A Crossroads of Big Data and Data Science- A Review

Hirdesh Sharma[1], Dr. Sachin Kumar[2], Dr. Gaurav Aggarwal[3]

[1]Assistant Professor, Dept. of MCA, Noida Institute of Engg. & Tech., Greater Noida, India

EmailID:- hirdesharma@gmail.com

[2]Associate Professor, Dept. of MCA, Noida Institute of Engg. & Tech., Greater Noida, India

EmailId:- sachinks.78@gmail.com

[3]Professor & Dean, Dept. of CSE, Jagannath University, Bahadurgarh(Jhajjar), Haryana, IndiaEmailId:-- gauravaggarw@gmail.com

Abstract

Big Data Technology and Data Science are changing the environment in ways that are raising new questions among social scientists, such as the effects of the internet on people and the media, the ramifications of smart cities, the potential for cyber-warfare and cyber-terrorism, the implications of precision medicine, and the consequences of artificial intelligence and automation. Effective new data science approaches support analysis using administrative, internet, textual, and sensor-audio-video data in tandem with these societal changes. By providing new ways to construct concepts from data, do descriptive inference, make causal inferences, and produce predictions, burgeoning data and creative methods make it easier to answer previously difficult-to-address questions about society.They also pose difficulties for social scientists, who must understand the significance of concepts and predictions produced by complicated algorithms, weigh the relative value of prediction versus causal inference, and deal with ethical issues as their methods, such as algorithms for voter mobilization or bail determination, and are adopted by policymakers.

Keywords:big data, data science, artificial intelligence, cyberinfrastructure, causality, prediction, text analysis, internet, smart cities, cyber-warfare, automation.

Introduction about Big Data and Data Science

Big data and data science are buzzwords that are made up of a range of ideas. Despite their flaws and the hyperbole that often surrounds them, these words point to real shifts in political science. The following advances and trends are explored in this article thanks to big data, data science, and related ideas like artificial intelligence, cyberinfrastructure, and machine learning: Big data and data science are triggering social and political changes.The enormous increases in computing capacity and advances in data science techniques have combined to transform society in fundamental ways, thanks to the amount, velocity, variety, and veracity of data produced by and accessible to governments, armies, companies, non-profits, and citizens.Big data and data science are spawning new phenomena and posing fundamental questions about the monitoring and exploitation of individuals and populations, the future of privacy, the veracity of knowledge, the future of work, and a slew of other issues that concern political scientists. All scientists, including political scientists, would have access to greater volumes of data. These changes have an effect on all sciences. The Thirty Meter Telescope, which will be operational in 2022, will generate 90 terabytes per night; genomic data doubles every nine months and is currently generated at a rate of approximately 10 terabytes per day; and the Large Hadron Collider at CERN generates 140 terabytes per day. Every day, the World Wide Web generates about 1,500,000 terabytes of data, allowing social scientists to investigate the "sinews of society" and the "nerves of government" in ways that were previously impossible. Political scientists can now observe and analyze (sometimes in real time) the information that people choose to absorb, the information provided by political actors, their living environment, and many other aspects of people's lives. New approaches to the organization of political science research.With the influx of data, political scientists will rethink how they perform research by studying new technologies that make data access, management, cleaning, study, and archiving easier. Political scientists are posing new types of questions. Political scientists must ask themselves what they want to achieve through idea formation, explanation, causal inference, prediction, and future projection. New approaches and insights into political behaviour will be created as a result of this process, as well as new designs for political institutions.*Ethical questions of political science research are being addressed. Finally, political scientists must consider difficult ethical problems such as knowledge access, use, and dissemination, as well as the potential abuse of their models and findings.*I explain the rapid increase in data and computational capacity that has led to the popularity of so-called big data and data science, accompanied by descriptions of these untidy words, before moving on to these five changes and their consequences for political science[1-2].

Definitions of Big Data and Data Science

More efficient algorithms are needed to process the large amounts of data. The use of data analytics makes this possible. Data Analytics is the application of structured statistical and mathematical techniques on collected data in order to detect underlying patterns as well as make predictions.Many attempts to understand and use data have been launched as a result of the growth of data and the development of vast databases in industry, government, everyday life, and science research. Big data and data science have recently gained popularity [11].

Big Data

“Big data” may sound like a moving target to those of us who recall when computer memory was calculated in kilobytes rather than terabytes (a factor of a billion more), but the concept has emerged amid advancements in computer power since data appears to be increasing faster than our ability to process it. Data volume, variety (text, images, audio, video, sensor, social media, and other forms), and regular velocity (Laney 2001) are all the at a faster rate than computing power. The vast amount of data causes storage and management issues. The increase in variety makes it more difficult to convert data from one format to another, and the increase in velocity necessitates on-the-fly data editing and prioritization. On top of length, variety, and velocity, a fourth concern, data veracity, has recently emerged, introducing a new layer of complexity.

