Data science: A Review towards the Big Data Problems

Mahyuddin K. M. Nasution, Opim Salim Sitompul, Marischa Elveny, and Rahmad Syah
Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Sumatera Utara, USU, Padang Bulan 20155, Medan, Sumatera Utara, Indonesia
E-mail: mahyuddin@usu.ac.id

Abstract. Data science is the naming of science that can change when dealing with its subject, big data, into big data science. Extraction as the main task of and based on the definition of data science requires an interpretive way of big data. This interpretation follows the characteristics of big data, namely a review of several problems that arise concerning the characters of big data is as an approach. The goal is that data and information, in information extraction or knowledge extraction from the information space, can well organize as is the case in social networks. This paper aims to provide a brief description of it.

1. Introduction
Data Science, based on its definition, is a science that concerned with the extraction of useful knowledge from large and complex data sets [1], takes the direction of revealing information as the extraction of knowledge from information sources, namely big data [2]. Big data has characteristics that require a way to overcome constraints when extracting knowledge from it because most big data contains unstructured data [3, 4, 5, 6, 7, 8]. Web, as a document containing unstructured data/information, represents the entire global information world [9]. The contents of the Web indirectly give useful information for making decisions [10]. In general, a collection of information expressed as an information space, which requires infrastructure for management to become a source of knowledge [11, 12]. This paper reveals several problems related to big data by its characteristics. Then, with some cases, specifically relating to the extraction of social networks, explanations, and reviews were added.

2. Background
A review about data science, it based on some of the literature available relating to data. However, it will pay particular attention to literature with titles that contain phrases: Data Science.

2.1. Data Science
A review of data science begins with one definition of data science [1]:

Data science is the extraction of knowledge from high-volume data, using skills in computing science, statistics and the specialist domain knowledge of experts [13].
Emphasis on extracting knowledge from information sources is an approach that follows the pyramid concept of the relationship for data-information-knowledge-wisdom as the outcome of the input-process-output system [2]. These terms have interpretation briefly as follows [14]:

D₁ Data is a basic fact as the smallest unit that cannot be separated.

D₂ Information is data that has been processed and is in a structure that is useful for decisions making.

D₃ Knowledge is things that are true as familiarity, awareness, or understanding, not opinions.

D₄ Wisdom is the state of something or quality related to the right knowledge with a fair (unbiased) assessment of action, intelligence, or insight.

However, the presence of wisdom is as a result of mastery of knowledge. The mastery starts with the presence of data and information. Data recorded in the database in this case, the database is an information room [15]. Data also exists either in a structured or unstructured form. In general, data is not always alone in the information space but mixed with that information. The information space, let’s call it as Ω [11, 12], contains a variety of information, which in general refers to as multimedia, namely data, databases, text, hypertext, images, audio (sound stream), and video (frame stream). Therefore, even though each of the multimedia content has a text description, it is still necessary to have a method of gaining knowledge from each content. In general, it involves extraction [16].

An extraction involves a method. The information or knowledge extraction method adapts to the types of information sources. For example, social network extraction from Web. There are two categories of approaches, namely classification and clustering. The supervised research stream organizes knowledge classifications from the dataset or corpus by involving class labels as training data [17]. The unsupervised research stream conducts the clustering of knowledge from information sources without the involvement of human supervision [18, 19]. Therefore, dataset as the training data is to produce an inferred function is to predict new things and prove the results in the information space as a more concise implementation [20].

3. An Approach

The keyword for the definition of data science, high-volume data [21], establishes that the information space has the following characters: value, valence, veracity, variety, velocity, and volume. Therefore, the approach to data science reviews involves problems from big data, namely discussions around the character of big data.

