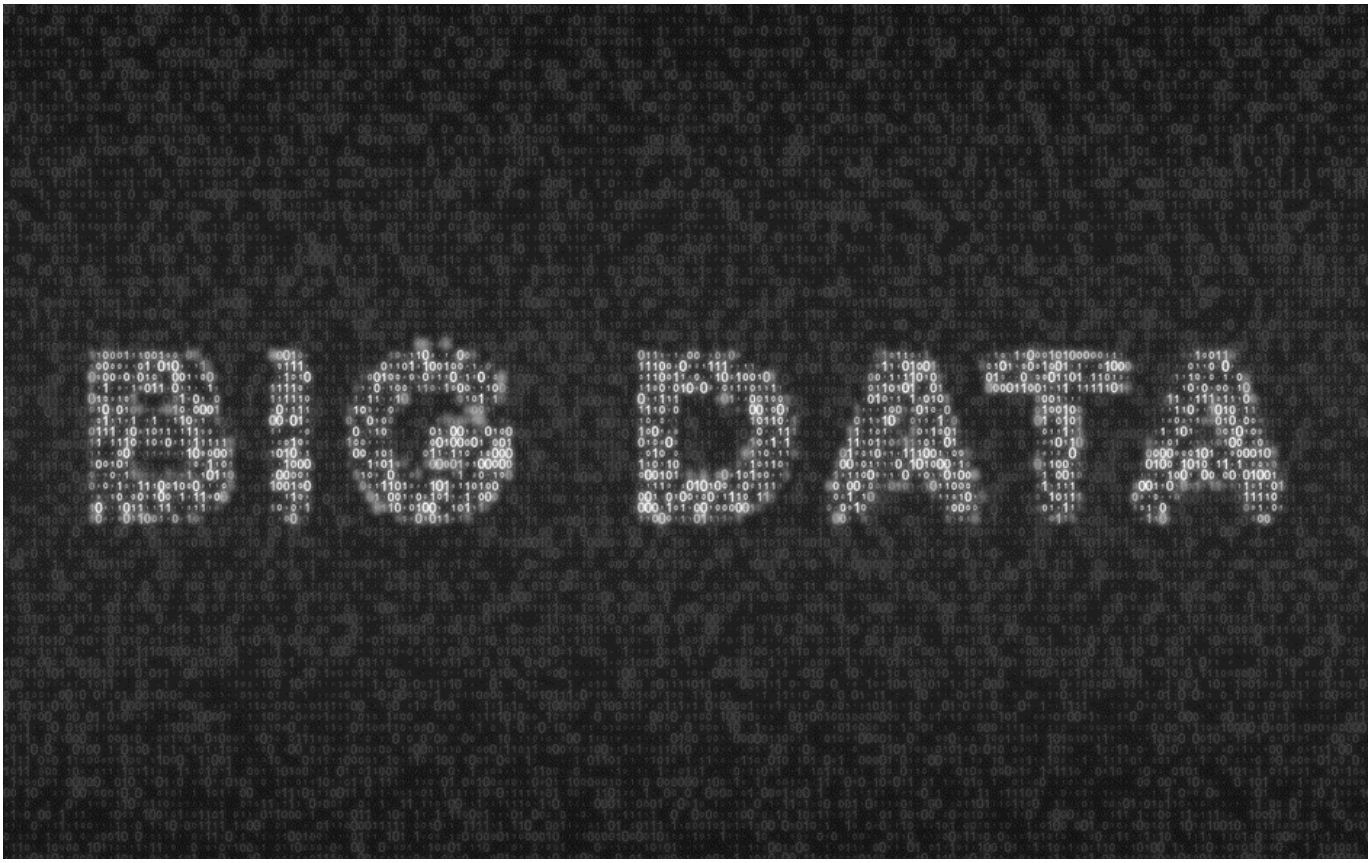
Big Data: Applications, Benefits and Limitations

Dhruv Nyati
asdnyati@gmail.com



Abstract

This research proposal aims to carry out a comprehensive investigation of the use of big data in the healthcare industry, stock market, surveillance systems, and smart cities. This research work looks into the transformative outcomes, possibilities, and difficulties related to using big data in these fields. This research seeks to provide a thorough knowledge of the possibilities and constraints of big data technology by examining real-world case studies. The results also promote innovation and efficiency in healthcare, finance, security, and urban development while assisting in strategic planning and informed decision-making in these vital industries.

