EasyChair Preprint

№ 11009

Big Data Analytics: Review of Application of Data Mining Techniques for (Lean) Six Sigma Methodology

William Pontius and Mark McMurtrey

October 3, 2023

Table of Contents

Big Data Analytics: Review of Application of Data Mining Techniques for (Lean) Six Sigma Methodology

William D. Pontius III
University of Central Arkansas
wpontius@cub.uca.edu

Mark E. McMurtrey Ph.D.
University of Central Arkansas
markmc@uca.edu

ABSTRACT

Lean Six Sigma has proven to be an invaluable performance measure among competing firms. The Data-driven methodology imposed by LSS defines metrics to measure, analyze, improve, and control processes. As our digital footprint increases by the day, there is more data, hence more opportunities for companies to gain a competitive advantage. Firms are beginning to look into ways of capturing more data in an effort to capitalize on its value. Estimations suggest that roughly 80% of all data (Big Data) goes unaccounted, alluding to a wealth of insights for leveraging the market. Advanced data analytics can enhance LSS to improve operational efficiency and facilitate innovation. Through research, we hope to provide a thorough analysis of LSS and BDA in support of combining methods. This paper investigates existing literature to rationalize the application of advanced analytics into the LSS methodology. Also discussed are existing case studies of BDA integration for process improvements.

LEAN SIX SIGMA

LSS combines both Lean and Six Sigma methodologies, essentially sharing the common goal of eliminating waste and creating the most efficient system possible. The ideas behind lean were initially introduced by Henry Ford and later inspired by the production system installed by the Japanese automobile company Toyota. Lean's goal, defined by Toyota, was to reduce production times within processes and response times from suppliers to customers. The Lean methodology uses JIT inventory management, automated quality control, 5S, and value stream mapping to achieve performance parameters such as on-time delivery with the correct quantity and quality. However, rather than a set of tools, firms should adopt lean as a performance improvement philosophy.

Unlike Lean, Six Sigma is a data-driven approach. Introduced in the 1980s by Motorola, Six Sigma was implemented to eliminate variations within their business processes. The Six Sigma methodology utilizes statistics and data analysis to analyze and reduce variation or defects for optimal quality control. According to Six Sigma, defects encompass anything that does not meet customer expectations concerning quality. The etymology of Six Sigma is derived from the Greek symbol "sigma" or "σ," a statistical term for measuring process deviation from the process mean or target. In simpler terms, Sigma measures how far a process deviates from total accuracy or perfection. In statistics, Sigma, also known as process sigma or sigma level, is visualized by the bell curve, where one Sigma represents one standard deviation away from the mean.

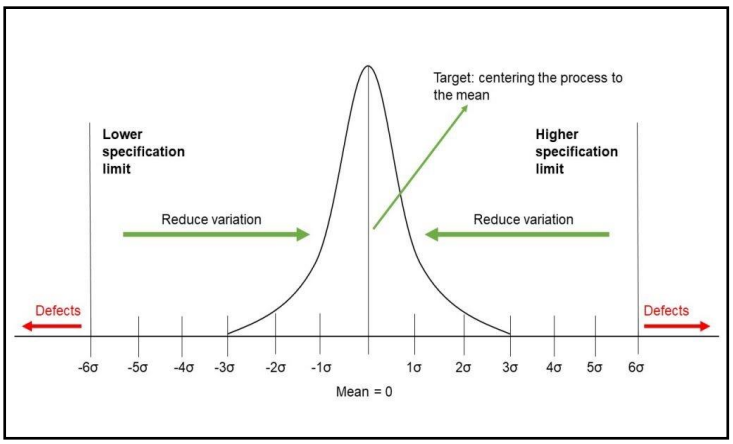

Figure 1. *Six Sigma normal distribution (bell curve).*

When a process exhibits 6 Sigma, the defect rate is purported to be extremely low. Companies that obtain a sigma level of 6 achieve 3.4 defects per million opportunities (DPMO). Although, Sigma level 4 and 5 companies achieve a success rate of 99. +% they cannot adequately compete with Sigma level 6 companies. Defects per opportunity (DPO) is calculated by dividing the total number of defects by the total number of units, times the number of opportunities for error per unit. To calculate DPMO, multiply DPO by 1,000,000. A DPMO of 3.4 means units are produced correctly at 99.999%, representing the probability of success for each process/product characteristic.

Sigma Level	Defects per million	Produced correctly (%)
1	690,000	31.000%
2	308,537	69.146%
3	66,537	93.319%
4	6,210	99.379%
5	233	99.967%
6	3.4	99.999%

Figure 1.1 *Six Sigma performance level*.

While Six Sigma focuses on process improvement and variation reduction, it is driven by the DMAIC model. DMAIC is an acronym for define, measure, analyze, improve, and control. The DMAIC methodology is a data-driven quality strategy utilized for process improvement by integrating a Six Sigma quality initiative. Firms implement the DMAIC model into their Six Sigma projects to improve and maintain their sigma level and current processes. The DMAIC method cannot be supported without sufficient data. Firms will compile and analyze relevant data through data analytics when deciding which performance measures to utilize, such as LSS. While structured data has been considered the norm for data analysis, the focus is shifting in organizations to start harnessing unstructured data, hence big data.

BIG DATA

From creating the punch card in 1725 to forming data lakes to store and process big data, accessing, storing, and utilizing vast amounts of data has become an integral step in the evolution of organizational competition. The vast amounts of data generated daily are too large to handle with conventional data management, and traditional data-processing application software is no longer viable. This reality alludes to the emergence of Industry 4.0, a concept based on the emergence of new technologies such as cloud computing, the internet of things (IoT), cyber-physical systems, and big data (Moeuf et al., 2017). Data scientists and organizational leaders are looking toward advanced data analytics or big data analytics to capture the value hidden in raw data. BDA has been traced to the mid-1990s, first coined by John Marshey, retired former Chief Scientist at Silicon Graphic, to refer to the handling and analysis of massive datasets (Diebold, 2012). In 2001, Doug Laney, Vice President and analyst for Gartner Management consulting company, ascribed big data as having three dimensions: volume, variety, and velocity. Laney's idea became known as the Three V's of big data, each posing a challenge to firms who wish to exploit it.