V₁ Value is a characteristic of the information space where there is interest in any information [3]. For example, information is in the form of text, and in the information space, there is a document (denoted by d) or a web page (note by ω). Each of these may consist of multimedia content. Suppose ω ∈ Ω consists of a set of words as a token, written as \( w = \{w_1, w_2, \ldots, w_K\} \), \( k \in K \) is the number of words uniquely in the document or \( |w| = k \), while the document size \( |\omega| = l \), where \( k \leq l \). Therefore, each word in the document has three times, \( |w| = h \), where \( h \) is the frequency (occurrence) of the words \( w \) [22]. The word "abstract" in this paper appears twice, so |"abstract"| = a = 3. Meanwhile, the same word in the information space is 1,410,000,000 hits with a quotation time by Google’s search engine of 0.52 seconds, for example. In the information space, both as a document file and as a database is the first protection of information. The most valuable information is information that is trusted and is always protected by security. The level of trustworthiness of information, starting from the highest class is documents of research results or academic documents such as books, articles in journals, and papers in proceedings. Meanwhile, the lowest document is web-blog. Even so, every personal web-blog can use as a study material that reveals this private behavior. However, the most valuable information is information
that is trusted and is always protected by security. Therefore, knowledge extraction from the information space requires validation. That is, by conducting surveys and borrowing recall-precision. Value is the nature of data [23].

\textbf{V}_2 \textit{Valence} is a characteristic of the information space which is naturally inherited from various activities by information creators/users [8]. The information contains data items. It is words that build sentences and then become documents. Is it structured data in a database where each one of the data items has a relationship with another. Is it an image or other stream of information that has identity descriptions in the form of text as well. This property is the basis for the birth of the network concept in information space. The network consists of: a collection of points, abstracted as \( V = \{ v_1, v_2, \ldots, v_n \} \), \( V \neq \emptyset \); a collection of edges, written as \( E = \{ e_1, e_2, \ldots, e_m \} \); and label sets, written as \( L = \{ l_1, l_2, \ldots, l_k \} \) [24]. The formal network is \( G(V, E) \) indicating the structure of the associated object. The development of the notation \( G(V, E(L)) \), \( G(V(L), E) \) or \( G(V(L), E(L)) \) aims to express the network by completing the label as meaning giver structure. The network provides an interpretation of the existence of the relationship between objects. The words network shows that, in a document, there are words that connect to another. Via the query \( q \) the relationship can be expressed via the search engine, for example, where occurrence and co-occurrence apply. Occurrences represent one target meaning in the query or contained in a subspace of information. Co-occurrence completes a pair of target meanings in the query, or it is also called the presence of two objects in the information sub-space simultaneously. Formally, for the information subspace \( \Omega_x \in \Omega \), \( |\Omega_x| = \{(t_x, \omega_x)\}_h \), \( h = 1, \ldots, H \), where \( t_x \) is a search term consisting of \( \{w_1, w_2, \ldots, w_k\} \) and \( \Omega_x \) acting as an occurrence [11]. The information subspace \( \Omega_x \cap \Omega_y \in \Omega \) as a presentation of query-based co-occurrence and search engines, namely \( \Omega_x \cap \Omega_y = \{(t_x, \omega_x)\}_h \cap \{(t_y, \omega_y)\}_h \) [12]. The use of the intersection operator \( \cap \) expresses the principle of overlapping information as a valence generator. This principle can also express by the presence of the same URL (Uniform Resource Locator) address in two occurrences. The composition of the URL contains a set of tokens \( U = \{ s, d_1, \ldots, d_m, p_1, \ldots, p_{n-1} \} \) that satisfies the structure

\[ s://d_m, \cdots d_2, d_1/p_1/p_2/\cdots/d_{n-1}, \]

namely a string consisting of a scheme, priority, and path. With the canonical form consisting of components such as in the set \( U = \{ s, d, p, q \} = \{ \text{scheme, authority, path, query} \} \), which is a string form:

\[ s://d_m, \cdots d_2, d_1/p_1/p_2/\cdots/d_{n-2}/x \]

where \( x = p_{n-1} \), or \( x = p_{n-1}/q \), such until the URL has a \( n \) layer, each part is separated by a slash, "/" [25]. \( \Omega \) is as a source of information consists of various valences, which provide reliable predictions regarding the occurrence and co-occurrence or the relationship between data items [26].

\textbf{V}_3 \textit{Veracity} is a characteristic of an information space that specifically contains quality information, for example, avoiding errors in spelling. The quality of data in the information space can vary widely and depends on the creator of the information [6]. Garbage in garbage out (GIGO) is an inherent characteristic of computers, where users are human, and humans are negligent in many ways. Humans are as user operators who enter data into the computer have the opportunity to make mistakes reduction of this error can only do through verification from a different person. The information that the computer will present to the user depends on the data that enters the computer. Therefore, it always requires verification. Data verification is performed by experts who compare what enters into the computer with data printed on the hard copy. Meanwhile, the information contained in
the document in a textual manner and online form on the web has a comparative value. A computer program can represent experts specifically comparing each word on the web, for example, against the related words in an online dictionary. A comparison between the number of misspelled words with the correct words will give a reliable percentage of the performance of the operator entering textual data. Entry errors are one of the determinants of information quality besides fraudulent information from buzzers [27].

