Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2021, 8(8):74-79**

Review Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Big Data in Natural Sciences

Matthew N. O. Sadiku[1], Uwakwe C. Chukwu[2], Abayomi Ajayi-Majebi[3], Sarhan M. Musa[1]

[1]Roy G. Perry College of Engineering, Prairie View A&M University, Prairie View, TX, USA
[2]Department of Engineering Technology, South Carolina State University, Orangeburg, SC, USA
[3]Department of Manufacturing Engineering, Central State University, P.O. Box 1004, Wilberforce, OH, USA
Email: sadiku@ieee.org; uchukwu@scsu.edu; ajayi-majebi@centralstate.edu; smmusa@pvamu.edu

Abstract Big data is a phenomenon revolving around the endeavor of accumulating massive amounts of data. It has recently become the buzzword and is helping many organizations to become data-driven. Scientific big data will become a new solution in scientific research as the paradigm changes from being model-driven to data-driven. Government agencies and privates sector must prepare for a new decade where big data is the norm rather than the exception. This paper is a primer on the application of big data in natural sciences.

Keywords big data, data analytics, natural sciences, physics, chemistry, biology

Introduction

Data is the raw information collected from any study, while data science is the study of this data. Data plays vital role in every field. The units of data are illustrated in Figure 1 [1]. Big data refers to huge data sets with such complexity that it would be impossible to process, manipulate, and present using traditional database and software tools. It is an industry jargon that denotes a vast amount of information currently being generated by websites, social media, network sensors, etc.

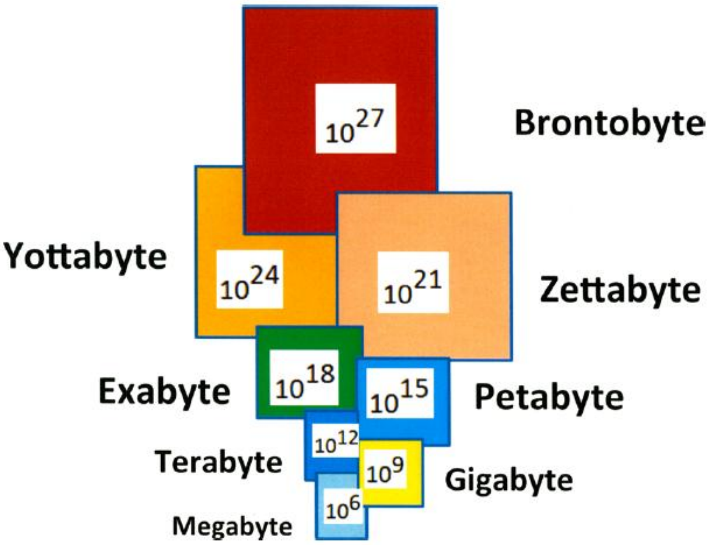

Figure 1: Data units in terms of bytes [1].

Big data is used in several areas such as academia, business, finance, government, healthcare, engineering, manufacturing, agriculture, social media, tourism, industry, etc. It is now shaping several fields including sciences. The word “science” denotes the deepening of and distribution of knowledge of a subject. When the subject is a natural phenomenon, we call it natural science; if the subject is society we call it social science [2]. Natural sciences are empirical sciences in that knowledge must be based on observable phenomena that can be verified by other researchers working under the same conditions. They are a subset of science concerned with the description, prediction, and understanding of natural phenomena. Essentially, natural sciences consist of physics, biology, and chemistry. They seek to understand how the universe around us works. Big data science has been described as a “fourth paradigm” of science, after experimental, theoretical, and computational paradigms [3,4].

Review on Big Data

Big data (BD) refers to a collection of data that cannot be captured, managed, and processed by conventional software tools. It is a relatively new technology that can help many industries. The three main sources of big data are machines, people, and companies. Big data can be described with 42 Vs [5]. The first five Vs are volume, velocity, variety, veracity, and value [6].

