The Impacts of Data Science in Everyday Life

Abraham Neme Alvarez* Hannes Lötsch†

Berliner Hochschule für Technik

ABSTRACT

The field of data science has brought a significant transformation in decision-making processes and our interaction with the digital realm. This paper investigates the profound impact of data science on our everyday lives, delving into its core principles, diverse applications across various domains, technical obstacles encountered, and the emergence of novel fields of study and employment opportunities. The paper highlights the crucial stages of data science, which play a vital role in facilitating informed decision-making. Furthermore, it explores how data science has positively influenced sectors across the hole spectrum of industries while simultaneously addressing the challenges posed by it. Moreover, it sheds light on the birth of data analytics and artificial intelligence as exciting areas of study that are shaping the future job market. Lastly, ethical concerns pertaining to privacy and algorithmic bias are examined, considering their tangible impact on individuals' daily lives.

Index Terms: data science—big data—data center—data analytics artificial intelligence—machine learning—data privacy—

1 INTRODUCTION

Data science has emerged as a revolutionizing field that harnesses the power of data to drive decision-making across multiple fields and industries, as well as transforming our interaction with the digital world. This paper seeks to explore its impact on our daily lives and address crucial issues. To begin the fundamentals of data science will be explained, exploring as well its process from collection to interpretation, and its importance in informed decision making. It will be then examined how it is being implemented in various fields, such as medicine, transportation and security, improving the quality of life and providing innovative solutions. Data science faces nonetheless significant technical challenges, such as handling large volumes of data and ensuring its quality and integrity. This paper will also inquire in some innovative approaches that are being developed to overcome these challenges and how data science has given rise to new branches of study and work, such as data analytics and artificial intelligence, which are shaping the employment landscape and fostering innovation. Finally, ethical concerns will be addressed, such as privacy and algorithmic bias, and how they affect people's daily lives.

2 UNDERSTANDING DATA SCIENCE

Data science encompasses the multidisciplinary field that combines statistical analysis, machine learning, and computational techniques and processes to unlock insights and knowledge from data. This involves several interconnected stages, each building upon the previous one to deliver meaningful outcomes. Data science is composed by the processes of data collection, cleaning, preparation, analysis, modeling, and interpretation. Let's explore each of the stages that compose the journey of data science:

*s91232@bht-berlin.de

†s91281@bht-berlin.de

- **Data collection:** It all begins with gathering data from various sources, such as databases, sensors, social media, or web scraping. This diverse collection allows for a comprehensive understanding of the subject under study and ensures a holistic view.
- **Data cleaning and preparation:** Raw data is often messy, with inconsistencies, missing values, and errors. Data scientists employ techniques to clean and preprocess the data, ensuring its quality, accuracy, and reliability for subsequent analysis.
- **Data analysis:** This stage involves exploring the data using statistical methods, visualizations, and machine learning algorithms. Through data exploration, patterns, correlations, and insights are discovered, providing a deeper understanding of the data and potential relationships.
- **Modeling and interpretation:** In this stage, data scientists build models and algorithms that encapsulate the patterns and relationships identified during analysis. These models allow for predictions, classifications, or recommendations based on the data, enabling a deeper understanding of complex phenomena.
- **Iterative process:** Data science is an iterative process, meaning that each stage informs the next, and insights gained from one iteration can lead to refining the entire process. Feedback loops and continuous refinement are critical to enhancing the accuracy and effectiveness of data science outcomes.

3 LEVERAGING DATA SCIENCE

In today's data-driven world, organizations across various industries have recognized the immense value of data science techniques in making informed decisions and gaining competitive advantages. By harnessing the power of data, these organizations can gain valuable insights and drive strategic initiatives. Data science enables organizations to gain a deeper understanding of their customers. Via the analysis of vast amounts of customer data, including demographics, purchase history, and online behavior, businesses can identify patterns and preferences. This insight allows them to personalize products, services, and marketing campaigns, resulting in enhanced customer satisfaction and loyalty. By analyzing operational data, such as supply chain processes, manufacturing metrics, or logistics data, businesses can identify bottlenecks, streamline workflows, enhance productivity, and reduce costs. Through data science, organizations can leverage historical data to predict future trends and behaviors. Companies employing techniques like regression analysis, time series forecasting, and predictive modeling, businesses can anticipate market demands, customer preferences, and potential risks. This foresight enables proactive decision-making and strategic planning. Product development and quality improvement is another area in which through analysis of customer feedback, user behavior data, and product usage metrics, businesses can uncover insights to enhance product features, usability, and overall user experience. These data-driven improvements help organizations stay ahead of the competition while benefiting their customers.

