A method for constructing a dataset to reveal the industrial behaviour of big data

Mahyuddin K M Nasution

Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, Kampus USU, Padang Bunan 20155, Medan, Indonesia

E-mail: mahyuddin@usu.ac.id

Abstract. Dataset is a set of data that becomes the standard for showing the behaviour of something. In this case, the industry. For this situation, the industry. Be that as it may, each dataset consistently has restrictions to help the prediction. The prediction is about the current conduct of the industry, where yet just exists in enormous information like in big data. This paper aims to construct a concept that describes a method for developing a dataset.

1. Introduction

Data has become the basis for every activity of individuals, communities, and either profit or non-profit organizations [1,2]. In industry 4.0, the industry and commerce are undergoing a shifting where the data becomes so important to make a strategic decision [3]. Any industry produces products or services. On the one hand, both production and services require markets, and on the other hand production requires both labour and raw material resources, and services require expertise [4,5]. Furthermore, based on the world of commerce, the market requires both seller and buyer. Whereas, a buyer requires the flow of funds for buying any product [6]. The funds come from the results of his/her hard work or as a performance [7]. It states that industry and commerce are two different worlds but are as interrelated as two sides of a coin. However, the two are not related if the data related to them are not mutually supportive [8,9].

The industry will continue to take place when its production is following market demand, or its production behaves to increase welfare so that it changes market behaviour [10]. In other words, the behaviour of interested stakeholders, either individuals or organizations, determine the sustainability of industry and commerce [11]. That behaviour comes from interpreting the data. On the one hand, that interpretation of data is not as easy as a consequence of its composition. Moreover, the integrated data needs to describe the behaviour between industry and commerce, between products and markets, and between sellers and buyers [12,13]. On the other hand, this integration has an impact on the size of the enlarged data, but also the data changes size into big data as a result of the development of information technology [14]. Big data has characteristics that show its problems with handling, but it illustrates the current reality of the behaviour of all stakeholders.
2. Review towards the problem
Big data is a term used to express characteristics of data in volume, speed, and diversity from many separate sources to be processed, searched, summarized, and visualized in an integrated manner involving thousands of microprocessors [15, 16, 17].

The sources can be in the form of a computer with its users who embed both programming codes and pure data into the main-computer [18]. The main-computer is also not one, and sometimes each is in a different place [19]. So the data may be on a different and scattered computer [20]. Instead, these main-computers become a source of information for many users. Therefore, computers with each user build big data in many places [21]. Big data had, is, and will grow exponentially [22]. It relates to the principle of computation. When a computer connects to other computers in computer networks, where each computer involves a user, then indirectly it is social communities and also collections of data [23]. From the point of the characteristic of big data, the size of the data is big to refer to the volume characteristic [24]. The information space as a result of the results of the Internet and web user activities has changed the one-way information flow into two-way reciprocity [25]. As a result, the information space as big data sets [26,27]. The data recorded in the information space has a dynamic of change, starting from the one that remains unchanged, slowly, to fast changes [28]. The speed of change in data refers to the velocity characteristic of the big data [29]. Because of the diversity of data sources either technologically or technically, the data are not uniform in their order. Moreover, in the information space, there are unstructured data and structured data. Online databases are structured, but textually information is unstructured, and are arranged in various shapes and forms [30,31,32]. It results in complex data. Data complexity refers to the variety characteristic of big data [33].

Big data consists of elements that differ in both structure and composition, and the elements compose different units of data [34]. However, because of that, the units have a connectedness with each other. That connectedness is as a cause of a two-way flow of information: interactions between individuals, interactions between individuals and organizations, interactions between organizations [35]. In addition to the nature of the data, elements have a relationship. The number of elements, they available in data units, forms the valence characteristic of the big data [36]. On the other hand, the reliability of the data must also have a measurement to get the truth [37, 38]. It concerns the quality of the data. In the information space, each data has a level of reliability. The highest reliability of data is data derived from research results. Then the data collected by the organization. The lowest data reliability is data from sensors but does not include data from Internet of Thing (IoT) [39]. While data in social media is the second lowest data [40]. Thus, data quality forms the veracity characteristic of big data [41].

