meana-Llorián, D.; G-Bustelo, B.C.P.; Lovelle, J.M.C. A review about Smart Objects, Sensors, and Actuators. *Int. J. Interact. Multimed. Artif. Intell.* **2017**, *4*, 7–10.
137. Bell, G.; Hey, T.; Szalay, A. Beyond the Data Deluge. *Science* **2009**, *323*, 1297–1298.
138. Doctorow, C. Big data: Welcome to the petacentre. *Nature* **2008**, *455*, 16–21.
139. Beaver, D.; Kumar, S.; Li, H.C.; Sobel, J.; Vajgel, P. Finding a needle in Haystack: Facebook's photo storage. In Proceedings of the 9th USENIX Symposium on Operating Systems Design and Implementation (OSDI 10), Vancouver, BC, Canada, 4–6 October 2010; pp. 1–8.
140. Trewe, M. How carriers gather, track and sell your private data. The American Genius, 2012. Available online: https://theamericangenius.com/tech-1363 news/how-carriers-gather-track-and-sell-your-private-data/ (accessed on 27 October 2022).
141. Sharp, A. Dispatch from the Denver debate. Available online: https://blog.twitter.com/2012/dispatch-from-the-denver-debate (accessed on 27 October 2022).
142. Zapponi, C. GitHut. Available online: http://githut.info/ (accessed on 27 October 2022).
143. Sawant, N.; Shah, H. Big Data Application Architecture Q&A A Problem—Solution Approach. In *Intergovernmental Panel on Climate Change*; Cambridge University Press: Cambridge, UK, 2013; p. 172, ISBN 978-1430262923.

144. World Data Group. The World Bank. Available online: http://data.worldbank.org/indicator/ (accessed on 27 October 2022).
145. Twitter Inc. Twitter: Company. Available online: https://about.twitter.com/es/company (accessed on 27 October 2022).
146. Michel, F. How Many Public Photos are Uploaded to Flickr Every Day, Month, Year? Available online: https://www.flickr.com/photos/franckmichel/6855169886/ (accessed on 27 October 2022).
147. YouTube. YouTube: Statistics. Available online: https://www.youtube.com/yt/press/en/statistics.html (accessed on 9 June 2016).
148. Savitz, E. Gartner: 10 Critical Tech Trends for The Next Five Years. Available online: http://www.forbes.com/sites/ericsavitz/2012/10/22/gartner-10-critical-tech-trends-for-the-next-five-years/ (accessed on 27 October 2022).
149. Google. Google Photos: One Year, 200 Million Users, and a Whole Lot of Selfies. Available online: https://googleblog.blogspot.com.es/2016/05/google-photos-one-year-200-million.html (accessed on 27 October 2022).
150. Facebook. Newsroom. Available online: https://web.archive.org/web/20160609081220/https://newsroom.fb.com/company-info/ (accessed on 27 October 2022).
151. Cisco. Cisco Visual Networking Index: Forecast and Methodology, 2016. Available online: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-wh (accessed on 9 June 2016).
152. Warner, J. GitHub Blog. Available online: https://github.blog/2018-11-08-100m-repos/ (accessed on 27 October 2022).
153. Alvi, P.; Ali, K. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, the World's Largest and Most Powerful Generative Language Model. *arXiv* **2022**, arXiv:2201.11990.
154. Floridi, L.; Chiriatti, M. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds Mach.* **2020**, *30*, 681–694.
155. Dewdney, P.E.; Hall, P.J.; Schilizzi, R.T.; Lazio, T.J.L.W. The Square Kilometre Array. *Proc. IEEE* **2009**, *97*, 1482–1496.
156. Lazer, D.; Kennedy, R.; King, G.; Vespignani, A. The Parable of Google Flu: Traps in Big Data Analysis. *Science* **2014**, *343*, 1203–1205.
157. Boyd, D.; Crawford, K. Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Inf. Commun. Soc.* **2012**, *15*, 662–679.
158. ACM SC08 International Conference for High Performance Computing, Austin, TX, USA, 15–21 November 2008. IEEE Computer Society: Austin, TX, USA. Available online: http://sc08.supercomputing.org/ (accessed on 27 October 2022).
159. Astrophysical Research Consortium. The Sloan Digital Sky Survey SDSS. Available online: https://www.sdss.org/ (accessed on 27 October 2022).
160. Gayo-Avello, D. No, you cannot predict elections with twitter. *Internet Comput. IEEE* **2012**, *16*, 91–94.
161. Thusoo, A.; Sarma, J.S.; Jain, N.; Shao, Z.; Chakka, P.; Anthony, S.; Liu, H.; Wyckoff, P.; Murthy, R. Hive—A warehousing solution over a map-reduce framework. *Proc. VLDB Endow.* **2009**, *2*, 1626–1629.
162. Apache Software Foundation. Hue. Available online: http://gethue.com/ (accessed on 27 October 2022).