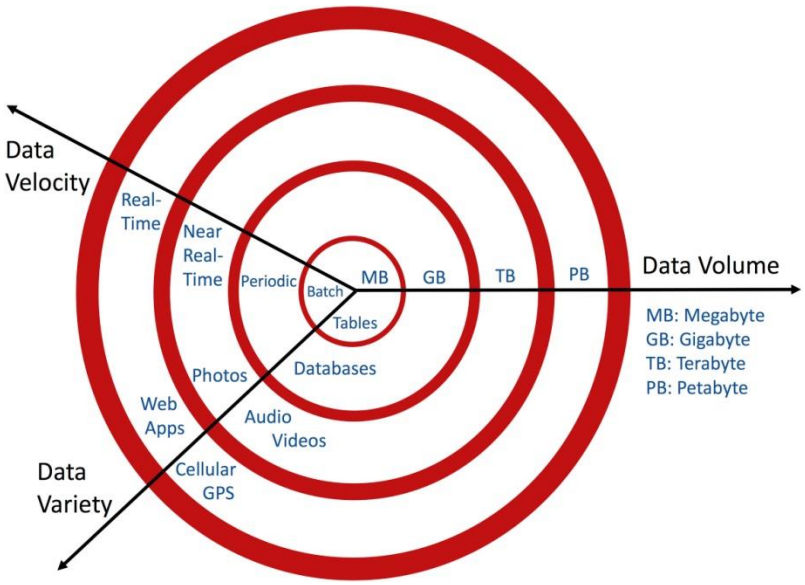

Figure 2. *The 3 Vs of big data* (Soubra et al., 2018, as cited in Wehner et al., 2017)

Albeit the references to the mid-nineties and early 2000s, the term big data was not as ubiquitous until 2011. The hype surrounding big data can be attributed to the promotional initiatives by IBM and other leading technology companies that invested in building the niche analytics market (Gandomi & Haider, 2015).

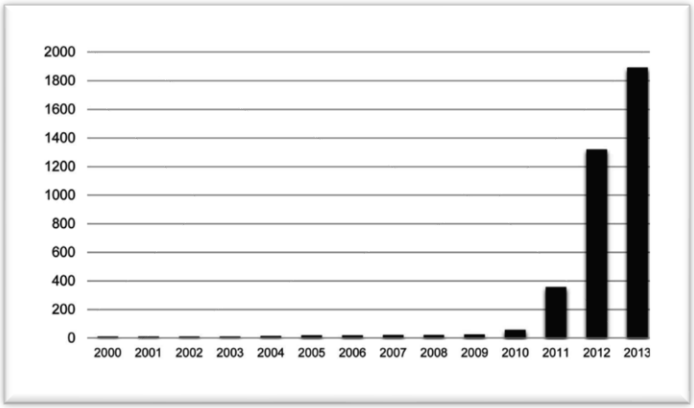

Figure 3. *Frequency distribution of documents containing the term "big data" in ProQuest Research Library*

Big data is comprised primarily of semi-structured and unstructured data. Unstructured data, typically categorized as qualitative data, does not have a pre-defined data model or is not organized in a pre-defined manner. Data generated by firms include textual, multimedia, and network data. Unstructured data is typically text-heavy. Multimedia data consists of alphanumerics, images, graphics, animation, video, and audio. Network data is generated from equipment to equipment communication, cyber-physical systems, and sensor networks (Gandomi & Haider, 2015; Li & Wang, 2017; Sivarajah et al., 2017, as cited in Gupta et al., 2020). Cyber-physical systems range from robots, intelligent buildings, implantable medical devices, self-driving automobiles, and drones. Sensor networks are integrated into home security systems, environmental monitoring, municipal surveillance, and earthquake detection. The generally accepted maxim is that structured data represents only 20% of the information available to an organization (Gutierrez, 2015).

Consumers and enterprises generate about 2.5 quintillion bytes of data per day. Data units are no longer being reported as gigabyte or terabyte, but rather in petabytes, exabytes, and zettabytes.

PB ($1PB = 2^{10}TB$)

EB ($1EB = 2^{10}PB$)

ZB ($1ZB = 2^{10}EB$)

In 2022, it is estimated that about 97 zettabytes of data currently make up the entire digital universe and is predicted to exceed 181 zettabytes by 2025 (Statista, 2021). There is a noticeable correlation between 2011, the widespread usage of the term big data, and the uptick in the volume of big data created.