All data has a value, and it depends on its importance [42]. However, the importance of data does not always exist first [43]. Benefit, reward, or profit is not always the motivation of research on data [44], but the value is the characteristic of big data [45]. Three last characteristics of big data are inherent characteristics of data [46]. The interpretation of data follows on value. By involving statistics, optimization, data mining, or other mathematical formulations, the implementation of interpretation provides meaning, for example, that functions in decision making [48]. However, as a sample of data from a population, the data must meet all the requirements statistically. That is not biased, efficient, and consistent. An estimator is said to be unbiased if the expected value of the statistics is the same as the parameter value [49]. An estimator is said to be efficient if the estimator can produce the smallest standard error compared to the standard error of other estimators [50]. An estimator is said to be consistent if the opportunity to obtain a difference between statistics and parameters are close to zero if the number of individual samples increases [51]. Also, data are randomly derived from the population or have sufficient size [52]. The minimum number of samples for correlation research, experimental research, descriptive research, and causal-comparative research are 50 entities, 30 per group, 100 entities, 30 entities, respectively [53]. Likewise, it relates to the quality of the data, aside from depending on the statistical properties, the result of testing to data needs to achieve trust [54]. Usually, it conducts by the appropriate expert. Meanwhile, the relationship of data elements, in particular, is illustrated by binding data items in the database. In other words, the
relationship of data elements in the text is like the relationship between words based on grammar [55]. Therefore the sample, in particular, does not need to be so large. In general, the results of statistical data processing are under the assumption. In specific, it gives a clue that it is possible to make predictions. As a result, there is no need to consider the speed of data changes [56]. Likewise, concerning complexity, samples are always homogeneous, and certainly do not vary. Informing samples, three characteristics: volume, velocity, and variety, have not been considered in the statistical process [57]. In contrast, these six characteristics of big data to meet the terms and conditions [58]. Of course, it provides many obstacles to produce reliable information, especially related to industry and commerce.

3. An approach toward a method

There are challenges in using big data directly in solving problems related to industry and commerce [59, 60, 61, 17]. Dataset is an alternative to understanding the reality available in big data [62, 63], but transforming big data into a dataset requires a strategy. Besides, there are problems in building datasets that can represent big data [64].

![Diagram of dataset structure](image)

**Figure 1. Structure of a dataset**

The first approach to the problem is to state the dataset and then reveal its characteristics. A dataset is a collection of data objects, namely records, points, vectors, events, cases, samples, observations, or entities [65, 66]. Data objects usually consist of several attributes. Each attribute captures the character of the entity. Therefore, this definition states that the dataset contains characteristics. However, entities have different characteristics. The characteristics of the industry are different from the commerce, but
two entities have an intersection that connects them. These characteristics reveal the following definitions: The industry is an economic activity that processes materials, raw materials, semi-finished or finished goods, into goods of high quality in their use [67]. While commerce is the economic activity of exchanging goods or services or both, based on mutual agreement, but is not an implementation [68].

The definition has revealed boundaries associated with each entity, where related entities implicitly have their characteristics. The difference between big data and dataset creates a request for a method. A method, for example, to transform big data into a dataset. This transformation method brings it the characteristics of big data into the dataset in whole or in part. The method also delivers the characteristics of the entity as the target object into the dataset.

4. Design a method
In big data, information about an entity and its attributes identified by a discrete method. Structurally, the attributes capture the characteristics that reveal limited behaviour. But to enrich the entity’s behaviour, and in unstructured cases are descriptions in text. For example, on Facebook, there is a profile composed of data items structurally, but on Facebook, there are descriptions in a variety of different media: in text, pictures, sounds, video, and other information flows un-structurally. Big data has six characteristics that are valence, veracity, volume, velocity, variety, and value. Whereas, the dataset only confirms three of the six characteristics as part of it, based on the sample concept of the population. Therefore, to build a dataset that considers the details of related entities is by transforming characteristics of big data into a dataset [69, 70]. Suppose $f$ is a transformation function, $\Omega$ is information space as big data, and $D$ is a dataset. That is

$$f : \Omega \rightarrow D$$

(1)

where $V_i = \{Volume, Velocity, Variety, Veracity, Valence, Value\}$, $i = 1, 2, 3, 4, 5, 6$. The method for developing a dataset is as follows.