Keywords: Big Data, Stock Market, Surveillance Systems, Healthcare Industry

Introduction

In a world with never ending growth of data production, data usage and harnessing the great power of this data, Big Data in Data Science has emerged as the biggest technological advancement across all the sectors. The confluence of technology-driven data collection, processing capabilities, and advanced analytics has paved the way for tremendous development. There are several applications of Big Data in various fields such as the healthcare industry, share markets, surveillance systems, smart cities, and urban planning.

In the healthcare industry, the impact of Big Data will transform patient care, treatment outcomes and disease management. Through the analysis of patient information, medical records and diagnostic data healthcare, professionals can gain valuable insights into disease trends, personalize treatment plans, and predict outbreaks.

In share markets, Big Data analysis can have one of the greatest impacts by accurately predicting the flow of trade.

Surveillance systems, smart cities, and urban planning all result in a better society, safer neighborhoods, cleaner environment, and a more sustainable growth. This means that with the help of various data sources, anyone's security will not be compromised. The urban planning will lead to optimization of transportation systems, energy consumption and infrastructural growth. This will ultimately lead to a more sustainable place to live.

Literature Review

Big Data, basically, is the collection of various small data sets. These data sets may range from simple to complex data. These can therefore be grouped under three categories of data types. The data types are-

- Structured Data - This is a very organized type of data and so is the easiest to analyze. The data present is grouped by a certain structure and in fixed fields like a row or a column. The structured data is stored in data warehouses which filter the data to fit the structure. This data type includes spreadsheets and formed databases. A practical usage of this type is in online transactions and banking systems.
- Unstructured Data - This type of data is very unorganized and unstructured. Finding patterns and analyzing unstructured data sets is a very difficult job as it may have very complex algorithms or would have to be converted to structured data sets and then analyzed. Both of these ways are very time consuming. The unstructured data is stored in data lakes which store data in the original state with all of its information. This data type includes videos, audios, images, chats and much more.
- Semi-structured Data - This type of data is organized only to a key-value pair limit and so is different from structured and unstructured data types. This data is not very rigid in its structure but has the assignment of data sets into different groups. This includes location tags, device ID stamps and metadata making semi-structured data very helpful in machine learning.

Without Big Data processing, Big Data is just a very complex collection of data with no particular use. The processing of it is what makes it useful. There are various stages of this processing of Big Data.

- Data Extraction - As the name suggests, it is the extraction of useful data from Big Data sources. This has to be uniform and accurate so that the results in the later stages are accurate and thus give useful information.
- Data Transformation - This is the stage where the data is modified to fit some criteria necessary for the planned project. Unstructured and semi-structured data are transformed to structured data and then to a format easily understandable by the developer.
- Data Loading - The transformed data is then loaded in the central database where all the algorithms are performed. This may happen in real-time to give results in a very less amount of time.
- Data Visualization - This helps with visualizing the data to get the whole overview of the planned project. In easy terms, the required information is highlighted for an overview of the whole project.
- Machine Learning Application - This is the stage where models and algorithms are created to form a response for the data. Analyzing large amounts of data and making sense out of it as well as providing useful insights and steps to further take so that the profits are maximum.

For processing of Big Data, frameworks like Hadoop and Kafka are widely used. Hadoop is a framework used for managing and processing large amounts of data. The technique it uses is parallel processing where it bifurcates the data into smaller loads and processes them at high speeds. Kafka is another software used to stream data in real-time. Both of these softwares are integrated by developers to use the advantages of both and an increase in efficiency. Big Data is always exponentially increasing and with all the data sets, it becomes more and more complex. To process this in real-time, great speeds are required which sometimes are not achieved leading to choking of the network.

Finally, this research yearns to study the role of Big Data in four discrete sectors that are interrelated through Big Data: the health care sector, share markets, surveillance systems, and smart cities. With the growth of each sector, use of Big Data has increased, utilizing its power for informed decision-making, optimizing resource allocation, identifying disease and the individualized medicine for it, among various other advantages and applications of Big Data. Moreover, this research will also discuss the inherent limitations Big Data possesses.

This paper aims to help society by noting various applications of Big Data in the healthcare industry, stock market, surveillance systems, and smart cities. By analyzing the nature of Big Data and its usefulness in the sectors mentioned before, this paper throws light at the benefits our society and mankind in general can utilize to improve the standard of living. This standard includes better health, safer trading, better environment, and better living conditions.