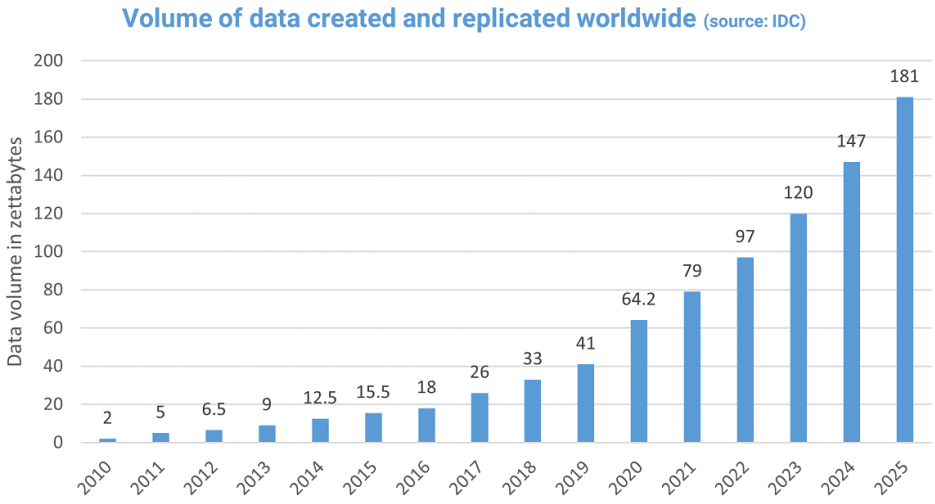

Figure 3.1 *Volume of data created and replicated worldwide.*

Fully leveraging all the data generated is essential for organizational success, especially in the wake of the fourth industrial revolution. In addition, more data leads to more accurate and confident decision-making. Firms can utilize big data tools and software such as Hadoop to process, mine, integrate, store, track, index, and report business insights from raw unstructured information (Dialani, 2020). The big data and LSS capabilities of a firm depend upon the resources, which are tangible and intangible (Braganza et al. 2017; Wamba et al. 2016, as cited in Gupta et al., 2020). This aspect represents a firm's resource-based view of big data and LSS (Gupta et al., 2020).

Furthermore, firms must equip the necessary BDA tools and have sufficient storage available for appropriate data analysis to target LSS projects effectively. IDC (International Data Corporation) states that storage capacity worldwide was 6.7 zettabytes in 2020. The IDC predicts a five-year annual compound growth rate or CAGR of 19.2% throughout the forecast period.

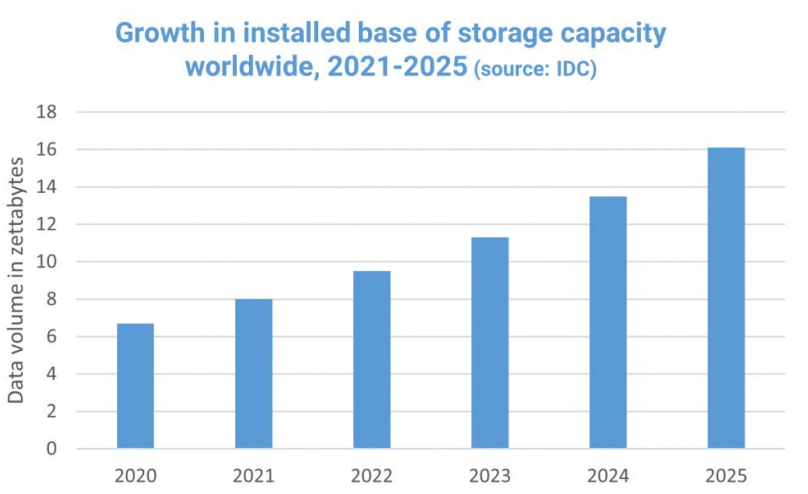

Figure 3.2 *Growth in the installed storage capacity base worldwide, 2021-2025*

BDA TECHNIQUES

Using BDA and available tools, organizations can exploit data for many reasons. Sectors such as manufacturing utilize BDA to obtain asset performance and efficiency gains, improve production processes and supply chain performance, and create possible product customization. Financial services use BDA to inform better investment decisions, maximize portfolio returns, and identify and prevent fraudulent activities. In addition, the healthcare sector is utilizing BDA to predict and prevent epidemics, cure diseases, and reduce the costs of treatment for patients (Berisha et al., 2022).

To further help firms decide their objectives, BDA can be categorized into three sets of analytics – Predictive, Descriptive, and Prescriptive Analytics. Descriptive analytics identifies correlations through the use of current and historical data. Firms use descriptive analytics to figure out what happened or is happening in a process. Predictive analytics tells firms what will happen next in their business or process by analyzing patterns in current and historical data. Prescriptive analytics utilizes available data to determine an optimal course of action and aids firms in answering questions such as, "what should the firm do?".

Classifying BDA into different categories can aid firms in distinguishing their goals, deciding the optimal performance measure, and deciding which BDA techniques to employ. BDA techniques fall into the field of data science and encompass big data, data analysis, and artificial algorithms. Two of the most prominent BDA techniques include artificial intelligence and data mining. Each technique has subsets and is considered in the realm of data science.

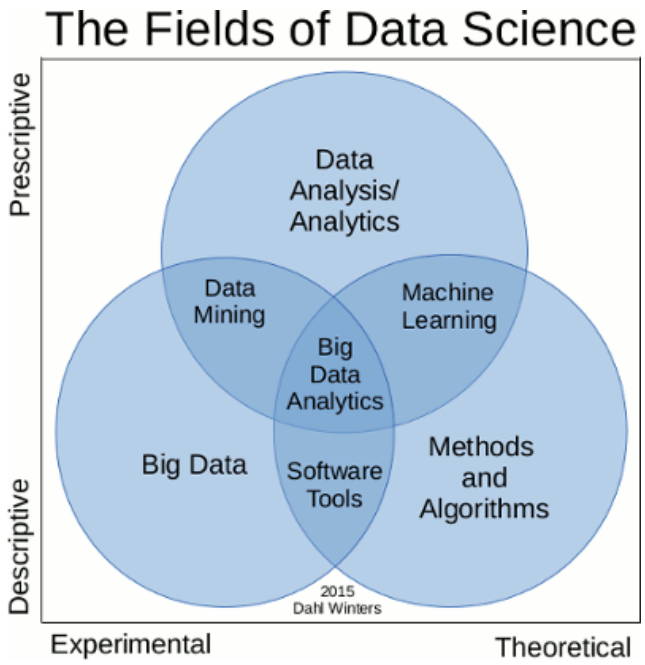

Figure 4. *The Fields of Data Science*

A sub-field of AI is machine learning (ML). Machine learning has many applications, including decision-making, forecasting, and predicting. It is a critical enabling technology in deploying data mining and big data techniques in the diverse fields of healthcare, science, engineering, business, and finance (Ali et al., 2016). Sub-fields of ML include supervised learning and unsupervised learning.

