Asian Journal of Information Technology 18 (3): 99-105, 2019
ISSN: 1682-3915

Role of Big Data in Chemistry Education Research

Florence O. Ezeudu, Ijeoma H.N. Nwoji and Anselem Abonyi Ugwuanyi
Department of Science Education, University of Nigeria, Nsukka, Nigeria

Abstract: This study examined the potential role of big data in chemistry education research. Big data refer to data sets that are too large or complex for traditional data processing application software to adequately deal with. With large amounts of information streaming in from countless sources, chemistry education researchers are faced with finding new and innovative ways to manage big data. Chemistry education researchers armed with data-driven insight can make a significant impact on school systems, students and curriculums. By analyzing big data, they can identify at-risk students, make sure students are making adequate progress and can implement a better system for evaluation and support of chemistry teachers and school administrators. With big data, chemistry education researchers and educators can identify root causes of chemistry student's failures issues and defects in near-real time and detecting fraudulent behaviour among chemistry students among others.

Key words: Big data, big data analytics, chemistry research, complex, large, adequate

INTRODUCTION

The term big data has launched a veritable industry of processes, personnel and technology to support what appears to be an exploding new field. The use of big data has become well established in business, entertainment, science, technology, engineering and educational institutions. Educational institutions use big data to meet their educational goals (Matteson, 2013). Generally, big data is large datasets and the category of computing strategies and technologies that are used to handle large datasets. In this context, large dataset means a dataset too large to reasonably process or store with traditional tooling or on a single computer. The common scale of big datasets is constantly shifting and may vary significantly from organization to organization (Ellingwood, 2016). Big data is a blanket term for the non-traditional strategies and technologies needed to gather organize, process and gather insights from large datasets. While the problem of working with data that exceeds the computing power or storage of a single computer is not new, the pervasiveness, scale and value of this type of computing have greatly expanded in recent years (Ellingwood, 2016). Big data is the information owned by a company, educational institutions, obtained and processed through new techniques to produce value in the best way possible (Matteson, 2013). Data with many cases (rows) offer greater statistical power while data with higher complexity (more attributes or columns) may lead to a higher false discovery rate (Breur, 2016). Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating information privacy and data source. The amount of computerized information that educational institutions collect and the process is growing, so, large that the term big data is commonly being used to describe the situation. Accordingly, big data is defined by a combination of the volume, variety, velocity and veracity of the data being processed.

Big data tools are already having an impact on the chemical industry and research (Pence and Williams, 2016). From the perspectives of Mills *et al.* (2012) and Sicular (2013), big data is a term that is used to describe data that is high in volume, high in velocity and with a high variety; requires new technologies and techniques to capture, store and analyze it and is used to enhance decision making, provide insight and discovery and support and optimize processes. Big data is an umbrella term for datasets that cannot reasonably be handled by traditional computers or tools due to their volume, velocity and variety. This term is also typically applied to technologies and strategies to work with this type of data. This implies that big data is a lot of data produced very quickly in many different forms. This may involve production databases, web traffic logs, online videos, social media interactions etc. Big data was originally associated with three key concepts: volume, variety and velocity (Laney, 2001). Other concepts later attributed to big data are veracity (i.e., how much noise is in the data) and value (Goes, 2014; Bernard, 2014). Laney (2001) first presented what became known as the three Vs of big data to describe some of the characteristics that make big data different from other data processing, they include volume, velocity and variety.

The increasing volume of data in chemistry education requires the development of new methods and approaches for their handling. Big data in chemistry education refers to considerably larger databases than commonly used ones in orders of magnitude which become recently available (Tetko *et al.*, 2016). In addition, big data in

Corresponding Author: Ijeoma H.N. Nwoji, Department of Science Education, University of Nigeria, Nsukka, Nigeria