This research work aims to address the following research questions for each domain:

1. In the healthcare sector, how has the integration of big data analytics improved patient care and healthcare management, while also considering ethical and privacy implications and identifying adoption challenges?
2. In the stock market, what is the impact of big data analytics and machine learning on stock market predictions, trading strategies, sentiment analysis, and the potential for market manipulation?
3. For surveillance systems, how does big data enhance security and threat detection, and what ethical concerns arise?
4. In the context of smart cities and urban planning, how are big data technologies utilized for urban development, resource optimization, and citizen engagement, and what challenges and opportunities do they present?

Methodology

To accomplish this research work, a sophisticated research strategy is employed. This research strategy is divided into four phases. First, a comprehensive literature review is conducted to better establish the foundational understanding of Big Data and its various benefits and limitations in the healthcare sector, stock market, surveillance systems, and smart cities/urban planning. Second, a structured data collection plan, which includes data acquisition from various sources such as research documents and surveys. This plan is utilized to analyze the collected data using statistical methods. A small-scale simulation is performed over the analyzed data, Big Data for this experiment, using the methods and algorithms used for actual Big Data. This stimulation allows us to learn the possible developments which may be replicated on a full-scale operation. Third, the findings in all the domains are compared to each other, drawing parallels and distinctions to find and derive meaningful insights. Fourth, the findings are represented through data visualizations, charts, reports, case studies and conclusions will be drawn to address the research questions. The mentioned case studies are prevalent in the real world and the observations from them immensely help with the research.

Case Studies

1. AlphaFold

AlphaFold is an artificial intelligence program developed to predict protein structures. This program has been very useful in drug discovery because of its prediction of the human proteome, a full set of proteins. It uses previously identified protein structures as a reference to find the correct orientation of proteins. This input of identified protein structures was done with the help of Big Data. This proves the necessity of Big Data to provide artificial intelligence and machine learning programs with sample data to understand and be able to find patterns for future use.

AlphaFold is one of the most impactful programs which uses Big Data in the Healthcare Sector. This program shows that Big Data truly comprises everything known and can lead to finding the unknown, showcasing the impeccable use of Big Data. This use case has reduced the time of finding one structure with 10^{300} possibilities to mere minutes with more than 92% accuracy.

2. Traffic Management and Surveillance

The Government of Goa had implemented an artificial intelligence traffic management system in March 2023. This system has reportedly reduced traffic and corruption by a significant amount while also increasing security on the road. The system has been efficient in identifying traffic and then changing the signal timing accordingly to reduce the traffic. It has also reduced corruption in the State of Goa by issuing challans online to the vehicles that were identified breaking rules, thus also reducing human intervention.

The tedious job of traffic management can now be automated which not only reduces the workload but also increases the efficiency with which these tasks are done. The system is also trained to identify accidents and call medical assistance for the same. Although small accidents happen more often than fatal ones, alerting medical assistance is not required for them and so proves a flaw with the identification of fatal accidents. Therefore, more training is required in identifying accidents and categorizing them.

The A.I. system uses images and videos of traffic as a sample to learn and identify real traffic in real-time. It also accesses the vehicle database to send e-challans to the respective vehicle owners. These data values come under Big Data, thus proving the importance of Big Data in surveillance also.

3. CLARIFY

CLARIFY Project (Cancer Long Survivors Artificial Intelligence Follow Up) is a research project aimed to determine the personalized factors that predict poor health status after cancer treatment. Their vision is: “Big Data and Artificial Intelligence techniques will be used to integrate all available patient´s information with publicly available relevant biomedical databases as well as information from wearable devices used after the treatment”.

This is another program which uses Big Data impeccably and has found several relations between cured cancer patients and the poor health they face after treatment. Their findings provide insights for more research towards the betterment of cancer survivors and also becomes a source for insightful data in Big Data.

The data they provide is beneficial for everyone and puts in light another advantage of Big Data.

4. **Tay, Microsoft A.I. Chatbot**

Tay was an A.I. chatbot created by Microsoft that was able to chat with individuals on Twitter (now X). Every interaction of Tay with different individuals acted as a source of Big Data for its machine learning processes. Within the span of 24 hours, Tay had to be shut down because of its racist and abusive comments. These comments were a result of the millions of people tricking Tay and providing those comments to Big Data. Tay eventually learnt to utilize harmful information from this Big Data.