However, the advances in this field not only benefit large companies, but also ordinary people all around the world in a wide variety of areas. Many significant advancements in healthcare have been achieved by leveraging large datasets and developing algorithms that help medical professionals make more accurate diagnoses, predict disease outcomes, and personalize treatment plans. Machine learning algorithms, which are another result of advances in data science, can analyze medical records, genetic data, and clinical research to improve patient care and outcomes. Najat Khan Chief Data Science Officer and Global Head at the Janssen Pharmaceutical Companies of Johnson & Johnson talked in an interview about how parmaceuticals are helping patients with this new technologies: "Imagine that a patient has a specific mutation of cancer for which there is a targeted therapy. However, these mutations are not picked up in a routine way because they are new and therefore are not sequenced on a regular basis. But every patient gets their tumor biopsied, so what we have done is used data science to digitize those biopsy slides, and then we use machine learning model to predict what mutation a patient might have just from those images. The impact of that is to find diseases earlier in patients who could benefit from a targeted therapy and may otherwise have been missed." [6] The financial industry has too been revolutionized , enabling faster and more accurate risk assessments, fraud detection, and personalized financial advice. Via the analysis of transaction data and patterns, banks can detect suspicious activities and protect customers' financial assets. Data science techniques are also vital in cybersecurity to detect and prevent threats. By assessing network logs, user behavior patterns and anomaly detection algorithms, governments and companies can identify potential security breaches and fraud attempts. This proactive approach strengthens cybersecurity measures and safeguards sensitive data from unauthorized access.

4 Technical challenges of Data Science

Looking back to the processes of data science it is possible to realize that as the size of the data to be used increases, multiple challenges arise. Analyzing massive datasets, commonly known as big data, presents a unique set of challenges due to the volume, velocity, variety, and veracity of the data. While big data offers immense potential for generating valuable insights, it also requires specialized approaches and technologies to overcome these challenges. First of all to generate big amounts of data, multiple types of hardware are required such as sensors, microprocessors and computers. Incrementing the amount of collection devices requires a considerable commitment and big investments that can limit the hability of most people and small companies to obtain the desired or needed results. After collecting the information, it is necessary to ensure the quality of the data used for the analysis. Data sets may contain errors, outliers or incomplete information, which can affect the accuracy of the models and the results obtained. To address this problem, multiple specialists are needed to select the data to be used and to filter out what is not relevant or may have a negative effects and to further streamline processes and eliminate repetitive tasks, data cleaning and preprocessing techniques are also being developed, including error detection and correction, imputation of missing values and identification of outliers.

Big data often exceeds the storage and computational capabilities of traditional systems. Managing and storing massive volumes of data requires scalable infrastructure, such as distributed file systems and cloud-based solutions. Similarly, processing and analyzing such vast amounts of data necessitate distributed computing frameworks, like Hadoop or Apache Spark that can parallelize computations across clusters of machines. A distributed computing framework refers to a software framework that allows the coordination and utilization of multiple computers (also known as nodes) connected over a network to work collectively. In addition to all these hardware and trained personnel requirements, it is necessary to build infrastructure to house all these systems, also known as data centers. Construction requirements include choosing a suitable site that meets security standards, access to reliable electrical power services, access to high-speed communication networks and scalability for future expansion. The physical infrastructure must be robust and include redundant power supply systems, efficient cooling systems for the computers, and security systems to protect the equipment and stored data. According to an analysis by the commercial real estate services company Lang LaSalle Incorporated "the average-powered base building (defined as foundation, four walls and roof along with common areas for security, loading dock, restrooms, corridors, etc. . .) of a data center facility typically ranges from \$125 US-dollars per ft^2 to upwards of \$200 per ft^2." [3] On top of these costs, getting a data center ready to function with all IT capabilities can cost between "\$280 and \$350 US-dollars per ft^2." [3] These costs can, of course, be much higher depending on the location of construction as well as power and availability requirements. To provide an example, when Facebook started constructing the Prineville Data Center in Oregon USA back in 2009 the company invested more than a billion US-dollars to construct the 1.25 million-square-foot ($116{,}128m^2$) facility.

Data science involves extracting meaningful information from large volumes of data. Nevertheless, proper interpretation of the results can be challenging. Machine learning models and analysis techniques can generate complex results that require deep and expert understanding for interpretation. This is why multiple companies and governments are constantly looking for educated and skilled workers to help them get the right results. Another way data scientists are addressing this challenge is through the development of visualization methods and model explainability techniques that help understand and communicate results more effectively.