The new technologies for capturing, communicating, networking, and generating information, rather than the sheer volume of data, are truly distinguishing features of the big-data revolution. Phone calls, emails, messages, tweets, social media posts, and other technical methods are now digitally registered, time- and location-stamped, and attributed to nodes in networks in ways that far outweigh the far more ephemeral media of the past. FedEx monitoring services, Online searches and transactions, parking meter payments, car journeys, tax payments, photos of social events, weather and environmental measurements, digital images from microscopes and telescopes, and many other business, legislative, social, and science tasks now have digital trails. When these variables are applied to the fact that the World Wide Web is an excellent forum for social networking and knowledge access, and that machines can now author information and communicate with us—possibly even creating artificial intelligence, autonomous robot-like beings, and virtual realities—the perception is of immersive data that encompasses us in our everyday lives, rather than just large data. The NIST-identified "decentralization of data" may be more than a collection of strategies for dealing with massive computing problems, but the future shape of computing and the internet is still uncertain. As a result, the real impact of the big-data revolution is a shift in our cognitive climate (Lugmayr et al. 2016, Neumann 2016, Schroeder 2018) that necessitates new perspectives to deal with datafication, connectedness, networking, and machine authoring (Lugmayr et al. 2016, Neumann 2016, Schroeder 2018). These occurrences are the result of the development of emerging technology, including data science approaches that are cutting-edge [12].

Data Science

Data Science is a field of research that entails using various scientific methods, algorithms, and processes to derive information from large volumes of data. It assists you in uncovering secret trends in raw data. Because of the evolution of statistical analytics, data mining, and big data, the term Data Science has evolved.

The companion concept of big data, data science, is based on a description of a method for discovering new information in an age when data is aplenty and cries out to be analyzed. In 2001, statistician William S. Cleveland proposed a proposal to “enlarge the main fields of technical work in the field of statistics” by allocating more resources to “computing with data” (Cleveland 2001, pp. 21, 22), and dubbed the new discipline “data science.” In a 2007 speech to the National Research Council's Computer Science and Telecommunications Commission, computer scientist Jim Gray called for "data-driven science" as a modern scientific paradigm that allows scientific discoveries using vast collections of data. Gray (2009, p. xxv) suggested that there was a “need for tools to help scientists collect their data, curate it, and then visualize it,” with the aim of “unifying all science data with all literature to create a world in which the data and literature interoperate with one another.”

The steam engine was only one aspect of the nineteenth-century transportation revolution, which also included the discovery of new sources of energy (oil and electricity), the invention of new types of motors (internal combustion and electrical), the development of rail, road, and river networks, and the establishment of new social norms such as standard time zones. Similarly, the information revolution is about far more than computers or some other single object. Sensors, databases, programming languages, artificial intelligence, telecommunications, machine learning, social media, the internet, and a variety of other technologies are all part of it. None of these advances are covered by the terms "big data," "data science," or some other name. The phrase "cyber infrastructure" may be useful, but it has yet to catch on. Jordan (2018), a leading data science researcher, advocates for the word "intelligent infrastructure," which is wider than "artificial intelligence," but has its own set of limitations. We're left with actual phenomena but insufficient terminology [13-14].

Quantity of data accessible to all scientists

Many new data sources have also helped social scientists. Political scientists had access to a small number of data sets, mainly about the United States but also about other nations, as early as 1980: Historical election numbers, typically by county but sometimes by precinct; polls dating back to the 1930s; census data;Data on campaign donations from the Federal Election Commission (FEC); roll-call data from legislatures and the UN; data from the Correlates of War Project, the Global Handbook of Political and Social Indicators, and a few other outlets. Beyond these regions, the volume and variety of data has grown dramatically in the last 30 years, thanks in large part to-

- data from the government
- access to the internet
- information in text form

- Data from sensors, as well as audio and video [15].

Literature Review

Mariani, M et al. (2021) is the first attempt to capture how the presence and depth of hospitality services customers' environmental discourse affect e-WOM helpfulness through various digital channels, using a broad sample of online feedback from multiple countries and destinations.As such, it adds to the field of big data analytics, e-WOM, and sustainable tourism research [1].

R.J. Smith (2014) contributes to current conversations and debates about "big data" by developing long-lasting criticisms of sociological methods and study.It concludes by outlining the contributions and approaches that such an analytical qualitative research program might render in the light of the ‘digital turn,' and is useful to those employed at the intersection of conventional and digital(ised) inquiries and methods [2].