4.1 Discretization
The discretization process quantifies continuous properties in big data. The continuous property causes unstructured data. The streaming of textual information, for example, flows into big data. Therefore, discretization involves modelling data, which uses structuring as in Figure 1. For example, the entities of the industrial world derived from materials, raw materials, processing machines, and products. The raw material consists of various entities depending on the type of industry. Each entity has attributes.

V1 The function $f$ brings the valence characteristic of big data to the dataset. It is trivial. How not, in general, a particular entity will have a direct or indirect relationship with other entities. For example, a machine has a direct relation to an energy source, raw materials, or indirectly to the environment [71]. In one entity, attributes have a relationship with one another. So, entities have a relationship between one to another. For example, the name of raw material, place of origin of raw materials, types of raw materials (solid, liquid, gas, or mixture), and others. Data modelling is to determine each metadata or variable that holds the attributes of each entity. Discretization also relates to the type, size, position of the attribute. There is a relationship between data units that explains how valence automatically built into the dataset. Also, there are files as structures that contain unstructured text. This text is an obedient description of a grammar of a natural language but implicitly contains metadata. Files arranged as smart documents. Smart documents have implicit instructions for using their content. One clue is how this file has a relationship with the entity and its attributes. The structure arranged in Figure 1.

V2 The function $f$ brings the veracity characteristic of big data to the dataset. This transformation also naturally occurs for all data sets. Both online and single alone databases have the same status at certain angles, but in terms of data expansion, they vary. The database is one form of relationship
between the attributes of an entity. Between the attributes, there is a close relationship that causes the data to be almost homogeneous. It shows the quality of the attributes is the same. Besides, the content of smart documents is also validated by experts so that each file contains decent information, and this allows the data not to be homogeneous. The quality of the data maintained in the light of its truth, but it gives a random picture as conditions of flexibility and heterogeneity.

V3 The function $f$ also brings value characteristic of big data to the dataset directly. If the acquisition of information from big data confirmed by using information retrieval, the dataset requires verification and validation from the appropriate experts. Characteristics of dataset value are the ability to produce behaviours, together with the detail of an entity as a guide to decision making. Then, smart documents provide meaning that enriches behaviour. Moreover, the heterogeneity of data has more impact on the value characteristics of the dataset.

V4 The next $f$ function brings the variety characteristic of big data to the dataset. Databases, in general, have a rigid relationship as a result of the arrangement of rows and columns. One-level relationships occur in one database table. Multilevel relationships occur between tables in a database. However, the flexibility of the relationship must occur in the dataset. Therefore, columns and rows based on metadata are implicit in the dataset, whereas they can explicitly arise when access occurs. Thus, the relationship between data elements can change flexibly, also does not lose the meaning. That is, it includes the original relationship. Meanwhile, support for the characteristics of the variety for the dataset comes from smart documents attached to each entity. It is not only enrich meaning but also diversity. Therefore, the dataset is not increasingly complex, but the flexibility reduces complexity. It is as a consequence of the order of the content of the dataset that has followed a design. A design that also reduces the diversity of structures.

V5 The function $f$ must also bring velocity characteristics from big data to the dataset. However, it is related to the speed of change that continues to occur. That is a consequence of the connection and interaction of the device with the user. This characteristic is going well. In contrast, about the dataset, velocity characteristics are related to the speed of change by adding appropriate information to the dataset. This change is related to regulations and patterns of implementing changes both manually and automatically. Of course, this involves verification and validation. Therefore, the velocity characteristics of the dataset are closely related to the addition of information.

V6 The function $f$ brings the volume characteristics of big data to the dataset. Both structured and unstructured, the contents of the dataset have limitations, but the minimum size of the dataset must be under the terms and conditions. Each attribute has a maximum limit to accommodate the contents of the data item. These properties are naturally carried into tables, and so on into databases. Besides, the main reason for this limitation is that it comes from available data processing hardware. However, the flexibility of the structure is the basis for fulfilling the size of the dataset. Moreover, the presence of smart documents provides reasons for fulfilling the size of the dataset. It does not become an obstacle to a data collection be dataset can represent big data. Moreover, the addition of information continues to occur as a result of patterns and regulations for adding data. Therefore, the size of the dataset’s content qualifies to predict the behaviour of related entities, including problems related to industry and commerce, for example.