In ML, supervised learning refers to a class of algorithms set to determine a predictive model utilizing data points with known outcomes. Supervised learning is defined by its use of labeled input and output data. If the output (or prediction) belongs to a continuous set of values, then such a problem is called regression. At the same time, if the output assumes discrete values, then the problem is called classification (Ali et al., 2016). *Classification* is a data mining function that separates data points into different classes according to their attributes. Types of classification techniques include Naïve Bayes Classifiers, Support Vector Machines, K-Nearest Neighbors, Logistic Regression, and, most famous of the methods, Decision Trees.

Unsupervised learning algorithms work autonomously to discover the inherent structure of unlabeled data without using labeled input and output data. Unsupervised algorithms can be classified into three tasks – clustering, association, and dimensionality reduction. *Clustering* is a data mining technique for grouping unlabeled data based on correlations. *Association* is another technique used to find relationships between variables in a data set and is commonly used for basket analysis.

Necessary to BDA and in tandem with ML, data mining can be referred to as the techniques used for finding and describing structural patterns in large volumes of data (Witten & Frank, 2005). Data mining and ML are used to perform predictive, descriptive, and prescriptive analytics to uncover hidden patterns in the data. Like process discovery, the process of extracting useful information from relational and non-relational data using data mining and ML techniques is called knowledge discovery (Ali et al., 2016).

BDA TECHNIQUES APPLIED TO LSS

Through BDA, firms obtain and suggest the best performance measures (Gupta & George, 2016). Of many performance measures, lean and six Sigma are the most practiced among firms (Antony, 2011; Hines, Holweg, and Rich, 2004, as cited in Gupta et al., 2020). LSS provides a structured and measured approach, whereas advanced and BDA can predict and analyze business problems more appropriately (Fogarty 2015b, as cited in Gupta et al., 2020). Furthermore, LSS utilizes the DMAIC methodology to develop process improvement, but many traditional tools are insufficient to handle the amount of data now available to firms. By combining traditional analytical tools with BDA techniques, companies can apply data-driven applications in real time, with quicker data processing and manipulation. In a study conducted by (Gupta et al., 2020), a basic framework is proposed for applying BDA techniques and traditional analytical tools in each phase of LSS DMAIC methodology via inspiration from Moefuf et al. (2018).

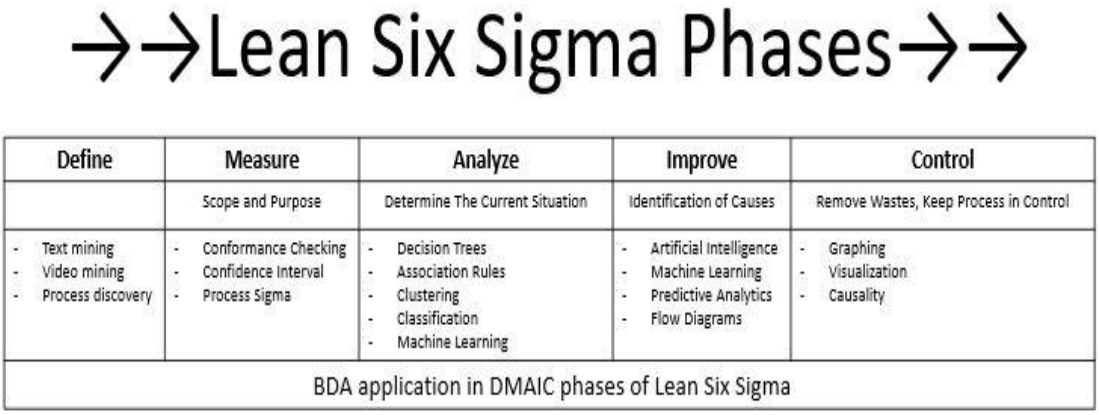

Figure 5. *BDA applied to the DMAIC phases of LSS*

In the framework proposed by (Gupta et al., 2020), data mining techniques that can be applied to the define phase of LSS include ML, text mining, video mining, and process discovery. According to (Gupta et al., 2020), text mining can translate large data sets into a meaningful summary with the help of proof (data) centered decision-making. Text mining data can be generated from social media, emails, event logs, weblogs, and work orders (Gandomi and Haider, 2015, as cited in Gupta et al., 2020). Video mining can aid and support analysis to clarify the problem and define the scope by providing meaningful clues about the surrounding environment. Process mining, also called process discovery, extracts data from event logs relevant to a process. As LSS focuses on improving the process, *process mining* can be defined as a BDA technique for discovering and tracking actual transactions by mining the information from event logs to help minimize the variations and wastes in the process (Rovani et al. 2015, as cited in Gupta et al., 2020).

The second phase of measure is mapped with the help of confidence interval, process sigma, and conformance checking (Gupta et al., 2020). Advanced statistical techniques are considered a more efficient approach to understanding a process's behavior and identifying waste, unhidden bottlenecks, and operational rigidities (Soofastaei A., 2020). The third phase, analyze, can utilize supervised and unsupervised ML techniques such as classification, decision trees, clustering, and association. ML can learn patterns through artificial neural networks to aid in analysis, resulting in more confident conclusions.

The fourth phase of improve, according to (Gupta et al., 2020), can utilize AI, ML, predictive analysis, and flow diagrams for better results. Utilizing BDA techniques, such as AI algorithms, aids in optimizing the process parameters of a problem targeted in the LSS project. The last phase of control can be visualized through the utilization of graphics and visualization derived from process mining. Process mining used for visualization can view and detect potential failures which could impact the improvement and optimal solution obtained in the improvement phase. At the same time, an AI algorithm can monitor process control (Gupta et al., 2020).