The influence of the educational intelligent economy from the perspective of digital frontierism is discussed using a decolonial context, with a special emphasis on Big Data and data sharing in Comparative and International Education, by B.H. Nordtveit et al. (2019). (CIE).Latest controversies regarding CIE's past records and current directions are discussed in order to assess the consequences for data sharing. Finally, the authors argue that, rather than the stagnancy of Big Data and data mining, decolonial participatory research designs that strive for constructive, long-term transitions should be used to resolve the problems inherent in the Educational Intelligent Eco[3].

By focusing on the perspectives of service science and activity theory, Uden, L. et al. (2018) attempts to establish a conceptual framework for translating Big Data into organizational value. The study also provides impetus for researchers and academics to discuss emerging patterns and consequences of Big Data for intellectual capital (IC), which is in line with the agenda for evolutionary research on IC.Big Data transformation for IC management focuses on the value creation process using a collection of essential dimensions to define priorities, key players and stakeholders, processes, and motives. The framework offers guidance for companies looking to take advantage of the new Big Data model for IC management to gain a competitive advantage [4].

L. Ardito et al. (2019) review and categorize the literature on Big Data analytics and management. An objective bibliometric analysis is carried out, which is augmented by subjective evaluations focused on research on the intersection of Big Data analytics and management fields. Descriptive statistics and document co-citation analysis are presented in greater detail. Four clusters depicting literature linking Big Data analytics and management phenomena are exposed as a result of the document co-citation review and evaluation: theoretical development of Big Data analytics; management transition to Big Data analytics; Big Data analytics and firm capital, skills, and performance; and Big Data analytics for supply chain management [5].

By exploring the possible impacts of Big Data Analytics on knowledge use in an organizational and supply chain sense, Kache, F. et al.(2017) contribute to theory growth in SCM. The exploratory research will provide insights into the opportunities and challenges arising from the implementation of Big Data Analytics in SCM, as it is critical for companies in the supply chain to have access to up-to-date, reliable, and meaningful information. Despite the fact that Big Data Analytics is gaining popularity in industry, empirical research on the subject is still limited. The current knowledge base is small due to the lack of comparable content at the intersection of Big Data Analytics and SCM, and no empirical study has been provided so far directly evaluating opportunities and challenges on a corporate and supply chain level with a particular emphasis on the implications placed by Big Data Analytics [6].

R. Salazar-Reyna et al. (2020) review and synthesize the literature on data analytics, big data, data mining, and machine learning in healthcare innovation systems. The research area continues to be expanding, with new research areas emerging and applications being explored.A social network analysis was also used to identify the key authors and collaboration groups publishing in this field. This could lead to the discovery of researchers with similar interests in the field by new and existing authors [7].

S. Virkus et al. (2020) present the findings of a study looking at the new field of data science from the viewpoint of library and information science (LIS). Data science education and training; data professional knowledge and skills; the role of libraries and librarians in the data science movement; data science methods, techniques, and applications are the six general categories. Data science from the viewpoint of information management, as well as data science from the standpoint of health sciences. The authors concentrated on data science methods, techniques, and applications the most, followed by data science in the sense of health sciences, data science education and training, and data technical knowledge and skills [8].

M. Mariani addresses the evolution of Big Data (BD) and Analytics in the tourism and hospitality industry. It explores the significant role that BD has played in tourism and hospitality research to date and predicts how it will change in the future.According to the findings, tourism and hospitality researchers are becoming more aware of and embracing BD approaches to retrieve, capture, analyze, report, and visualize their data. However, as both sets of approaches and technologies, there are a range of ways to enhance the use and analysis of BD and BD analytics. Furthermore, BD analytics have the potential to improve a range of data-driven emerging technologies in tourism and hospitality, such as AI and IoT. As a result, the authors assume that the tourism and hospitality literature will form a new digital entrepreneurship sector. Future research directions at the intersection of BD, tourism, and hospitality are outlined [9].

N., S., P. Samuel, and others (2019) Telecommunications plays a critical position in the modern era's technological advancement. Every second, the number of smartphone users with multiple SIM cards grows. As a result, telecommunications is an important field where big data technologies are needed. Due to high consumer turnover, telecommunication companies compete intensely. Customer retention is a big issue in the telecom industry. The IRA's success in predicting mobile user attrition when combined with an agent-based model. The following are the benefits of this proposed model: the user churn prediction method is easy, cost-effective, scalable, and distributed with good business profit [10].