- *Volume*: This refers to the size of the data being generated both inside and outside organizations and is increasing annually. Some regard big data as data over one petabyte in volume.
- *Velocity*: This depicts the unprecedented speed at which data are generated by Internet users, mobile users, social media, etc. Data are generated and processed in a fast way to extract useful, relevant information. Big data could be analyzed in real time, and it has movement and velocity.
- *Variety*: This refers to the data types since big data may originate from heterogeneous sources and is in different formats (e.g., videos, images, audio, text, logs). BD comprises of structured, semi-structured or unstructured data.
- *Veracity*: By this, we mean the truthfulness of data, i.e. weather the data comes from a reputable, trustworthy, authentic, and accountable source. It suggests the inconsistency in the quality of different sources of big data. The data may not be 100% correct.
- *Value*: This is the most important aspect of the big data. It is the desired outcome of big data processing. It refers to the process of discovering hidden values from large datasets. It denotes the value derived from the analysis of the existing data. If one cannot extract some business value from the data, there is no use managing and storing it.

On this basis, small data can be regarded as having low volume, low velocity, low variety, low veracity, and low value. Additional five Vs has been added [7]:

- *Validity:* This refers to the accuracy and correctness of data. It also indicates how up to date it is.
- *Viability:* This identifies the relevancy of data for each use case. Relevancy of data is required to maintain the desired and accurate outcome through analytical and predictive measures.
- *Volatility:* Since data are generated and change at a rapid rate, volatility determines how quickly data change.
- *Vulnerability:* The vulnerability of data is essential because privacy and security are of utmost importance for personal data.
- *Visualization:* Data needs to be presented unambiguously and attractively to the user. Proper visualization of large and complex clinical reports helps in finding valuable insights.

Figure 2 shows the 10V’s of big data. In addition, the 10V’s above, some suggest the following 5V’s: Venue, Variability, Vocabulary, Vagueness, and Validity) [8].

To thrive in today’s complex business environment, businesses must adopt a data-driven culture and leverage analytics platforms to make key decisions that improve productivity. Industries that benefit from big data include the healthcare, financial, airline, travel, restaurants, automobile, sports, agriculture, manufacturing, and hospitality industries.