Table 1: Web-based data types

Database	Main data types
ChEMBL V.21	Data mined from literature
PubChem	HTS assays
BindingDB	Experimental protein-small molecule interaction data
PubChem	Bioactivity data from HTS assays
Reaxys	Literature mined property, activity and reaction data
SciFinder (CAS)	Experimental properties, 13C and 1H NMR spectra, reaction data
GOSTAR	Target-linked data from patents and articles
AZ IBIS	AZ in-house SAR data points
OCHEM	Mainly ADMET data collected from the literature

chemistry education refers to humongous volumes of data that cannot be processed effectively with the traditional applications that exist. The processing of big data begins with the raw data that isn't aggregated and is most often impossible to store in the memory of a single computer. This implies that in chemistry education research, big data is a buzzword that is used to describe immense volumes of data, both unstructured and structured which can be used to analyze insights which can lead to better decisions and strategic educational moves. How to efficiently mine the large scale of data in chemistry education becomes an important problem for the future development of the educational sector. Furthermore in chemistry education, big data-simply means large unstructured data. Education sector now a day's have become technology oriented. There is a lot of Massive Open Online Course (MOOC) which are generating a huge amount of data. Big data in chemistry education is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage and process the data with low-latency. And it has one or more of the following characteristics high volume, high velocity or high variety. Big data in chemistry education comes from sensors, devices, student's record, video/audio, networks, log files, transactional applications, web and social media-much of it generated in real time and in a very large scale.

Web-based data types in chemistry education research: Web-based data types according to researchers (Papadatos *et al*., 2015; Gilson *et al*., 2016; Kim *et al*., 2015; Muresan *et al*., 2011; Sushko *et al*., 2011) are shown in Table 1.

MATERIALS AND METHODS

Process involved in big data: While approaches to implementation differ, there are some commonalities in the strategies and software generally. While the steps presented below might not be true in all cases, they are widely used. The general categories of activities involved with big data processing according to Ellingwood (2016) are:

- Ingesting data into the system
- Persisting the data in storage
- Computing and analyzing data
- Visualizing the results

Ingesting data into the system: Data ingestion is the process of taking raw data and adding it to the system. The complexity of this operation depends heavily on the format and quality of the data sources and how far the data is from the desired state prior to processing. One way that data can be added to a big data system is dedicated to ingestion tools. Technologies like Apache Sqoop can take existing data from relational databases and add it to a big data system. Similarly Apache Flume and Apache Chukwa are projects designed to aggregate and import application and server logs. Queuing systems like Apache Kafka can also be used as an interface between various data generators and a big data system. Ingestion frameworks like Gobblin can help to aggregate and normalize the output of these tools at the end of the ingestion pipeline.

During the ingestion process, some level of analysis, sorting and labelling usually takes place. This process is sometimes called ETL which stands for extract, transform and load. While this term conventionally refers to legacy data warehousing processes, some of the same concepts apply to data entering the big data system. Typical operations might include modifying the incoming data to format it, categorizing and labelling data, filtering out unneeded or bad data or potentially validating that it adheres to certain requirements. With those capabilities in mind, ideally, the captured data should be kept as raw as possible for greater flexibility further on down the pipeline.

Persisting the data in storage: The ingestion processes typically hand the data off to the components that manage storage, so that, it can be reliably persisted to disk. While this seems like it would be a simple operation, the volume of incoming data, the requirements for availability and the distributed computing layer make more complex storage systems necessary. This usually means leveraging a distributed file system for raw data storage. Solutions like Apache Hadoop's HDFS file system allow large quantities of data to be written across multiple nodes in the cluster. This ensures that the data can be accessed by computing resources can be loaded into the cluster's RAM for in-memory operations and can gracefully handle component failures. Other distributed file systems can be used in place of HDFS including Ceph and GlusterFS.

Data can also be imported into other distributed systems for more structured access. Distributed databases, especially, NoSQL databases are well-suited for this role because they are often designed with the same fault-tolerant considerations and can handle heterogeneous data. There are many different types of distributed databases to choose from depending on how you want to organize and present the data.

Computing and analyzing data: Once the data is available, the system can begin processing the data to surface actual information. The computation layer is perhaps the most diverse part of the system as the requirements and best approach can vary significantly depending on what type of insights desired. Data is often processed repeatedly, either iteratively by a single tool or by using a number of tools to surface different types of insights. Batch processing is one method of computing over a large dataset. The process involves breaking work up into smaller pieces, scheduling each piece on an individual machine, reshuffling the data based on the intermediate results and then calculating and assembling the final result. These steps are often referred to individually as splitting, mapping, shuffling, reducing and assembling or collectively as a distributed map reduce algorithm. This is the strategy used by Apache Hadoop's MapReduce. Batch processing is most useful when dealing with very large datasets that require quite a bit of computation.