4.2 Discretization Build a template
Based on the results of discretization, the next design forms a template as a flexible placeholder for information. This step involves things as follows: Searching all entities, determine and establish their attributes, identify the identity of each, and presenting them in an implicit structure in a file. Each entity is in one file. While, the attributes are arranged linearly from left to right in a file, where each item has a marker boundary. The split () instruction in Python, for example, discretizes text data. The filename.split ("\n") instruction generally recognizes the order of data in filename line by line, where
('n') contains the ENTER parameter or switches the cursor with 13h and 14h of ASCII. If () contains the parameter "," or (';') for rows.split (';') there is a marker boundary to separate data items. Each row of data has an identity, which is a unique item. Therefore, each row has a unique item as a key to connect to the related smart documents. These smart documents arranged in one or more files for one data row. So each entity is connected to many smart documents. Each template requires verification from experts in the field of data science. Therefore, the template implicitly contains entities, relations between them, attributes and their relations, and descriptions.

4.3 Collecting and verifying data
Templates can be created automatically through a computer program by translating the design into a program code [72]. Meanwhile, the contents of the dataset template manually or automatically add up one by one from the information source. Sources of information are sources of quotations from the dataset derived from the information space. Sources of information can come from manual documents such as books, journals, proceedings, or other reports. Collecting information into the dataset is manually carried out by the collector and then one by one verified by experts. Automatically, the contents of the dataset can come from the information space by extracting it. A computer program to obtain dataset content from the information space is a capable program of translating the dataset structure into a pattern of access to information in an unstructured information space. The patterns forms clusters based on the query that produces information according to the data items in the template. In different cases, the pattern also adheres to the rules based on the classification. Thus, data collection takes place either manually, semi-automatically, or automatically into the dataset. It is necessary to fulfil the Vi characteristics in an integrated manner.

4.4 Testing and validation
Testing a set of data or all data from the dataset is necessary to obtain the validity of the data. When the data is tested and valid, the dataset becomes standard. Thus, adding information to the dataset results in retesting. The test is to perpetuate the characteristics of big data in the dataset. Testing on the relationship of data items, tables, files in an integrated function is to avoid the nature of outliers or broken links between data elements. In general, there is always the relationships between data elements: row by row, following the column by column, or diagonally row and column, like in the matrix so that the degree of each matrix element is at least one. However, heterogeneity is the responsibility of randomness, namely testing the data randomly. Thus the other characteristics of the dataset have their tests and are integrated. It includes testing the speed of change that may be done by predicting it based on that dataset, which then validated against big data. Validation carried out to adjust the inherent characteristics of the data through the consistency of changes both through classification and classification.
Prediction is a behaviour that continues to be part of the industrial world. An industry will stop if it cannot meet market demand, or if production is excessive. Along these lines, the expectation of the measure of creation is significant in the industry world.

5. Conclusion
A dataset that adheres to the characteristics of big data needs a way to represent it. The strategy comprises of the accompanying advances: discretization, building templates, data collection, verification, testing, and validation. An idea that depicts the technique above has portrayed, next is to formalize and demonstrate it hypothetically.

References
[1] R Ibrahim, M Q Shafiq 2018 On the measurement and analysis of safety in a large city Proceedings - 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 2017 IEEE 15th International Conference on Pervasive Intelligence and Computing, 2017 IEEE 3rd International Conference on Big Data Intelligence and
Computing and 2017 IEEE Cyber Science and Technology Congress, DASC-PICom-DataCom-CyberSciTec 2017

[2] M K M Nasution, I Aulia and M Elveny 2019 Data Journal of Physics: Conference Series 1235(1)

[3] T Latinovic, D Preradovic, C R Barz, A Pop Vadean and M Todic 2019 Big Data as the basis for the innovative development strategy of the Industry 4.0 IOP Conference Series: Materials Science and Engineering 477(1)

[4] E W McCullough 1928 The Relation of the Chamber of Commerce of the United States of America to the Growth of the Simplification Program in American Industry The Annals of the American Academy of Political and Social Science 137(1).