SUPPLEMENTAL DATA

Concerning the framework introduced by (Gupta et al., 202) as a foundation to address the use of BDA techniques applied to the DMAIC methodology of LSS, we have undertaken a systematic literature review to further support this framework by introducing supplementary techniques to the DMAIC phases. The DMAIC methodology is a data-driven solution to problem-solving that implements operational practices to produce quasi-perfect products and services by reducing variations (Soofastaei A., 2020). The DMAIC process, by definition, is highly analytical and profoundly rooted in statistical analysis (Soofastaei A., 2020).

→→Lean Six Sigma Phases→→

Define	Measure	Analyze	Improve	Control
Scope and Purpose	Current Situation	Identify Causes	Implement Solutions	Keep Process in Control
Dimensionality Reduction • Text mining • Image Recognition Big Data Visualization Artificial Intelligence • Natural Language Processing (NLP) • Document (text) Segmentation Video mining Process discovery	Conformance Checking Confidence Interval Process Sigma Artificial Intelligence • Machine Learning	Data Mining • Association Rules • Clustering • Classification ▪ Decision Trees Machine Learning • Text Mining Video Mining	Artificial Intelligence • Machine Learning Data mining • Predictive Analytics Flow Diagrams Digital Twin Model Advanced Process Simulation Virtual Process Model Process Simulation	Artificial Intelligence • Machine Learning Graphing Visualization Causality Anomaly detection Prescriptive Analytics Augmented Reality AI-based automated decision
BDA application in DMAIC phases of Lean Six Sigma				

Figure 5.1 *BDA applied to the DMAIC phases of LSS - Expanded*

According to (Soofastaei A., 2020), natural language processing can aid in supporting analysis to define the problem and scope. Natural language processing allows computers to interpret text and spoken words similar to what humans can. NLP utilizes computational linguistics with statistical, ML, and deep learning algorithms. Another AI algorithm that could help define the problem is Document (text) Segmentation, which subdivides unstructured text from digital documents and images into meaningful parts. Similar to text mining, document segmentation can aid in extracting meaningful information from raw data – PDF text documents and images. Document segmentation could be beneficial in the healthcare industry for identifying insights from unstructured clinical records and doctor's notes.

The measure phase typically requires the use of statistical measures. The measure phase and analyze phase emphasize the analytical methods used. To further enhance the measure phase, LSS practitioners can utilize AI, ML, and other data mining techniques to handle data quality issues. The analyze phase can utilize text and video mining in accommodation with supervised and unsupervised ML techniques for more thorough process analysis.

A virtual process model could be utilized in the improvement phase to reflect an existing process accurately. The physical process is equipped with sensors that share data surrounding the process's performance. The data gathered through the sensor networks aid in predicting how a product or process will perform. Another technique that can aid in process improvement is process simulation. Process simulation provides a mechanism for robust validation under realistic conditions. If the simulation outcomes fail to sync with the expected results, it can substantially reduce the risk of deploying a new process (Shukla, 2022). Simulators allow businesses to dynamically examine process flows, validate the model and collect timing and resource information on proposed or legacy processes to improve them (Shukla, 2022). Lastly, the control phase can utilize AI for process control through real-time sensor networks that provide feedback, feedforward, and

predictive control analysis. AI algorithms trained using historical process data and ML can continuously improve process performance, monitoring variations.

USE CASES OF SUCCESSFUL BDA INTEGRATION FOR PROCESS IMPROVEMENTS

Fogarty (2015) evaluated a large global financial firm case where an advanced analytics team incorporated Six Sigma into their current process. In addition to realizing that Six Sigma helped improve their analytics projects by having a more structured and measured approach to executing analytics activities, they also discovered that analytic activities enabled projects not directly related to business analytics throughout the firm. According to (Fogarty, 2015), BDA will enable practitioners to take advantage of the massive stores of information accumulated by firms to measure the process better and to search for insights that fuel process improvements and innovation. Furthermore, Fogarty hypothesized that some insights would only be obtained through analyzing big data.

A case study by Manenti (2014) focused on the benefits of using BDA for process improvements. It noted that Intel, in 2012, saved $3M by implementing BDA for preventative analysis on a single microchip production line. Intel estimated that integrating BDA into more chip production lines would save over $30M throughout the next few years. In addition, General Electric estimates they could increase production speeds by 25% in their Aviation sector by implementing BDA to enable in-process inspection.

Another manufacturing BDA case study was a biopharmaceutical company that sought to reduce inconsistency in the capacity and quality of its manufacturing processes which could attract regulatory attention (Forgaty, 2015). Using BDA, the team conducted a segmentation analysis of its manufacturing processes based on activity, revealing process interdependencies and non-parameters that could impact vaccine yield (Fogarty, 2015). The team increased vaccine production by 50% by modifying target processes per the insights generated from BDA. The increase resulted in annual savings between $5M and $10M.

(Fogarty, 2015) also mentions that the service side is utilizing BDA for process improvements. In a case study, Infinity, a property and casualty company, utilized BDA and predictive analytics to identify signs of insurance fraud. Using BDA and predictive analytics, Infinity increased the success rate of pursuing fraudulent claims from 50-80%. Infinity also reduced the cycle time required for referring questionable claims for investigation by 95%, leading to an underwriting profit every year since implementing BDA.