While batch processing is a good fit for certain types of data and computation, other workloads require more real-time processing. Real-time processing demands that information be processed and made ready immediately and requires the system to react as new information becomes available. One way of achieving this is stream processing which operates on a continuous stream of data composed of individual items. Another common characteristic of real-time processors is in-memory computing which works with representations of the data in the cluster's memory to avoid having to write back to disk. Apache Storm, Apache Flink and Apache Spark provide different ways of achieving real-time or near real-time processing. There are trade-offs with each of these technologies which can affect which approach is best for any individual problem. In general, real-time processing is best suited for analyzing smaller chunks of data that are changing or being added to the system rapidly. The above examples represent computational frameworks. However, there are many other ways of computing over or analyzing data within a big data system. These tools frequently plug into the above frameworks and provide additional interfaces for interacting with the underlying layers. For instance, Apache Hive provides a data warehouse interface for Hadoop, Apache Pig provides a high-level querying interface while SQL-like interactions with data can be achieved with projects like Apache Drill, Apache Impala, Apache Spark SQL and Presto. For machine learning, projects like Apache SystemML, Apache Mahout and Apache Spark's MLlib can be useful. For straight analytics programming that has wide support in the big data ecosystem, both R and Python are popular choices.

Visualizing the results: Due to the type of information being processed in big data systems, recognizing trends or changes in data over time is often more important than the values themselves. Visualizing data is one of the most useful ways to spot trends and make sense of a large number of data points. Real-time processing is frequently used to visualize application and server metrics. The data changes frequently and large deltas in the metrics typically indicate significant impacts on the health of the systems or organization. In these cases, projects like Prometheus can be useful for processing the data streams as a time series database and visualizing that information. One popular way of visualizing data is with the Elastic stack, formerly known as the ELK stack. Composed of Logstash for data collection, Elastic search for indexing data and Kibana for visualization, the Elastic stack can be used with big data systems to visually interface with the results of calculations or raw metrics. A similar stack can be achieved using Apache Solr for indexing and a Kibana fork called Banana for visualization. The stack created by these is called Silk. Another visualization technology typically used for interactive data science work is a data notebook. These projects allow for interactive exploration and visualization of the data in a format conducive to sharing, presenting or collaborating. Popular examples of this type of visualization interface are Jupyter Notebook and Apache Zeppelin (Ellingwood, 2016).

Today, many educational institutions are collecting, storing and analyzing massive amounts of data. This data is commonly referred to as big data because of its volume, the velocity with which it arrives and the variety of forms it takes. Big data is creating a new generation of decision support data management. Businesses, educational institutions are recognizing the potential value of this data and are putting the technologies, people and processes in place to capitalize on the opportunities. A key to deriving value from big data is the use of analytics. Collecting and storing big data creates little value; it is only data infrastructure at this point. It must be analyzed and the results used by decision makers and organizational processes in order to generate value (Watson, 2014).

Big data analytics: By itself, stored data does not generate academic value and this is true of traditional databases, data warehouses and new technologies such as Hadoop for storing big data. Once the data is appropriately stored, however, it can be analyzed which can create tremendous value. Big data analytics is the term used to describe the examination of large amounts of data to see what patterns or other useful information that can be found (Whiting, 2018). In addition, big data analytics is the process of examining large data sets containing a variety of data types; to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful information (Galetto, 2016). Big data analytics gives analytics professionals such as data scientists, educational researchers and predictive modellers, the ability to analyze big data from multiple and varied sources including transactional data and other structured data. A variety of analysis technologies,

approaches and products have emerged that are especially, applicable to big data such as in-memory analytics in-database analytics and appliances.