[5] A Ronzano 2017 Bill on transparency, fight against corruption and modernisation of business life: The French Senate passes the Sapin II bill without modifying the text passed by the National Assembly on the topic of industry-commerce relations Concurrences 2017(1)

[6] K Mazouz, A Mohamed and B Saadouni 2016 Stock return comovement around the Dow Jones Islamic Market World Index revisions Journal of Economic Behavior and Organization 132

[7] A Umar, A H Sasonogko, I T Widjyastuti and Y Christanti 2020 Improving the social enterprise-based business performance from the aspect of social business model canvas International Journal of Entrepreneurship 24(1)

[8] S E Fkili and R Faizi 2018 How big data is reshaping the e-commerce industry Proceedings of the 31st International Business Information Management Association Conference, IBIMA 2018: Innovation Management and Education Excellence through Vision 2020

[9] Y Hu and H Ma 2019 Big Data Information Technology and the Development of E-commerce in Cultural Industry Advances in Intelligent Systems and Computing 929

[10] S S Oren and S Oren 2001 Market structure, market rules and market behavior in a competitive electricity industry Proceedings of the IEEE Power Engineering Society Transmission and Distribution Conference 1(WINTER MEETING)

[11] P Nisbet and W Thomas 2000 Attitudes, Expectations and Labour Market Behaviour: The Case of Self-Employment in the UK Construction Industry Work Employment & Society 14(2)

[12] A Balar, N Malviya, S Prasad and A Gangurde 2013 Forecasting consumer behavior with innovative value proposition for organizations using big data analytics 2013 IEEE International Conference on Computational Intelligence and Computing Research IEEE ICCIC 2013

[13] C Lehrer, A Wieneke, J vom Brocke, R Jung and S Seidel 2018 How Big Data Analytics Enables Service Innovation: Materiality, Affordance, and the Individualization of Service Journal of Management Information Systems 35(2)

[14] R R Shker and M N Hussin 2020 The impact of information technology and big data in business intelligence Journal of Physics: Conference Series 1530(1)

[15] A Wegener 2012 HPC and “big data” apps tap floating-point number compression Electronic Design 60(2)

[16] S Soares, T Deutsch, S Hanna and P Malik 2012 Big data governance: A framework to assess maturity IBM Data Management Magazine Issue 3

[17] S R Tiwari and K K Rana 2021 Feature Selection in Big Data: Trends and Challenges Lecture Notes on Data Engineering and Communications Technologies 52

[18] K E Zannat and C F Choudhury 2019 Emerging Big Data Sources for Public Transport Planning: A Systematic Review on Current State of Art and Future Research Directions Journal of the Indian Institute of Science 99(4)

[19] I A Kamil and S O Ogundoyin 2019 A big data anonymous batch verification scheme with conditional privacy preservation for power injection over vehicular network and 5G smart grid slice Sustainable Energy, Grids and Networks 20
[20] L Qian, J Yu, G Zhu, R Xie, Z Chen, L Wang, H Pang, Z Mei, M Xu and H Chen 2019 Review of the Basic Technology of Big Data on Analysis and Storage Applied in Power Grids IOP Conference Series: Earth and Environmental Science 440(3)

[21] J Tang and H Sui 2017 Application technology of big data in smart grid and its development prospect Proceedings - 2017 International Conference on Computer Technology, Electronics and Communication ICCTEC 2017

[22] M K M Nasution, M Elveny, R Syah and S A Noah 2015 Behavior of the resources in the growth of social network Proceedings - 5th International Conference on Electrical Engineering and Informatics: Bridging the Knowledge between Academic, Industry, and Community ICEEI 2015

[23] M K M Nasution 2013 Superficial method for extracting academic social network from the Web Ph.D. Thesis.