LIMITATIONS OF BDA

i. *Storage*
Storage and maintenance cost for large data sets are costly and large servers require significant space. Firms have the option to outsource operations. Cloud hosting and cloud storage utilizes the host company's infrastructure and can provide a potential solution to this problem.

ii. *System Design and Integration*
The risk of uncertainty, complexity, and variety pose challenges to the design and integration of big data systems (Gupta et al., 2020). The need for big data is different among firms, and variations in data management, data extraction, and data alterations will create discrepancies in data synchronization. Furthermore, choosing the correct framework for a firm's specific needs can

become confusing regarding the compatibility of various approaches. The challenge of integration is further stipulated by the lack of talent in the field of BDA.

iii. *System Performance*

Big data systems are typically horizontally scaled distributed systems and operate over thousands of storage systems nodes(Gupta et al., 2020). Reading and writing big amounts of data and execution of parallel topologies leads to highly variable response latency (Dean and Barroso 2013, as cited in Gupta et al., 2020). Once the big data system is deployed in an LSS environment, there is no regulation to the input sources and rate of data flow. Therefore data grows exponentially beyond the predicted volume(Gupta et al., 2020). Finally, the shared infrastructure such as cloud may not provide the anticipated quality of service due to the disagreement of resources at different levels (Klein and Gorton 2015, as cited in Gupta et al., 2020).

iv. *Quality Control*

Big data is a new concept, and academia hasn't established a uniform definition of its data quality and quality criteria. The literature differs on a definition of data quality, but one thing is certain: data quality depends not only on its own features but also on the business environment using the data, including business processes and business users. Only the data that conform to the relevant uses and meet requirements can be considered qualified (or good quality) data (Cai & Zhu, 2015). Extracting genuine high-quality data, especially in real time from massive data sets poses many challenges. Data variety generated from a diverse array of data sources makes data integration significantly more complex. Furthermore, data volume is extensive, making it difficult to sort data in a set amount of time. Timeliness of data is transitory, suggesting the need to for advanced processing technologies.

v. *Security and Privacy Concerns*

Big data is comprised of both enterprise and consumer data. There is potential for proprietary information to be leaked to unauthorized personnel such as competitors. Along with corporate data, firms collect confidential information about their client and customers Firms can utilize tools such as data cleansing algorithms, data masking measures, and data encryption to secure information. While these methods provide a layer of security, they each have limitations and risks.

CONCLUSION

Given the existing literature surrounding the topic of big data and its corresponding analytical techniques being applied to Lean and Six Sigma, the DMAIC methodology, and its success in real-world case studies, it is probable to suggest BDA in LSS holds great promise. From the rise in technology and the expansion of data generation, BDA holds the potential for renewing LSS into a modern-day tool with advanced capabilities. A limitation of this study is that few case studies currently exist surrounding the application of BDA in LSS. Furthermore, less research regarding BDA applied directly to the LSS DMAIC phases is available. As AI and data mining algorithms evolve, processes will become more automated leading Six Sigma companies to undergo BDA integration. For future research it is suggested that studies be conducted on specific Six Sigma companies who implement BDA techniques for process improvements relevant to their industry.

REFERENCES

Ali, A., Qadir, J., Rasool, R. ur, Sathiaseelan, A., Zwitter, A., & Crowcroft, J. (2016). Big data for development: applications and techniques. *Big Data Analytics*, *1*(1). https://doi.org/10.1186/s41044-016-0002-4

Antony, J. (2011). Six Sigma vs Lean. *International Journal of Productivity and Performance Management*, *60*(2), 185–190. https://doi.org/10.1108/17410401111101494

Berisha, B., Mëziu, E., & Shabani, I. (2022). Big data analytics in Cloud computing: an overview. *Journal of Cloud Computing*. https://doi.org/10/1186/s13677-022-00301-w

Braganza, A., Brooks, L., Nepelski, D., Ali, M., & Moro, R. (2017). Resource management in big data initiatives: Processes and dynamic capabilities. *Journal of Business Research*, *70*, 328–337. https://doi.org/10.1016/j.jbusres.2016.08.006

Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002

Dean, J., & Barroso, L. A. (2013). The tail at scale. *Communications of the ACM*, *56*(2), 74–80. https://doi.org/10.1145/2408776.2408794

Dialani, P. (2020). The Future of Data Revolution will be Unstructured Data. *Analyticsinsight*. https://doi.org/http://www-users.cs.umn.edu/~han/dmclass/cluster_survey_10_02_00.pdf

Diebold, F. X. (2012). On the Origin(s) and Development of the Term "Big Data." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2152421

Fogarty, D. (2015). Lean Six Sigma and Big Data: Continuing to Innovate and Optimize Business Processes. *Journal of Management and Innovation*. https://doi.org/file:///C:/Users/user/Downloads/8-Article%20Text-57-1-10-20150907%20(17).pdf

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2014.10.007

Gupta, M., & George, J. F. (2016). Toward the development of a big data analytics capability. *Information & Management*, *53*(8), 1049–1064. https://doi.org/10.1016/j.im.2016.07.004

Gupta, S., Modgil, S., & Gunasekaran, A. (2020). Big data in lean six sigma: a review and further research directions. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2019.1598599

Gutierrez, D. (2015, June 5). *Text Analytics: The Next Generation of Big Data*. InsideBIGDATA. https://insidebigdata.com/2015/06/05/text-analytics-the-next-generation-of-big-data/

Klein, J., & Gorton, I. (2015). Runtime Performance Challenges in Big Data Systems. *Proceedings of the 2015 Workshop on Challenges in Performance Methods for Software Development - WOSP '15*. https://doi.org/10.1145/2693561.2693563

Li, D., & Wang, X. (2015). Dynamic supply chain decisions based on networked sensor data: an application in the chilled food retail chain. *International Journal of Production Research*, *55*(17), 5127–5141. https://doi.org/10.1080/00207543.2015.1047976

Moeuf, A., Pellerin, R., Lamouri, S., Tamayo-Giraldo, S., & Barbaray, R. (2017). The industrial management of SMEs in the era of Industry 4.0. *International Journal of Production Research*, *56*(3), 1118–1136. https://doi.org/10.1080/00207543.2017.1372647

Rovani M., Maggi F.M., de Leoni M., van der Aalst W.M. Declarative process mining in healthcare. *Expert Syst. Appl.* 2015; **42:9236**–9251. https://doi.org/10.1016/j.eswa.2015.07.040

Shukla, B. (2022, July 29). *A guide to process simulation*. Process Excellence Network. https://www.processexcellencenetwork.com/tools-technologies/articles/a-guide-to-process-simulation

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of Big Data challenges and analytical methods. *Journal of Business Research*, *70*(1), 263–286. Sciencedirect. https://doi.org/10.1016/j.jbusres.2016.08.001

Soofastaei, A. (2020). Digital transformation of mining. In *Data analytics applied to the mining industry* (pp. 1-29). CRC Press.