**V**<sub>4</sub> *Variety* is a characteristic of heterogeneous information space, not only based on the type of data [5]. Traditional data types are neatly organized and organized in databases that rely on table or column and row format where there are data item identifier metadata. With that, in the information space, some sub-spaces have different structures and densities. Online databases have a high density, compared to documents composed of text. Thus, in data diversity, there is unstructured data as the content of the information space, such as text, audio, and video, which require pre-processing before obtaining metadata [28]. Data types such as strands have a length of presentation in memory, is as is the size of the text that is allowed as part of a query or the number of words in a single plaintext file. The type of numeric data, depending on the numeric type, integers different from the real number or has floating points. Data types vary not only from the software point of view but also from the hardware. This type of data has a limit on which software or hardware is a means.

**V**<sub>5</sub> *Velocity* is a characteristic of the information space regarding the rate at which data received [7]. This characteristic follows the speed associated with the hardware. That is the speed of the data stream when it pulled from memory or other storage [29]. Concerning data, both hardware and software or other means always related to the time used to stream data. For example, processor speed affects on both incoming and outgoing processing times. Server speed also depends on the installed processor speed. About network media, it will depend on the type of media, for example the speed of data flow in a coaxial cable will differ from the speed using optical fiber. For example, the size of a data stream that can move across the transmission medium at one time is a measure known as bandwidth. Thus, sometimes this requires follow-up.

**V**<sub>6</sub> *Volume* is a characteristic of the information space related to the size of the data, which demands high-speed processing [4]. Data sizes vary, but may contain tens of terabytes of data, some even up to the petabyte order. The size of the data also depends on the storage. To overcome this limitation, when data exceed a storage space, then data will be separated in the various available other storage [30]. The amount of data that can be handled by a computer depends on the capabilities of the processor. It follows the binary calculation concept of \(2^i\), \(i = 1, 2, 3, \ldots, n\). For example, the processor Intel 80386 DX is a processor 32 bit or \(2^5\) with 4 GB (Giga-Byte) of physical address memory. This processor has one additional instruction, namely BSF / BSR is as an instruction: *scan bit forward/reverse* [31].

Thus, the development of processor technology is a result of the need for data management. However, changes in hardware technology must also follow by changes in software else the processing speed will be the same but also the data volume.

With these characteristics, the information space is nothing but big data. Because of it, data science sometimes, known as, in a new name, is big data science.

### 4. Cases

Science, like statistics, is the implication of data, where the disclosure of data behavior, namely random, discrete, continuous, obscure, and others, is a necessary feature of each. However, statistics are not the implication of big data. The implication of big data is a change in the treatment of various methods that have been testing, processing, and then visualizing it as knowledge. Big data also changes several applications of technology, as well as changes that
occur in information technology [32, 33]. For example, the presence of the Internet of Things (IoT) is currently in many uses. After all, statistics cannot rely on manual computation to deal with massive data, the nature of the human ability to compute a computational constraint. Also, the limitations of both hardware and software require methods to expand human capabilities through other technologies such as artificial intelligence.

One interpretation of data science is extraction. Direct access to big data involves several tools, such as search engines, queries, and other ancillaries. This method is unsupervised, which relies on a superficial grasp of the dynamics of information in the information space. This measurement involved the singleton and doubleton in hits measure. By involving any search engine, such as Google search engine, for example, for the query \( a = "\text{Mahyuddin KM Nasution}" \) or the query \( b = "\text{Bayu Rahmad Syah}" \), the search engines return respectively \( |\Omega_a| = 6.160 \) hits and \( |\Omega_b| = 8 \) hits. In this way, it reveals the performance scores of two academic or scientific actors, which allow them to give temporary meanings: One is a senior scientist while the other is a junior scientist, for example. Meanwhile, for the query \( ab = "\text{Mahyuddin K. M. Nasution}, \text{Bayu Rahmad Syah}" \) search engine returns \( |\Omega_a \cap \Omega_b| = 4 \), and shows the relationship between the two scientists. It implies the extraction has yielded value for value in the interest of data science.