[24] Q Xu, Z X Zhao and W Wang 2013 Volume-based data representation of big data analysis Advanced Materials Research 798

[25] M K M Nasution 2011 Web: Teknologi yang mengubah aliran informasi Renungan

[26] P Golubtsov 2019 Information spaces for big data processing: Unification and parallelization of sequential information accumulation procedures Proceedings - 21st IEEE Conference on Business Informatics CBI 2019 1

[27] H M Walker 2020 Bias in Algorithms and the Misuse of Big Data Sets ACM Inroads 11(2)

[28] L Dupr’e and Y Demchenko 2016 Impact of information security measures on the velocity of big data infrastructures 2016 International Conference on High Performance Computing and Simulation HPCS 2016

[29] A M Al-Salim, T E H El-Gorashi, A Q Lawey and J M H Elmirghani 2018 Greening big data networks: Velocity impact IET Optoelectronics 12(3)

[30] B Shneiderman and C Plaisant 2015 Sharpening analytic focus to cope with big data volume and variety IEEE Computer Graphics and Applications 35(3)

[31] J W Williams, K S Aggour, J Interrante, J McHugh and E Pool 2015 Bridging high velocity and high volume industrial big data through distributed in-memory storage & analytics Proceedings – 2014 IEEE International Conference on Big Data IEEE Big Data 2014

[32] E D Valle, D Dell’aglio and A Margara 2016 Tutorial: Taming velocity and variety simultaneously in big data with stream reasoning DEBS 2016 - Proceedings of the 10th ACM International Conference on Distributed and Event-Based Systems

[33] R C Whetsel and Y Qu 2018 Quantifying the impact of big data’s variety 2017 3rd IEEE International Conference on Computer and Communications ICCC 2017

[34] R Dautov and S Distefano 2017 Quantifying volume, velocity, and variety to support (Big) data-intensive application development Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, 5th IEEE International Conference on Big Data, Big Data 2017

[35] M K M Nasution 2016 Social network mining (SNM): A definition of relation between the resources and SNA International Journal on Advanced Science, Engineering and Information Technology 6(6)

[36] M Ghasemaghaei 2020 The role of positive and negative valence factors on the impact of bigness of data on big data analytics usage International Journal of Information Management 50

[37] P Li and S Guo 2014 Load balancing for privacy-preserving access to big data in cloud Proceedings - IEEE INFOCOM

[38] W Q Meeker and Y Hong 2014 Reliability meets big data: Opportunities and challenges Quality Engineering 26(1)

[39] W-L Chin, W Li and H-H Chen 2017 Energy Big Data Security Threats in IoT-Based Smart Grid Communications IEEE Communications Magazine 55(10)
[40] C Huang and D Wang 2015 Spatial-Temporal Aware Truth Finding in Big Data Social Sensing Applications Proceedings - 14th IEEE International Conference on Trust, Security and Privacy in Computing and Communications TrustCom 2015

[41] L Berti-Equille and M L Ba 2016 Veracity of big data: Challenges of cross-modal truth discovery Journal of Data and Information Quality 7(3)

[42] P A Griffin and A M Wright 2015 Commentaries on big datas importance for accounting and auditing Accounting Horizons 29(2)

[43] J Park, Y Kim and W Jung 2018 Use of a big data mining technique to extract relative importance of performance shaping factors from event investigation reports Advances in Intelligent Systems and Computing 589

[44] J Amudhavel, V Padmapriya, V Gowri, K LakshmiPriya, K PremKumar and B Thiyagarajan 2015 Perspectives, motivations and implications of big data analytics ACM International Conference Proceedings Series 06-07-March-2015

[45] K Pantelis and L Aija 2013 Understanding the value of (big) data Proceedings - 2013 IEEE International Conference on Big Data Big Data 2013

[46] P Carruthers 2018 Valence and Value Philosophy and Phenomenological Research 97(3)

[47] J Debbattista, C Lange, S Scerri and S Auer 2016 Linked 'Big' Data: Towards a Manifold Increase in Big Data Value and Veracity Proceedings - 2015 2nd IEEE/ACM International Symposium on Big Data Computing BDC 2015

[48] H G Miller and P Mork 2013 From data to decisions: A value chain for big data IT Professional 15(1)

[49] S F Shah, S A Cheema, Z Hussain and E A Shah 2020 Masking data: a solution to social desirability bias in paired comparison experiments Communications in Statistics: Simulation and Computation

[50] J P Ziliak 1997 Efficient estimation with panel data when instruments are predetermined: An empirical comparison of moment-condition estimators Journal of Business and Economic Statistics 15(4)