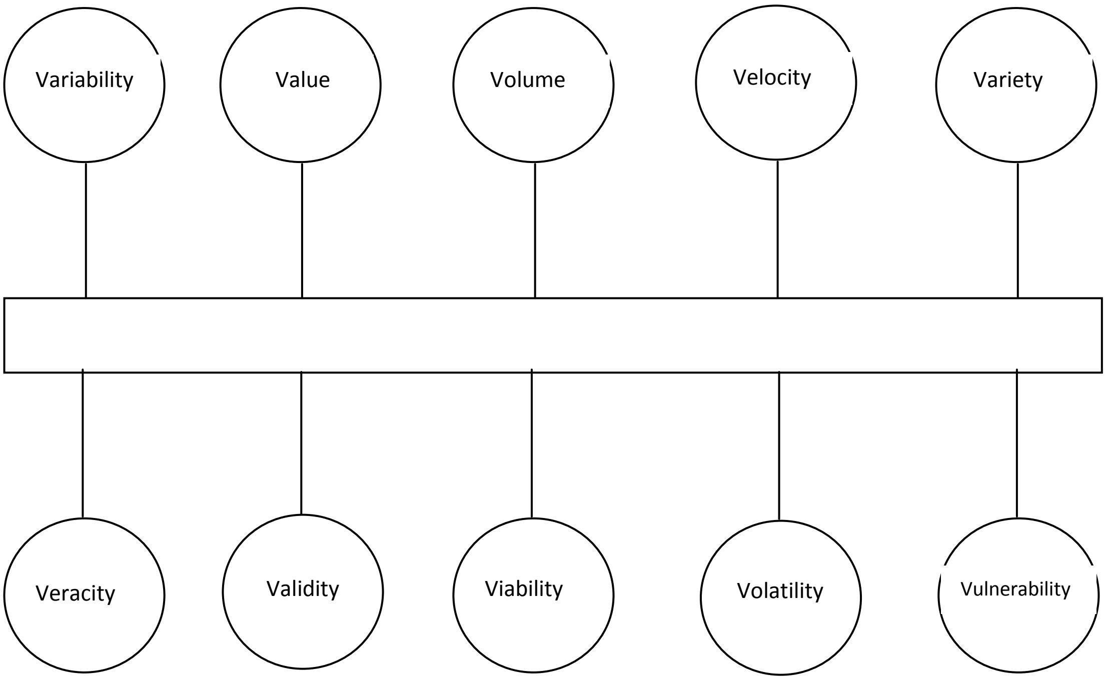

Figure 2: The 10V's of big data

Big Data Analytics

Every day, data is growing bigger and bigger, and big data analysis (BDA) has become a requirement for gaining invaluable insights into data such that companies could gain significant profits in the global market. Once the big data is ready for analysis, we use advanced software programs such as Hadoop, MapReduce, MongoDB, Spark, Cassandra, Apache Storm, and NoSQL databases [9]. Big data analytics refers to how we can extract, validate, translate, and utilize big data as a new currency of information transactions. It is an emerging field that is aimed at creating empirical predictions. Data-driven organizations use analytics to guide decisions at all levels [10].

Big data analytics is capable of processing massive amounts of dirty data and extract the gold information from it. It has the potential to predict performance, upcoming changes, and market trends with unprecedented accuracy. To maximize the value of data and make it useful to end users, natural scientists should become familiar with data analytics.

Big Data in Biology

The data is composed from a number of commencements of biology like bioinformatics in which data mountains. Biological data are much more heterogeneous than those in physics. They stem from a wide range of experiments. Much progress has been made in big-data biology and data handling [11]. Someone has predicted that every sector of human endeavor will soon emulate biology's example of identifying data-driven research and experiencing data-driven revolutions. In addition to biology, big data approaches can be applied to health data and health records.

Big Data in Chemistry

In recent years, the chemical industry has been facing increasingly severe data analytics challenges, which partly have to do with extracting valuable information from a wealth of raw data. Chemical companies use information technology and data gathered from different sources to improve the way they do business, support growth, and compete in the marketplace. Like other industries, the chemical industry is gathering more data (volume) from different sources (variety). A large volume of data is available to characterize products, customers, and operations [12].

Big Data in Physics

Science is growing rapidly in increasing capabilities of the computers and software tools used to handling big data. Physicists are joining the big-data club. Within the physics world, a lot of data comes from quantum systems and needs to be analyzed. The particle physics community has a long history of employing big data analysis. Data sets within computational physics are created with the sole intent of being examined.

From the computational perspective, data are simply too large to hold in memory or to store altogether. Predictive analytics can be used as a tool to calculate soft-body physics [13].

Applications

In general, application domains in big data is illustrated in Figure 3 [14].

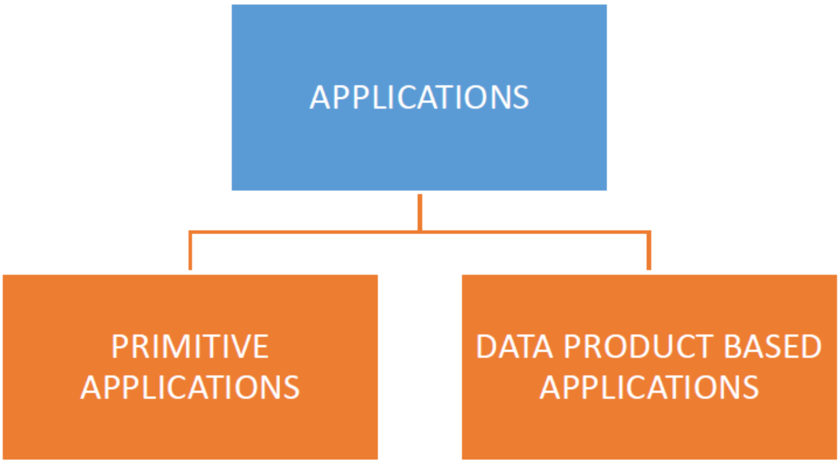

Figure 3: Categorization of application domains [14]

The natural sciences abound in examples of big data. The big data application may serve as mediators in natural science learning.