Statista. (2021, June 7). *Data Created Worldwide 2010-2025 | Statista*. Statista; Statista. https://www.mybib.com/tools/apa-citation-generator#:~:text=https%3A//www.statista.com/statistics/871513/worldwide%2Ddata%2Dcreated/

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S. J., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, *70*, 356–365. https://doi.org/10.1016/j.jbusres.2016.08.009

Wehner, M. R., Levandoski, K. A., Kulldorff, M., & Asgari, M. M. (2017). Research Techniques Made Simple: An Introduction to Use and Analysis of Big Data in Dermatology. *Journal of Investigative Dermatology*, *137*(8), e153–e158. https://doi.org/10.1016/j.jid.2017.04.019

Witten, I. H., & Frank, E. (2005). *Data mining : practical machine learning tools and techniques*. Morgan Kaufman.

APPENDIX

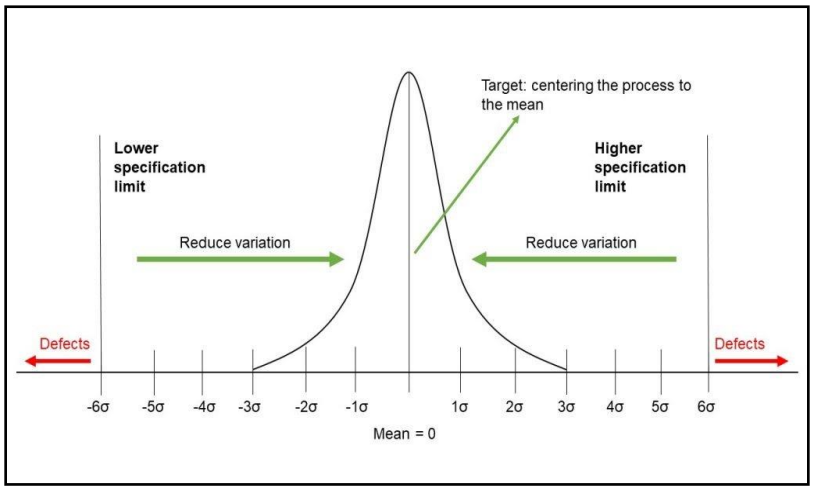

Figure 1. Six Sigma normal distribution (bell curve).

Sigma Level	Defects per million	Produced correctly (%)
1	690,000	31.000%
2	308,537	69.146%
3	66,537	93.319%
4	6,210	99.379%
5	233	99.967%
6	3.4	99.999%

Figure 1.1 Six Sigma performance level.

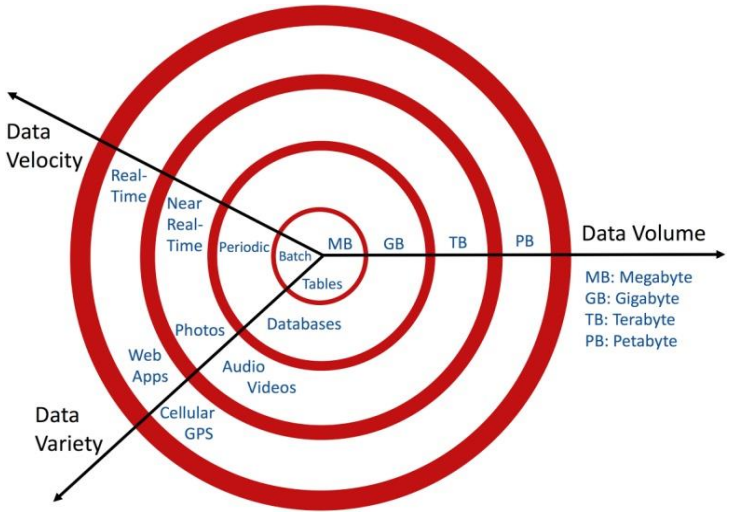

Figure 2. The 3 Vs of big data (Soubra et al., 2018, as cited in Wehner et al., 2017)

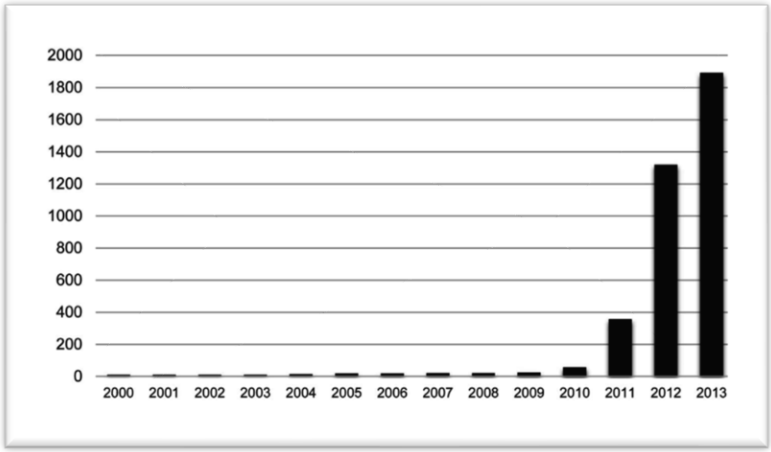

Figure 3. Frequency distribution of documents containing the term "big data" in ProQuest Research Library