Similar to Microsoft, Facebook also created A.I. chatbots. The Facebook A.I. Research Lab (FAIR) created chatbots with an aim to continue their A.I. research. Although the chatbots were programmed to use plain English as a medium to converse, these chatbots soon started to converse in a form of English unknown to humans. In other words, the chatbots created a language of their own without any human input.

Therefore, this arises a limitation: machine learning algorithms, when performed over Big Data that contains substandard data, will result in degrading and detrimental results, and ultimately the collapse of the A.I. model.

5. **ASCAPE**

ASCAPE, Artificial intelligence Supporting CAncer Patients across Europe, is a project which is aimed at taking the advantage of the recent advances in Big Data, A.I. and Machine Learning to support cancer patients' quality of life and health status.

This project aims to deploy many services for cancer patients like physiological and psychological support, improved patient and family counseling and guidance, early diagnosis and forecasts of ill-health, identification of disease trajectories, etc. These services will run on A.I. algorithms which process data to interact from Big Data. Additionally, with increasing number of patient interactions, new results will be shared to different A.I.s and Big Data for more breakthroughs in this field.

This is one more advantage of Big Data in the Healthcare Sector.

6. **Amazon's Recruitment A.I. Tool**

In the year 2015, Amazon disbanded its A.I. tool which was created to facilitate recruitment of applicants for jobs at Amazon. The major issue with this A.I. tool was that it separated the applicants according to a gender bias. The tool taught itself that male applicants were preferable and downgraded the application of female candidates. This gender bias was a result of the data provided to perform the algorithms. According to that data, most applicants were males and so the A.I. preferred males for the job.

This case presents another limitation, if poor-quality data is provided in Big Data, it will corrupt the A.I. and Machine Learning systems originally created to ease human work.

These case studies constitute a very small amount of the numerous practical uses of Big Data in A.I. powered programs which are capable of helping the society and increasing human ease.

Results

1. Healthcare

Big Data can help individual patients: predict diseases and illnesses they are prone to, prevent the diseases and illnesses beforehand, cure the diseases and illnesses with personalized remedy, and save them from potential side effects of the medicines used for the general masses. The data for the same may be collected from various sources that monitor the patient's health like wearable devices, previous surgery and medicinal records and lifestyle. This thus enables customized healthcare focused on curing as well as preventing.

An add-on to the previous results, a virtual assistant can be created using the A.I. and Big Data for each person on this planet. This provides faster and more accurate results. The limitation following this is that many non-practising doctors may lose their occupation to A.I.

Big Data can also contribute to A.I. and Machine Learning programs that can predict and detect disease outbreaks. The same program may also help in the generation of a cure or vaccine to eradicate the disease as early as possible. The data from this may be collected from public health records, animal health records, history of the disease, and similar sources. An advantage of this is that until the cure is found, the program may also generate necessary guidelines and rules for the patients as well as the public to follow and reduce the chances of a potential health crisis.

2. Stock Market

The use of Big Data in the stock market and trading sector has increased tremendously over the past few years. Algorithmic trading, a set of automated and pre-programmed trading instructions that account for price, timing, and volume to give the best result possible, has since proven better for the traders than traditional trading. Although algorithmic trading has been better for investors, it does have flaws which have resulted into market crashes like the flash crash of 2010, but since then the technology has evolved and with new and better data to be incorporated for the Machine Learning Programs which may result in a better version of algorithmic trading. This may include the incorporation of social media as one of the biggest influencing factors in today's time. For example, Elon Musk's tweets and the rise of Dogecoin. So, with incorporating influencing factors, a better version of automated trading can be installed for everyone to invest with. This in turn will reduce the overall risk and provide optimization to everyone using it.

3. **Surveillance Systems**

With newer technologies being invented every day, surveillance equipments are also being upgraded with better detection and information softwares. Big Data can help with the detection and information softwares as many A.I. prompts are now able to detect a person with great accuracy and so integrating these in the already present surveillance systems can ensure safer environments in the urban and rural areas. In extreme cases, these systems may also monitor criminals for their actions and inform the authorities of suspicious activities. These systems can also ensure road safety, and ethical usage of power by authorities, thus reducing crime rates. These regulations and advantages take our society one step closer to become a utopian society.