- *Scientific big data:* This will become a new solution in scientific research. It has a number of characteristics such as complexity, comprehensiveness, globalization, and high integration of information and communication technology. The scientific approaches are also transforming from single disciplines to multidisciplinary and interdisciplinary; from natural science to the integration of natural and social sciences. Scientific big data has made changes in the scientific world, and research has entered into a new paradigm—a paradigm of data-intensive science [15]**.**
- *Environmental science:* An important application of big data in environmental science is “Citizen Science.” This is the accumulation of data reported from people in geographic locations all over the world voluntarily offering information on conditions where they live. Big data can also be applied to examine problems areas for traffic, crime centers, health problems, pollution, poverty, etc. [16].
- *Food science*: This is the discipline that applies basic sciences and engineering to study the nature of foods and their harvesting, processing, distribution, storage, and preparation. It is essential to meeting the needs of a growing global population. It has a significant role to play in achieving food and nutrition security. Food security demands that we increase the amount of available food. Natural preservation through fermentation and separation technologies, such as forward osmosis, offer the potential to create new value-added food ingredients and bioactives from food loss and food waste [17,18].

Benefits and Challenges

The benefits of big data in natural sciences are numerous. Big Data and the analytical tools can process and analyze more past data than ever before. These also enable organizations to make data-driven decisions. Scientists can conceive, design, and implement their research based on the data.

Other benefits include [19]:

- Assess risk
- Assess effectiveness in clinical trials
- Personalize medicine
- Improve R&D efficiencies
- Assist in price control, budgeting research, and profit forecasts\
- Analyze wearable, implantable, and remote data
- Predict virus evolution

Big data also helps to optimize innovation; improve the efficiency of clinical trials and research; build new tools for consumers, physicians and other players in the industry; help regulators to come up with more individualized approaches

However, big data is not designed to be a one-size-fits-all answer. Like any other emerging technology, there are challenges and limitations to using big data in natural sciences. Big data brings complex issues to many organizations. Since some sciences concern data pertaining to humans and sciences are a moral undertaking, big data sometimes have ethical limitations. Therefore, it is important that data science be used ethically [20]. Although computers can perform just about every act conceivable, but the creativity, intuition and instincts involved in the process of developing theories and guiding their applications, is still beyond reach of the computer. Thus, human scientists are still necessary [21].

Conclusion

We live in the era of big data, and big data is hitting us from all angles. How we handle the emergence of an era of big data is critical. Every scientific discipline must find a way to tackle challenges in storing, handling, and interpreting large amounts of raw data. Big data is poised to change natural science. Big data has increased exponentially due to the rapid evolution of new technologies, devices, and communication means. To keep up, hardware in all areas of applications will need upgrading.

Big data is here to stay. Natural sciences will continue to require big data analytics [22]. How we handle the emergence of an era of big data is critical. Organizations that capitalize on big data stand apart from traditional data analysis environments. It is not hard to imagine a future in which the education of natural scientists will consist of teaching various disciplines such as programming, data analysis, and advanced statistics. More information about big data in natural sciences can be found in the book in [23].

References

[1]. H. E. Pence and A. J. Williams, “Big data and chemical education,” *Journal of Chemical Education*, vol. 93, 2016, pp. 504-508.

[2]. S. Ikeda, “Data science and natural science,” *Kavli IPMU News*, no. 36, December 2016, https://www.ipmu.jp/sites/default/files/imce/news/36E_ResearchReport.pdf

[3]. C. Wiggins, “Data science in the natural sciences,” November 2012, http://radar.oreilly.com/2012/11/data-science-natural-sciences.html

[4]. “Branches of science,” *Wikipedia*, the free encyclopedia, https://en.wikipedia.org/wiki/Branches_of_science

[5]. “The 42 V’s of big data and data science,” https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html

[6]. M. N. O. Sadiku, M. Tembely, and S. M. Musa, ”Big data: An introduction for engineers,” *Journal of Scientific and Engineering Research*, vol. 3, no. 2, 2016, pp. 106-108.

[7]. P. K. D. Pramanik, S. Pal, and M. Mukhopadhyay, “Healthcare big data: A comprehensive overview,” in N. Bouchemal (ed.), *Intelligent Systems for Healthcare Management and Delivery*. IGI Global, chapter 4, 2019, pp. 72-100.