Semantically, occurrence and co-occurrence give the meaning of the strength relationship between the two scientists by involving the measurement of similarity [34],

\[
N_c = \frac{2|\Omega_a \cap \Omega_b|^2}{|\Omega_a|^2 + |\Omega_b|^2}
\]

yielding the strength relationship between the two scientists is \( N_c = 0.000001 \). Social actors such as scientists and academics through their scientific works form social networks of academics. The singleton \( |\Omega_a| \geq 0 \) and \( |\Omega_b| \geq 0 \) indicate the existence of the social actor \( a, b \in A \), where \( A \) is the set of social actors.

The vertices of \( v_i \in V \) represent different social actors, \( i = 1, 2, \ldots, n \), or \( v_1 = a \) and \( v_2 = b \) for example, whereas co-occurrence underlies their formation the relationship if \( |\Omega_a \cap \Omega_b| \geq 0 \) as a doubleton, which has an implementation in the form of a strength relation. Based on the level of existence of the strength relation \( r \in R \) greater than the specified threshold, that is \( \alpha \), the measurement result \( r = N_c \) represents \( e_i \in E \), where \( R \) is the set of power relations. So extraction, in this case, reveals not only the value but also the valence of big data. Let \( \gamma \) represent the extraction, \( \gamma : G(V, E) \rightarrow G(A, R) \), where \( G(A, R) \) is a social network resultant with \( n \) social actors. It implies that there are a number of ties between one object and another, or a relationship between one social actor and other social actors. Valence measures aim to reveal the attractiveness of social actors, and this is related to adjacent in social networks, which is sometimes indicated by the degree of a vertex. Suppose that the extraction of a social network from the web generates a set of relations \( E_1 \), whereas a survey-based social network generates a set of relations \( E_2 \), the measurement of the percentage based on similarity of the two social networks is Jaccard coefficient, that is

\[
J_C = \frac{|E_1 \cap E_2|}{|E_1| + |E_2| - |E_1 \cap E_2|}
\]

with precision: \( Prec = \frac{|E_1 \cap E_2|}{|E_1|} \), recall: \( Rec = \frac{|E_1 \cap E_2|}{|E_2|} \), and \( F_{\text{measure}} \) is \( 2* \frac{|E_1 \cap E_2|}{|E_1| + |E_2|} \). On the other hand, the buzz information that appears in various media can reduce by involving social networks [36]. When a buzzer’s name appears, the subsequent buzzers are well recognized, and the forensic information acts as the veracity of big data and discards the buzz information.

The addition of labels to each component of social networks reveals the validity of values based on different meanings. The label associated with each social actor is an indication
that accompanies the social actor’s name, such as affiliation, the focus of study, and others. Meanwhile, labels relating to relationships are indications of relationships in social structures. Labels make it possible to express various social communities. Therefore, the labels of a social network are evidence for a variety of big data. The extraction, which is the task of data science, ensures the presence of related scientific fields to reveal this variety. Information extraction reveals the nature of the method for capturing the speed at which information changes in the information space through the speed of access. The methods specifically consider access times, processes, and changes optimally. Also, considering a large amount of data, the extraction capability by the method has constraints. For example, \( n \) social actor paired on a social network, there is \( n + \frac{1}{2}n(n - 1) = n + \frac{1}{2}n^2 - \frac{1}{2}n = \frac{1}{2}(n^2 + n) \) access symmetrically, where there are \( n \) singletons and \( \frac{1}{2}n(n - 1) \) doubletons. If the access limit to the information room, as a big data presentation, is only allowed 1000 times per day, or \( \frac{1}{2}(n^2 + n) = 1000 \), and a maximum of \( n = 44 \). The extraction task of data science has a maximum limit, and it depends on the method.

On the other hand, it has resulted in changes in other scientific fields and their scientific methods. Such as, there is no need to test the sample requirements for the information space. But what is important is to test the resultant correctness of the extraction method. As a consequence of the presence of big data is the emergence of facilities and infrastructure both in the form of systems and algorithms, which aim to make it easier to produce knowledge [35]. That means includes data science.