In addition, data analytics involves the science of examining raw data with the purpose of drawing conclusions about that information. Data analytics involves applying an algorithmic or mechanical process to derive insights (Monnappa, 2019). For example, running through a series of data sets to look for meaningful correlations between each other. It is used in a number of industries to allow the organizations and companies to make better decisions as well as verify and disprove existing theories or models (Monnappa, 2019). This implies that data analytics is used in educational institutions to allow educational institutions to make better decisions that will bring about effective teaching and learning. The focus of data analytics lies in inference which is the process of deriving conclusions that are solely based on what the researcher already knows. Artificial Intelligence (AI), mobile, social and Internet of Things (IoT) are driving data complexity, new forms and sources of data. Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data, from different sources and in different sizes from terabytes to zettabytes. Analyzing big data allows analysts, researchers and business users to make better and faster decisions using data that was previously inaccessible or unusable. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, statistics and natural language processing, educational institutions can analyze previously untapped data sources independent or together with their existing enterprise data to gain new insights resulting in better and faster decisions on students progress.

RESULTS AND DISCUSSION

Role of big data in chemistry education research: The role of big data in chemistry education research doesn't revolve around how much data available but the utilization of the data. With large amounts of information streaming in from countless sources, educational institutions are faced with finding new and innovative ways to manage big data. Educators armed with data-driven insight can make a significant impact on school systems, students and curriculums. By analyzing big data, they can identify at-risk students, make sure students are making adequate progress and can implement a better system for evaluation and support of teachers and principals.

Big data is used to enhance decision making, provide insight, discovery, support and optimize processes in chemistry education research (Fig. 1). In

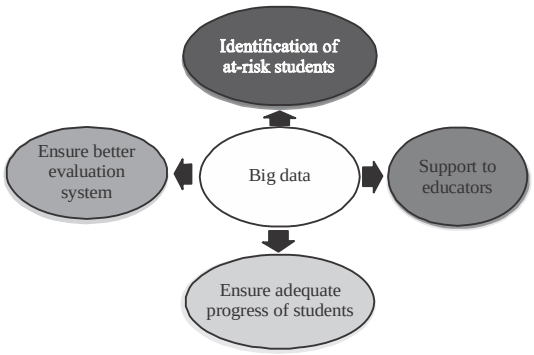

Fig. 1: The benefits of big data when applied in chemistry education

chemistry education research, big data enables decision support data management. Big data can be analyzed for insights that lead to better decisions and strategic educational moves. By analyzing big data, researchers, educators can identify root causes of student's failures issues and defects in near-real time and detecting fraudulent behaviour of students. When big data is managed effectively, teachers, researchers, educators can uncover hidden insights that can improve student's achievement and the educational institution. Students and teachers relationship building is critical in the educational sector and the best way to manage that is to manage big data. Teachers need to know the best way to communicate with the students and the most effective way to bring about effective teaching and learning and the most strategic way to bring back slow learners. Big data remains at the heart of all those things.

Technological and methodological advances have enabled an unprecedented capability for decision making based on big data in chemistry education. Big data is beginning to be utilized for decision making in chemistry education research (Dede *et al*., 2016). In addition, big data helps in collecting and analyzing student data by using much of the computational infrastructure, tools and human capacity required for effective collection, cleaning, analysis and distribution of large datasets, big data helps to measure these outcomes. Beyond the potential to enhance student outcomes through just-in-time, diagnostic data that is formative for learning and instruction, the evolution of higher education practice could be substantially enhanced through data-intensive research and analysis. Big data in chemistry education research potentially provide a variety of opportunities to improve student learning by individualizing a student's path to content mastery, through, adaptive learning or competency-based education. Big data potentially provide a variety of opportunities for better learning as a result of faster and more in-depth diagnosis of learning needs or course trouble spots including assessment of skills such as

systems thinking, collaboration and problem-solving in the context of deep, authentic subject-area knowledge assessments. Big data in chemistry education research enables using game-based environments for learning and assessment where learning is situated in complex information and decision-making situations; targeted interventions to improve student success and to reduce overall costs to students and institutions.