4. **Smart Cities and Urban Planning**

Big Data with A.I. and Machine Learning may help our society against natural disasters with better and accurate predictive models of cities and countries prone to respective natural disasters. The systems created may stimulate drills and prepare the masses for effective evacuation during these times. These may also allocate resources usefully in case a disaster strikes.

A.I. algorithms created using Big Data may manage energy resources with far greater efficiency than humans as these algorithms may analyze data from smart meters, weather forecasts, and fossil fuel management, and find patterns not visible to humans.

A.I. may also generate a centralized transportation system that integrates most transportation modes, optimizes the routing system, and sustains resources.

A limitation that may arise is that the capital needed to perform these will be very high but the cost will be a one-time cost and will provide ease with city planning.

Discussion

A.I. is a very prevalent technology and will make most of our future but it is important for us to use it carefully. For any experiment, good quality input materials will result in good quality final products and also a better yield. Similarly, the input we provide as Big Data to the A.I. tools largely determine the quality of the output we receive. To reduce the limitation of poor data quality in Big Data some measures should be followed while collecting the data. The measures:-

- the collection of data should not be biased nor favorable to any situation i.e. a fair and equal chance to every situation
- the collection of data should account for various variables and should be spread over large demographic areas that account for the variables.

Another limitation of Big Data is that while many people voluntarily provide their data, most of the data procured is without the consent. For example, while collecting health records, hospitals may provide the data without the knowledge of the patient because of which arises a

breach of personal privacy. This limitation questions the ethical consideration of data provided in Big Data.

A.I. and Big Data technologies are the future and the limitations they pose are operable to form the technology of the future.

Conclusion

Data is now commonly referred to as the 'Gold of the 21st Century' which makes sense given how data has become valuable and increased exponentially over the years. This exponential growth of data helps Big Data as the source or collection of data from which A.I. and Machine Learning programs receive data. These programs run on softwares like Hadoop and Kafka which give an easier work through from the data to the optimal results. These results; when customized for the various sectors like healthcare, stock market, surveillance systems, and smart cities and urban planning; can provide a better lifestyle and better society. Although for the most part, these results are helpful, and if implemented well, can change the sector altogether, the Big Data applications also have limitations that have the possibility to hinder our path to a better society. Although the advantages of Big Data outnumber the limitations of it, the limitations still have the potential to disrupt this technology.

References

[1] A. Oussous, F.-Z. Benjelloun, A.A. Lahcen, S. Belfkih (2018). *Big Data Technologies: A Survey* https://doi.org/10.1016/j.jksuci.2017.06.001

[2] J. Nahar, T. Imam, K.S. Tickle, Y.-P.P. Chen (2013). *Computational intelligence for heart disease diagnosis: a medical knowledge driven approach* https://doi.org/10.1016/j.eswa.2012.07.032

[3] H.B. Abdalla (2022). *A brief survey on big data: technologies, terminologies, and data-intensive applications* https://doi.org/10.1186/s40537-022-00659-3

[4] P.V. Desai (2018). *A survey on big data applications and challenges* https://doi.org/10.1109/ICICCT.2018.8472999

[5] H. Luan, P. Geczy, H. Lai, J. Gobert, S.J.H. Yang, H. Ogata, J. Baltes, R. Guerra, P. Li, C.-C. Tsai (2020) *Challenges and Future Directions of Big Data and Artificial Intelligence in Education* https://doi.org/10.3389/fpsyg.2020.580820

[6] A. David, S. Islam, E. Tankhilevich, M.J.E. Sternberg (2022) *The AlphaFold Database of Protein Structures: A Biologist's Guide* https://doi.org/10.1016/j.jmb.2021.167336

[7] https://www.labellerr.com/blog/6-smart-surveillance-use-cases-using-vision-ai/

[8] https://opengovasia.com/goa-india-unveils-ai-based-traffic-management-system/

[9] https://www.geeksforgeeks.org/types-of-big-data/

[10] https://cloud.google.com/learn/what-is-hadoop

[11] https://hevodata.com/learn/big-data-processing/

[12] https://www.reuters.com/article/idUSKCN1MK0AG/

[13] https://ascape-project.eu/the_project/overview

[14] https://www.clarify2020.eu/about_us/
[15] https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/
[16] https://www.firstpost.com/tech/news-analysis/facebook-researchers-shut-down-ai-bots-that-started-speaking-in-a-language-unintelligible-to-humans-3876197.html