5 The emergence of new branches of study and work

The rapid rise of data science has led to the emergence of new fields of study and has created exciting opportunities in the job market. As companies increasingly recognize the value of data-driven decision making, the demand for qualified data science professionals continues to grow.

5.1 Studies on data science

Educational institutions have responded to the demand for data science skills by introducing specialized programs and courses. These programs equip students with the necessary skills in statistics, mathematics, programming, machine learning and data analysis allowing them to acquire the knowledge and proficiency needed to excel in this evolving domain. Within data science there are several prominent career paths, some of which were born alongside this field and represent its foundations are:

- **Data scientist:** These professionals are responsible for collecting, cleaning, analyzing, and visualizing data to gain meaningful insights and make decisions based on it.
- **Data engineer:** Data engineers are specialists in designing, building and maintaining data infrastructures. They focus on developing systems and architectures for data storage, extraction and transformation, and ensuring data availability and quality.
- **Data analyst:** Data analysts focus on interpreting and analyzing data to identify meaningful trends, patterns and relationships.

- **Data architect:** Data architects are responsible for designing and overseeing the technology infrastructure needed to manage large volumes of data. They work on the planning and design of databases, storage systems and data flows.

As the field evolves new opportunities and specialized roles emerge that adapt to the specific needs of certain domains and technological advances. These domain-specific roles require a deep understanding of both the data science techniques and the unique challenges and nuances of the respective fields. Some examples of these study branches are:

- **Machine learning engineer:** With the recent rise of artificial intelligence, machine learning has become a popular field of study all over the world. These engineers play a crucial role in developing and implementing models that can learn from data and make accurate predictions or automate processes.
- **Businesse analyst:** A business analyst the processes in a company and analyses industry trends and markets. Business analysts process enormous amounts of data in order to find opportunities to improve business revenue and growth.
- **Clinical data scientist:** These scientists unite healthcare training with computer science and statistics to actively collect, assimilate, analyze and predict patient diagnoses as well as medical industry trends.

5.2 Labor demand in data science

After all the previous insights of how data science its being leveraged for a wide variety of purposes, it is easy to acknowledge that the demand for professionals in this area has witnessed significant growth and is expected to continue its upward trajectory in the coming years. According to Mckinsey & Company and the Synergy Research Group the global spending on the construction of data centers is forecast to reach $49 billion US-dollars by 2030. [1]

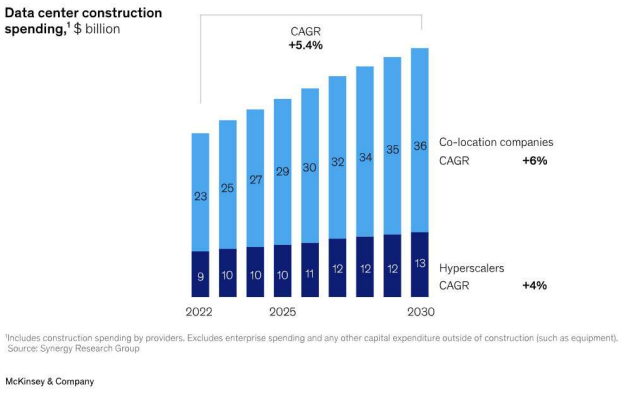

Figure 1: A visualization of the 2022–2030 forecast data of global spending on data centers. The image is from [1].

The increase of investments directly translates to an increase in the projected number of jobs not only for data science specialists, but also for construction workers, maintenance and IT support personnel who play a crucial role in building, upkeeping and upgrading data centers, ensuring the physical infrastructure is in optimal condition to house and protect data. Additionally, the US Bureau of Labor Statistics predicts that "data scientist jobs will experience an impressive 36% growth between 2021 and 2031 in the United States, making it one of the fastest-growing occupations." [5] This growth can be attributed to the expanding adoption of data-driven strategies across industries and the need for professionals who can derive actionable insights from complex data sets, and since the demand for specialists keeps increasing workers can expect a median anual salary of a $100 thousand US-dollars. Furthermore, in the ranking of the 100 best jobs of the digital media company U.S. News & World Report, data scientist ranks 22nd, with other occupations related in some way to data science like software developer, information security analyst and market research analyst ranking 1st, 5th and 15th respectively. [4]