The value of big data in assessing complex skills, conventional assessments in education classrooms are infrequent and constrained, both in their design (e.g., essay prompts, multiple-choice questions) and in their feedback (which is usually delayed and sometimes subjective). Progress in educational technology can provide tools for measuring student's performance on more authentic tasks such as engineering design problems and free-form text answers. Measuring these types of tasks can increase the relevance and the precision of the results regarding what students learn can allow the tailoring of instruction to specific student's needs and can give individualized feedback across a range of learning issues. Social interactions have increasingly moved from in-person to online. Big data can include detailed traces of student-to-student interactions. By integrating these and other sources of data, one may be able to measure more complex problem-solving and collaborative skills (Dede *et al*., 2016). Furthermore, researchers can collect fine-grained data about the actions of an individual student. This data can provide specifics about learning trajectories from both correct and incorrect answers and about the actions taken to get there. Extensive research shows differences in the problem-solving strategies of novices and experts. Experts can chunk information; for example, an expert looking at an analogue circuit will be able to remember that circuit whereas a novice will not, likely leading to differential patterns of behaviour such as scrolling. Data from rich assessments may provide information on the development of such expertise. We can also record how many times a student flips between pages of a problem set or looks up equations in a textbook and we can then investigate which of these variables contain data that can act as proxies for expertise.

In chemistry education research, many MOOCs include data from students who are completing large numbers of complex problems as part of their regular coursework. For example, 6.002x the first edX/MITx course, used assessments that consisted entirely of relevant performance tasks. Students completed design-and-analysis problems that required answers written as equations, numbers or circuits. Since, these types of questions have a near-infinite number of possible solutions, answers cannot be guessed. Students could submit an answer as many times as necessary in order to gain feedback and eventually solve a problem. The assessments were time-consuming: most weeks of the course had just four assessments but completing those four required 10-20 h of work. Similarly, relevant performance assessments have been used in courses such as chemistry, biology, physics and computer science. Such complex assessments if pooled for a given student across many courses, can give rich data about problem-solving skills and collaborative activity. Also as increased amounts of digital group work are introduced into courses, more traces of social activity appear in server log files. These logs can help to identify students who underperform or overperform in group tasks and can directly measure individual student's contributions to the group. These systems may provide enough data to begin to look for specific actions and patterns that lead to good overall group performance. Feedback can be provided to students by using these patterns to improve group performance. Natural language processing frameworks such as the open-source edX EASE and Discern are still used primarily for short-answer grading but they were designed to apply also to the analysis of social activities such as emails and forum posts. These frameworks promise to provide insights into writing processes and group dynamics. In chemistry education, big data is collected organized and analyzed for relevant information; the aim is to find patterns, correlations and information that can help educational institutions make important decisions. Scientists, modellers and many others in the analytics field use big data analytics to sift through large amounts of data that can come from a variety of sources like transactions, web servers, social media, surveys and even emails. Big data enables researchers to analyze their data in full context quickly and some offer real-time analysis. With high-performance data mining, predictive analytics, text mining, forecasting and optimization, the educational sector that utilize big data analytics are able to drive innovation and make the best decisions. This proactive approach is transformative because it gives analysts and decision makers the power to move ahead with the best knowledge and insights available, often in real time. This means that educational sector can improve their student's achievement, develop better and effective teaching and learning atmosphere and gain a competitive advantage by taking rapid action to respond to students need and other metrics that impact on the learners. In chemistry education research, utilizing big data with fidelity also have the ability to boost student's achievement, discover students in difficulty, new opportunities, improve student's achievement and reduce risk. Finally, aside from looking within individual courses, MOOC data systems allow longitudinal analysis across a student's educational trajectory. In most cases, a single time point does not provide interesting information about learning. However, reviewing all of the projects over the duration of a student's education can provide

more precise estimates of learning and proficiency. Data can be taken from any source, analyze to find answers that enable:

- Cost reductions
- Time reductions
- New product development
- Smart decision making

RECOMMENDATIONS

In chemistry education research, the amount of data generation increases day by day. So, handling this huge amount of data becomes a challenge. In chemistry education research, big data refers to large unstructured data. For big data management, researchers are encouraged to use various new databases like NOSQL. For the education data, a lot of analysis and researches are advised to handle the data with the proper techniques invent new tools to generate a real-time solution, prediction in the education sector. Researchers are encouraged and should be exposed to the use of big data. It is recommended that teachers should use big data, since, by analyzing big data, can identify at-risk students, student's progress and can implement a better system for evaluation and support of teachers and principals. By analyzing big data, researchers, educators can identify root causes of student's failures issues and defects in near-real time and detecting fraudulent behaviour of students. When big data is managed effectively, teachers, researchers, educators can uncover hidden insights that can improve student's achievement and the educational institution.