New related issues

1. Scalable architectures for parallel data processing: For offline or online data processing, a Hadoop or Spark environment is used. The industry is looking for scalable architectures to accommodate big data parallel processing. While there has been a lot of change in recent years, there is still a lot of room for improvement.

2. Using a distributed cloud to handle real-time video analytics:- Videos have become a popular medium of data sharing as internet connectivity has improved, even in developed countries. In this regard, telecom infrastructure, operators, IoT implementation, and CCTVs all play a part.

3. Large-scale graph processing efficiency: One field that necessitates efficient graph processing is social media analytics. The reference article [4] goes into great detail about the role of graph databases in big data analytics. Working on efficient graph processing on a large scale remains a fascinating problem.

4. Detect fake news in real time:- As fake news spreads like a virus in a bursty manner, dealing with it in real-time and at scale is a pressing problem. The details could come from Twitter, false URLs, or WhatsApp. It could seem to be an authenticated source, but it may also be a hoax, making the problem more difficult to solve.

5. Approaches to Dimensional Reduction for Large-Scale Data:- To handle large scale data, one can either expand existing dimensionality reduction approaches or suggest new ones. This includes elements of visualization as well. To begin, one can use existing open-source contributions and then contribute back to the community.

6. In the context of big data analysis, there are many ways to deal with uncertainty: - In big data analysis, there are a number of ways to deal with uncertainty [4]. This involves sub-topics like how to learn from low-veracity data and training data that is incomplete or inaccurate. When there's a lot of unlabeled results, how do you manage the uncertainty?

7. Detecting Anomalies in Extremely Large-Scale Systems: Anomaly detection is a common problem, but it is difficult to solve on a wide scale in real time. Health care, telecom, and financial domains are among the application domains available.

8. Algorithms are automatically converted to MapReduce problems: In the world of big data, MapReduce is a well-known programming model. It's not just a map and reduce function; it also gives applications scalability and fault tolerance. However, there aren't many algorithms that explicitly help map-reduce [3-6].

IDEAL SOLUTION & PRESENT STATUS

Data science is a discipline that deals with both unstructured and organized data and encompasses anything related to data cleansing, planning, and analysis. Statistics, statistics, programming, problem-solving, data capture in innovative ways, the desire to view things in new ways, and the activity of cleansing, planning, and aligning data are all part of data science. This umbrella term refers to a variety of techniques for extracting knowledge and information from data. Big data refers to large amounts of data that can't be processed efficiently with today's applications. Big data processing starts with raw data that hasn't been aggregated and is therefore too large to fit in a single computer's memory.Big data is a buzzword for massive amounts of unstructured and organized data that can inundate a company on a daily basis. Big data is used to analyze information that can contribute to more informed decisions and strategic business steps. “Big data” is characterized by Gartner as “high-volume, high-velocity, or high-variety information assets that necessitate cost-effective, creative forms of information processing to allow enhanced insight, decision making, and process automation [7-8].”

Suggestions to implement Research Activities

- Devise a Big Data Analytics Plan.
- Choosing the required information.
- Making use of the required data science applications.
- Develop a methodology for analysing data.
- Adhere to the cloud-based business model.
- Perform a pilot analysis.
- Make analytics a part of the decision-making process [9].

Conclusion

The authors of this paper have provided a comprehensive view of Big Data procedures and Big Data Analysis approaches as presented in a normative slice of literature. The aim of the presented study was to evaluate, synthesize, and present a detailed systematic review on Big Data and Big Data Analysis based on the results of previous research studies to help in the direction-finding of future research [16].

The SLR methodology used proved to be a useful tool for performing descriptive literature reviews, with benefits such as the synthesis of key observations, the identification of literature voids, and the creation of a basis for future research.The results of this systematic literature review will help academics and practitioners in both Big Data and Big Data Analysisin developing new ideas based on the problems presented in this paper [17].

Analyzing the various aspects of big data reveals that data science has a number of shortcomings that must be overcome by research and development in a range of fields Advances in data science are needed at all stages of the big data life cycle so that business and analytical users can navigate through a variety of data and knowledge, check theories, analyze trends, and make informed data-driven decisions.Only with an in-depth understanding of manageable and actionable data will the importance of big data be realized [10].

The most difficult aspect of Data Science technology is dealing with a wide range of data and knowledge. Data Science is a field of research that entails using various scientific methods, algorithms, and processes to derive information from large volumes of data. Data Science concepts such as statistics, visualization, deep learning, and machine learning are all significant. Discovery, Data Preparation, Model Planning, Model Development, Operationalize, and Communicate Findings are all phases in the Data Science Process. Business Intelligence forecasts are looking backwards, while Data Science predictions are looking forward. 1) Internet search is an essential application of data science. 2) Systems of Recommendation 3) Identification of images and expression 4) The gaming industry 5) Compare prices online [18].