[8]. J. Moorthy et al., “Big data: Prospects and challenges,” *The Journal for Decision Makers*, vol. 40, no. 1, 2015, pp. 74–96. https://www.grandviewresearch.com/industry-analysis/industrial-wireless-sensor-networks-iwsn-market

[9]. M. N. O. Sadiku, J. Foreman, and S. M. Musa, “Big data analytics: A primer,” *International Journal of Technologies and Management Research,* vol. 5, no. 9, September 2018, pp. 44-49.

[10]. C. M. M. Kotteti, M. N. O. Sadiku, and S. M. Musa, “Big data analytics,” *Invention Journal of Research Technology in Engineering & Management*, vol. 2, no. 10, Oct. 2018, pp. 2455-3689.

[11]. “The big challenges of big data in natural sciences big data in natural sciences,” https://www.fisheriesindia.com/2020/10/big-challenges-of-big-data-in-natural.html

[12]. M. N. O. Sadiku, S. M. Musa, and O. S. Musa, ”Big data in the chemical industry,” *International Journal of Advances in Scientific Research and Engineering,* vol. 3, no. 10, Nov. 2017, pp. 20-23.

[13]. P. Clarke et al., “Big data in the physical sciences: Challenges and opportunities,” https://indico.cern.ch/event/449964/attachments/1253648/1849677/Bigdata_physicalsciences.pdf

[14]. M. Sahoo, S. S. Rautaray, and M. Pandey, “The emergence of big data: A survey,” *International Journal of Computer Science and Mobile Applications,* vol.6, no. 6, June 2018, pp. 23-32.

[15]. H. Guo et al., “Scientific big data and digital earth,” *Chinese Science Bulletin,* vol. 59, no. 35, 2014, pp. 5066–5073.

[16]. “Big data: Explaining its uses to environmental sciences,” https://www.environmentalscience.org/data-science-big-data

[17]. M. N. O. Sadiku, T. J. Ashaolu, and S M. Musa, "Food science: A primer," *International Journal of Trend in Scientific Research and Development,* vol. 3, no. 4, May-June 2019, pp. 839-841.

[18]. M. B. Cole, et al., “The science of food security,” *npj Science of Food, vol. 2, no.*14, 2018.

[19]. “Life sciences data analytics: The importance of big data,” October 2019, https://treximo.com/blog/big-data-life-science-industries/

[20]. “Big data: Changing the way we do science,” May 2018, https://www.enago.com/academy/big-data-changing-the-way-we-do-science/

[21]. H. S. Sætra, “Science as a vocation in the era of big data: The philosophy of science behind big data and humanity’s continued part in science,” *Integrative Psychological and Behavioral Science*, vol. 52, no. 7, December 2018, pp. 508–522.

[22]. R. P. dos Santos, “Big data: Philosophy, emergence, crowdledge, and science education,” *Themes in Science and Technology Education*, vol. 8, no. 2, 2015, pp. 115-127.

[23]. S. C. Lewis (ed.), *Journalism in an Era of Big Data: Cases, Concepts, and Critiques.* Routledge, 2018.

About Authors

Matthew N. O. Sadiku is a professor emeritus in the Department of Electrical and Computer Engineering at Prairie View A&M University, Prairie View, Texas. He is the author of several books and papers. His areas of research interest include computational electromagnetics and computer networks. He is a fellow of IEEE.

Uwakwe C. Chukwu is an associate professor in the Department of Industrial & Electrical Engineering Technology of South Carolina State University. He has published several books and papers. His research interests are power systems, smart grid, V2G, energy scavenging, renewable energies, and microgrids.

Abayomi Ajayi-Majebi is a professor in the Department of Manufacturing Engineering at Central State University in Wilberforce, Ohio. In 2015 he was honored by the White House as a Champion of Change for his significant contributions to the engineering education of minority students. He is a senior member of both the Society of Manufacturing Engineers and the American Society for Quality.

Sarhan M. Musa is a professor in the Department Electrical and Computer Engineering at Prairie View A&M University, Texas. He has been the director of Prairie View Networking Academy, Texas, since 2004. He is an LTD Sprint and Boeing Welliver Fellow. His areas of research interest include computational electromagnetics and computer networks.