CONCLUSION

This review has shown that the role of big data in chemistry education research cannot be overemphasized. Chemistry education researchers armed with data-driven insight can make a significant impact on school systems, students and curriculums. By analyzing big data, they can identify at-risk students, make sure students are making adequate progress and can implement a better system for evaluation and support of teachers and school administrators. When big data is managed effectively, chemistry education researchers and educators can uncover hidden insights that can improve chemistry student's achievement and performance of the educational institution.

REFERENCES

Bernard, B., 2014. Big data: The 5 Vs everyone must know. LinkedIn Corporation, Sunnyvale, California, USA.

Breur, T., 2016. Statistical power analysis and the contemporary crisis in social sciences. J. Marketing Anal., 4: 61-65.

Dede, C., A. Ho and P. Mitros, 2016. Big data analysis in higher education: Promises and pitfalls. Educause Rev., 51: 1-14.

Ellingwood, J., 2016. An introduction to big data concepts and terminology. Digital Ocean, New York, USA. https://www.digitalocean.com/community/tutorials/an-introduction-to-big-data-concepts-and-terminology

Galetto, M., 2016. What is big data analytics?. NGDATA, Inc., Flanders, Belgium. https://www.ngdata.com/what-is-big-data-analytics/

Gilson, M.K., T. Liu, M. Baitaluk, G. Nicola and L. Hwang *et al*., 2016. BindingDB in 2015: A public database for medicinal Chemistry, computational Chemistry and systems Pharmacology. Nucleic Acids Res., 44: D1045-D1053.

Goes, P.B., 2014. Design science research in top information systems journals. MIS. Q. Manage. Inf. Syst., 38: 3-8.

Kim, S., P.A. Thiessen, E.E. Bolton, J. Chen and G. Fu *et al*., 2015. PubChem substance and compound databases. Nucleic Acids Res., 44: D1202-D1213.

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. MBA Thesis, META Group, Stamford, Connecticut.

Matteson, S., 2013. Big data basic concepts and benefits explained. CBS Interactive, San Francisco, California, USA. https://www.techrepublic.com/blog/big-data-analytics/big-data-basic-concepts-and-benefits-explained/

Mills, S., S. Lucas, L. Irakliotis, M. Ruppa and T. Carlson *et al*., 2012. Demystifying big data: A practical guide to transforming the business of government. Breaking Media, Inc., New York, USA.

Monnappa, A., 2019. Data science vs. Big data vs. Data analytics?. Simplilearn Company, Raleigh, North Carolina, USA. https://www.simplilearn.com/data- science-vs-big-data-vs-data-analytics-article

Muresan, S., P. Petrov, C. Southan, M.J. Kjellberg and T. Kogej *et al*., 2011. Making every SAR point count: The development of Chemistry connect for the large-scale integration of structure and bioactivity data. Drug Discovery Today, 16: 1019-1030.

Papadatos, G., A. Gaulton, A. Hersey and J.P. Overington, 2015. Activity, assay and target data curation and quality in the ChEMBL database. J. Comput. Aided Mol. Des., 29: 885-896.

Pence, H.E. and A.J. Williams, 2016. Big data and chemical education. J. Chem. Educ., 93: 504-508.

Sicular, S., 2013. Gartners big data definition consists of three parts, not to be confused with three vs. forbes. Media LLC., New York, USA. https://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/#2c5a77dc42f6

Sushko, I., S. Novotarskyi, R. Korner, A.K. Pandey and M. Rupp *et al.*, 2011. Online Chemical Modeling Environment (OCHEM): Web platform for data storage, model development and publishing of chemical information. J. Comput. Aided Mol. Des., 25: 533-554.

Tetko, I.V., D.M. Lowe and A.J. Williams, 2016. The development of models to predict melting and pyrolysis point data associated with several hundred thousand compounds mined from PATENTS. J. Cheminf., 8: 1-5.

Watson, H.J., 2014. Tutorial: Big data analytics: Concepts, technologies and applications. Commun. Assoc. Inf. Syst., 34: 1247-1268.

Whiting, B., 2018. What is big data analytics?-definition & examples. Council of Better Business Bureaus, Arlington, Virginia. https://study.com/academy/lesson/what-is-big-data-analytics-definition-examples.html