References

[1] Mariani, M. and Borghi, M. (2021), "Are environmental-related online reviews more helpful? A big data analytics approach", *International Journal of Contemporary Hospitality Management*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/IJCHM-06-2020-0548

[2] Smith, R.J. (2014), "Missed Miracles and Mystical Connections: Qualitative Research, Digital Social Science and Big Data", *Big Data? Qualitative Approaches to Digital Research* (*Studies in Qualitative Methodology, Vol. 13*), Emerald Group Publishing Limited, pp. 181-204. https://doi.org/10.1108/S1042-319220140000013011

[3] Nordtveit, B.H. and Nordtveit, F. (2019), "The Educational Intelligent Economy and Big Data in Comparative and International Education Research: A Decolonial Vision", Jules, T.D. and Salajan, F.D. (Ed.) *The Educational Intelligent Economy: Big Data, Artificial Intelligence, Machine Learning and the Internet of Things in Education* (*International Perspectives on Education and Society, Vol. 38*), Emerald Publishing Limited, pp. 33-48. https://doi.org/10.1108/S1479-367920190000038003

[4] Uden, L. and Del Vecchio, P. (2018), "Transforming the stakeholders' Big Data for intellectual capital management", *Meditari Accountancy Research*, Vol. 26 No. 3, pp. 420-442. https://doi.org/10.1108/MEDAR-08-2017-0191

[5] Ardito, L., Scuotto, V., Del Giudice, M. and Petruzzelli, A.M. (2019), "A bibliometric analysis of research on Big Data analytics for business and management", *Management Decision*, Vol. 57 No. 8, pp. 1993-2009. https://doi.org/10.1108/MD-07-2018-0754

[6] Kache, F. and Seuring, S. (2017), "Challenges and opportunities of digital information at the intersection of Big Data Analytics and supply chain management", *International Journal of Operations & Production Management*, Vol. 37 No. 1, pp. 10-36. https://doi.org/10.1108/IJOPM-02-2015-0078

[7] Salazar-Reyna, R., Gonzalez-Aleu, F., Granda-Gutierrez, E.M.A., Diaz-Ramirez, J., Garza-Reyes, J.A. and Kumar, A. (2020), "A systematic literature review of data science, data analytics and machine learning applied to healthcare engineering systems", *Management Decision*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/MD-01-2020-0035

[8] Virkus, S. and Garoufallou, E. (2020), "Data science and its relationship to library and information science: a content analysis", *Data Technologies and Applications*, Vol. 54 No. 5, pp. 643-663. https://doi.org/10.1108/DTA-07-2020-0167

[9] Mariani, M. (2019), "Big Data and analytics in tourism and hospitality: a perspective article", *Tourism Review*, Vol. 75 No. 1, pp. 299-303. https://doi.org/10.1108/TR-06-2019-0259

[10] N., S., Samuel, P. and Chacko, M. (2019), "Feature intersection for agent-based customer churn prediction", *Data Technologies and Applications*, Vol. 53 No. 3, pp. 318-332. https://doi.org/10.1108/DTA-03-2019-0043

[11] Williams BA, Brooks CF, Shmargad Y**.** (2018)**.** How algorithms discriminate based on data they lack: challenges, solutions, and policy implications. *J. Inf. Policy* 8: 78–115

[12] Yarkoni T, Westfall J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12(6): 1100–22

[13] Wilkerson J, Casas A. (2017). Large-scale computerized text analysis in political science: opportunities and challenges. *Annu. Rev. Political Sci.* 20: 529–44

[14] Voigt R, Camp NP, Prabhakaran V, et al. (2017). Language from policy body camera footage shows racial disparities in officer respect. *PNAS* 114(25): 6521–26

[15] Roberts M, Stewart B, Tingley D, Lucas C, Leder-Luis J, et al. (2014). Structural topic models for open-ended survey responses. *Am. J. Political Sci.* 58(4): 1064–82

[16] Alvarez RM, ed. (2016). *Computational Social Science: Discovery and Prediction (Analytical Methods for Social Research)*. Cambridge, UK: Cambridge Univ. Press

[17] Athey S. (2018). The impact of machine learning on economics. Draft chapter, Natl. Bur. Econ. Res., Cambridge, MA. http://www.nber.org/chapters/c14009.pdf

[18] Enos RD. (2016). What the demolition of public housing teaches us about the impact of racial threat on political behavior. *Am. J. Political Sci.* 60(1): 123–42