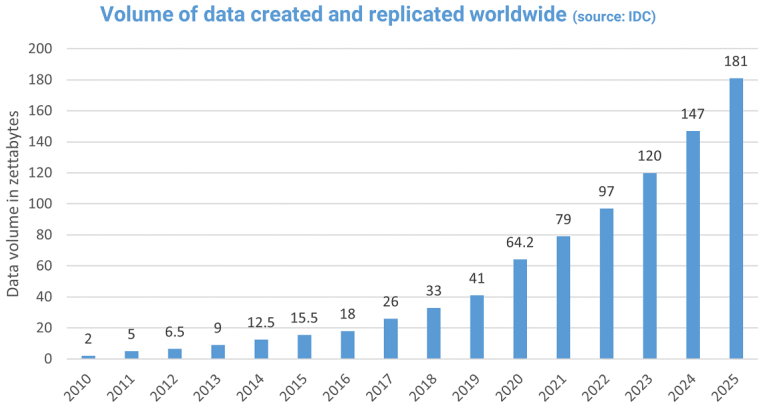

Figure 3.1 Volume of data created and replicated worldwide.

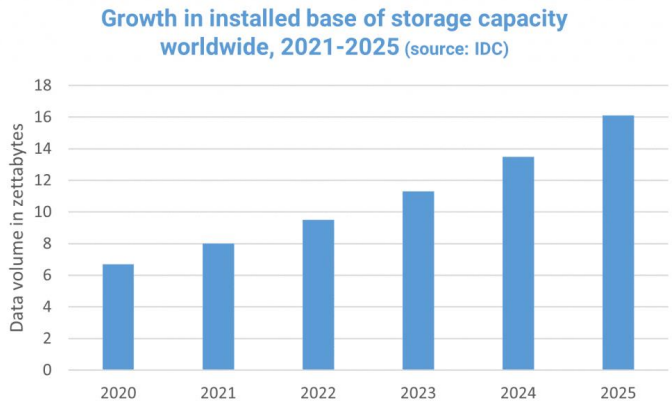

Figure 3.2 Growth in the installed storage capacity base worldwide, 2021-2025.

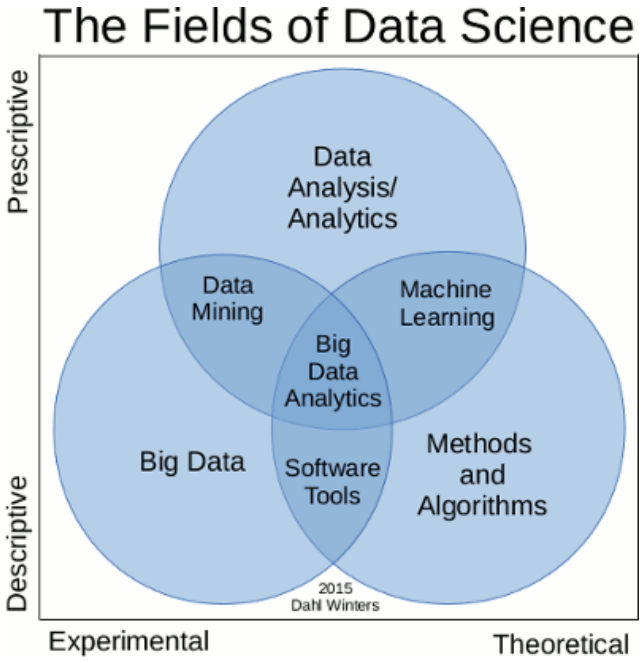

Figure 4. The Fields of Data Science

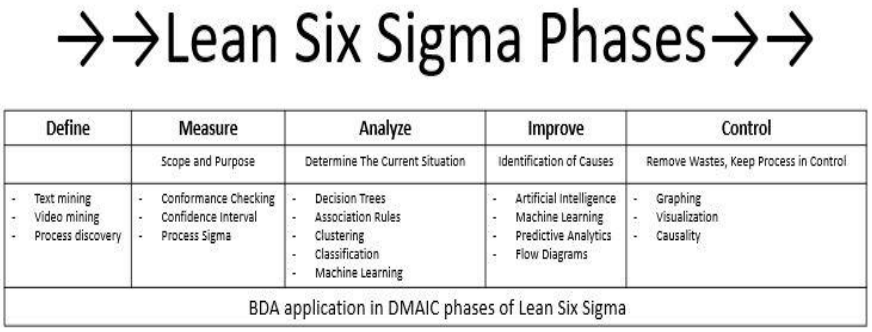

Figure 5. BDA applied to the DMAIC phases of LSS

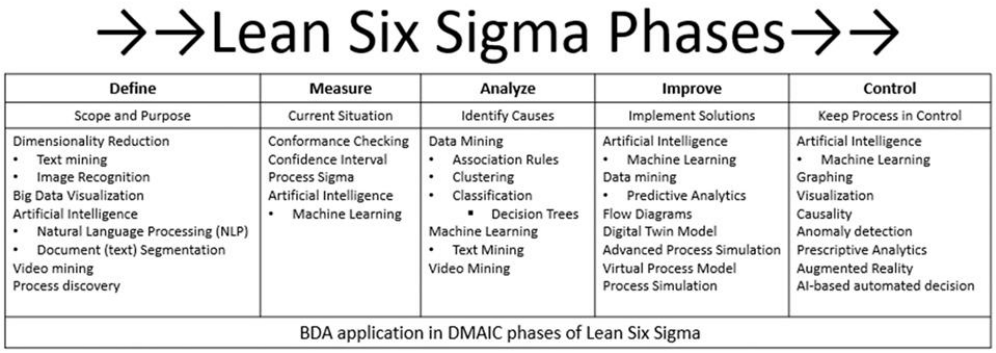

Figure 5.1 BDA applied to the DMAIC phases of LSS - Expanded