5. Conclusion
A review of the subject of study from data science, namely big data, has revealed several problems related to the character of big data that require special handling through interpretations of the potential of hardware and software. The extraction of social networks is a case that reveals that particular character. It also provides an overview of some of the interests related to the extraction method is as one of the tasks of data science.

References
[1] M K M Nasution, O S Sitompul, E B Nababan 2020 Data Science Journal of Physics: Conference Series 1566(1).
[2] M J Graham 2012 The art of data science Springer Series in Astrostatistics 2.
[3] K Ayankoya, A Calitz, J Greyling 2014 Intrinsic relations between data science, big data, business analytics and datification ACM International Conference Proceeding Series.
[4] C K-S Leung, F Jiang 2014 A data science solution for mining interesting patterns from uncertain big data Proceedings - 4th IEEE International Conference on Big Data and Cloud Computing, BDCloud 2014 with the 7th IEEE International Conference on Social Computing and Networking, SocialCom 2014 and the 4th International Conference on Sustainable Computing and Communications, SustainCom 2014.
[5] F Pop, V Cristea 2015 The art of scheduling for big data science. Big Data: Algorithms, Analytics, and Applications.
[6] K E Dierckx, A B Harrison, C K Leung, A V Pind 2017 A data science and engineering solution for fast k-means clustering of big data Proceedings - 16th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 11th IEEE International Conference on Big Data Science and Engineering and 14th IEEE International Conference on Embedded Software and Systems, Trustcom/BigDataSE/ICESS 2017.
[7] C E Concolato, L M Chen 2017 Data Science: A New Paradigm in the Age of Big-Data Science and Analytics New Mathematics and Natural Computation 13(2).
[8] 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.
[9] T. Berners-Lee, R. Cailliau, R., A Luotonen, H F Nielsen, A Secret 1994 The World-Wide Web Communications of the ACM 37(8).
[10] R-A Salas-Rueda, E-P Salas-Rueda, R-D Salas-Rueda 2020 Impact of the web application for the educational process on the compound interest considering data science Turkish Online Journal of Distance Education 21(3)7.
[11] M K M Nasution 2018 Singleton: A Role of the Search Engine to Reveal the Existence of Something in Information Space *IOP Conference Series: Materials Science and Engineering* **420**(1).

[12] M K M Nasution 2019 Doubleton: A Role of the Search Engine to Reveal the Existence of Relation in Information Space *IOP Conference Series: Materials Science and Engineering* **420**(1). 1, Issue 1, 7-26 (2003).

[13] V Dhar 2013 Data science and prediction *Communications of the ACM* **56**(12).

[14] M Houch, M Ganette 2019 Data, information, knowledge, and wisdom: A partnership to improve the management of forensic service organizations *Forensic Science International: Synergy* **1**. 1, Issue 1, 7-26 (2003).

[15] A V Dhar 2013 Data science and prediction *Communications of the ACM* **56**(12).

[16] M M Houck, M Gamette 2019 Data, information, knowledge, and wisdom: A partnership to improve the management of forensic service organizations *Forensic Science International: Synergy* **1**. 1, Issue 1, 7-26 (2003).

[17] V Jose, V P Jagathy Raj, S K George 2021 Ontology-based information extraction framework for academic knowledge repository *Advances in Intelligent Systems and Computing* **1184**.

[18] A V Dhar 2013 Data science and prediction *Communications of the ACM* **56**(12).

[19] M K M Nasution 2018 Singleton: A Role of the Search Engine to Reveal the Existence of Something in Information Space *IOP Conference Series: Materials Science and Engineering* **420**(1).

[20] M K M Nasution 2019 Doubleton: A Role of the Search Engine to Reveal the Existence of Relation in Information Space *IOP Conference Series: Materials Science and Engineering* **420**(1). 1, Issue 1, 7-26 (2003).

[21] V Dhar 2013 Data science and prediction *Communications of the ACM* **56**(12).

[22] M K M Nasution 2018 Singleton: A Role of the Search Engine to Reveal the Existence of Something in Information Space *IOP Conference Series: Materials Science and Engineering* **420**(1).

[23] V Dhar 2013 Data science and prediction *Communications of the ACM* **56**(12).

[24] M K M Nasution 2018 Singleton: A Role of the Search Engine to Reveal the Existence of Something in Information Space *IOP Conference Series: Materials Science and Engineering* **420**(1).