[51] J-J Ren 1997 On self-consistent estimators and kernel density estimators with doubly censored data Journal of Statistical Planning and Inference 64(1)

[52] N Roy Chowdhury, D Cook, H Hofmann, M Majumder, E-K Lee and A L Toth 2015 Using visual statistical inference to better understand random class separations in high dimension, low sample size data Computational Statistics 30(2)

[53] Jack R Fraenkel and Norman E Wallen 1990 How to Design and Evaluate Research in Education

[54] M Huberty 2015 Awaiting the Second Big Data Revolution: From Digital Noise to Value Creation Journal of Industry, Competition and Trade 15(1)

[55] K Englemeier 2016 Role and importance of semantic search in big data governance Big-Data Analytics and Cloud Computing: Theory, Algorithms and Applications

[56] B MacVicar, S Dilling and J Lacey 2014 Multi-instrument turbulence toolbox (MITT): Open-source MATLAB algorithms for the analysis of high-frequency flow velocity time series datasets Computers and Geosciences 73

[57] M R Evans, D Oliver, X Zhou and S Shekhar 2014 Spatial big data: Case studies on volume, velocity, and variety Big Data: Techniques and Technologies in Geoinformatics

[58] S Drisya and K Sreekumar 2017 Volume challenges of big data: Traditional vs big data Journal of Advanced Research in Dynamical and Control Systems 9

[59] Q Miao, J Jiang, H Zhang and C Luo 2019 Development of aviation intelligent engine under industrial big data: Chances and challenges Yi Qi Yi Biao Xue Bao/Chinese Journal of Scientific Instrument 40(7)

[60] B Zhang, Z Du, B Wang and Z Wang 2019 Motivation and challenges for e-commerce in e-waste recycling under Big data context: A perspective from household willingness in China Technological Forecasting and Social Change 144
[61] A Ashabi, S B Sahibuddin and M S Haghighi 2020 Big Data: Current Challenges and Future Scope ISCAIE 2020 - IEEE 10th Symposium on Computer Applications and Industrial Electronics

[62] Anonim (Editorial) 2019 Big data creates big opportunities: The promises and challenges of large veterinary health datasets Topics in Companion Animal Medicine 37

[63] T S B Abd Manan, T Khan, W H M W Mohtar, S Beddu, N L M Kamal, S Yavari, H Jusoh, S Qazi, S K B I Supaat, F Adnan, A A Ghanim, S Yavari, A Machmudah, A Rajabi, M Porhemmat, M Irfan, M T Abdullah and E S B A Shakur 2020 Dataset on specific UV absorbances (SUV A254) at stretch components of Perak River basin Data in Brief 30

[64] H Basim Alwan and K R Ku-Mahamud 2020 Big data: Definition, characteristics, life cycle, applications, and challenges IOP Conference Series: Materials Science and Engineering 769(1)

[65] M M Hasan, J Popp and J Ol´ah 2020 Current landscape and influence of big data on finance Journal of Big Data 7

[66] M Favaretto, E de Clercq, C O Schneble and B S Elger 2020 What is your definition of Big Data? Researchers understanding of the phenomenon of the decade PLoS ONE 15(2)

[67] M Carus, L Dammer, A Raschka and P Skoczinski 2020 Renewable carbon: Key to a sustainable and future-oriented chemical and plastic industry: Definition, strategy, measures and potential Greenhouse Gases: Science and Technology 10(3)

[68] X Lin, Y Li and X Wang 2017 Social commerce research: Definition, research themes and the trends International Journal of Information Management 37(3)

[69] R Boden and A Nucci 1982 The 1982 Characteristics of Business Owners dataset Small Business Economics 2(4)

[70] N Alshammari and S Alanazi 2020 An Arabic dataset for disease named entity recognition with multi- annotation schemes Data 5(3)

[71] W A Winarno and B Tjahjadi 2017 The Moderating Effect of Industry Environments on the Relationship between IT Asset Portfolios, Efficiency and Innovation in the ERP Context European Research Studies Journal XX(2A)

[72] F Muhammad and L Susilowati 2020 Algorithm and computer program to determine metric dimension of graph Journal of Physics: Conference Series 1494