6 Artificial Intelligence

Artificial intelligence (AI) is a sub-discipline of computer science focused on building computers with flexible intelligence capable of solving complex problems using data, learning from those solutions, and making replicable decisions at scale. Machine learning (ML) sometimes also refered to as applied AI is a branch of AI that focuses on developing algorithms and models that can learn from data and perform specific tasks without being explicitly programmed, in other words it offers a path to make artificial intelligence a reality. Data science has played a key role in the advancement of AI and ML in recent years. Since AI requires vast amounts of data, advances in data collection, analysis and application have driven the development of more efficient and sophisticated algorithms, leading to significant improvements in the performance of these technologies. "Fundamentally, machines can not hope to mimic humans' cognitive processes without information and Data scientists are tasked with feeding machines accurate, empirical data and statistical models that enable machines to learn autonomously." [7] There is no definitive answer to the exact number of machine learning algorithms, as the field of machine learning is continuously evolving and new algorithms are being developed. However they can be broadly categorized into three types:

- **Supervised learning:** In supervised learning, the algorithm learns from labeled training data. It is provided with input features and corresponding target labels, and its goal is to learn a mapping function that can predict the correct label for new, unseen inputs. Some popular supervised learning algorithms include linear regression, logistic regression, decision trees, support vector machines (SVM), and neural networks. These algorithms are widely used for tasks such as classification and regression (predicting continuous values).
- **Unsupervised learning:** Unsupervised learning deals with unlabeled data, where the algorithm explores the underlying structure and patterns in the data without any predefined target labels. The goal is to discover meaningful information, such as clusters, associations, or dimensions, that can provide insights into the data. Common unsupervised learning algorithms include k-means and hierarchical clustering.
- **Reinforcement learning:** This is a type of machine learning in which an agent learns to make sequential decisions by interacting with an environment. The agent learns through trial and error, receiving feedback in the form of rewards or penalties for its actions. The goal of reinforcement learning is to maximize the cumulative reward over time by discovering the optimal actions to take in different situations.

Nowadays AI models and systems can be found in a broad range of devices and are used for multiple applications, from recommendation algorithms in search engines and digital platforms such as youtube to models that can predict climate change on Earth based on historic data of weather patterns and pollution emissions. A prime example that demostrates how fast AI is evolving thanks

to the unprecedented quantity of data availability and to the tools and techniques provided by data science is Chat GPT (Generative Pre-trained Transformer) model. Chat GPT is a language model based on the transformer architecture that has been trained on a large amount of textual data from the internet. The transformer architecture is a neural network model that that can understand and generate human language. It does this by processing input sequences in parallel, which helps it understand how words are related to each other. This model is able to generate coherent and contextually relevant responses from an input text. Chat GPT has shown the general population the power of AI and is enabling millions of people to create their own use cases, such as automating email responses, integrating intelligent chatbots into websites, translating text and more. Another notable example is the field of computer vision, where advances in data science have enabled the development of systems for object recognition and pattern detection in images and videos. Computer vision is used in many processes in industry, agriculture and even medicine to automate repetitive tasks such as data collection, product sorting and identification of potential pests or diseases.

7 Ethical Issues

Ethical issues surrounding data science have become a significant concern as the field continues to advance and play a pivotal role in shaping society and the world. One of the foremost ethical dilemmas in data science revolves around the ownership and control of user data. With the proliferation of digital technologies and online platforms, vast amounts of personal information are collected and stored by organizations. The question arises: Who should own this data, and how should it be used? The misuse or unauthorized access to user data can lead to privacy violations, identity theft, and the manipulation of individuals and communities. The lack of clear regulations and standards for data ownership and privacy protection creates a power imbalance between users and organizations, often leaving individuals at a disadvantage. In addition, algorithms used in data science applications may inadvertently promote harmful content and misinformation. Recommendation systems, for example, are designed to personalize content and provide users with relevant information. However, without careful design and oversight, these algorithms can reinforce pre-existing biases and amplify extremist views. This can have far-reaching consequences, such as polarizing public discourse, fostering disinformation campaigns, and undermining democratic processes. For example, in 2019 Facebook researchers created three dummy accounts to study the platform's technology for recommending content in the News Feed. The first was for a user in India, then they created two more test accounts to represent a conservative american user and a liberal one. All three accounts engaged exclusively with content recommended by Facebook's algorithms. Within days, the liberal account, started seeing critics of Republican Senator Mitch McConnell after he blocked bills to protect American elections from foreign interference. The conservative account, was guided toward QAnon conspiracy theories. Meanwhile, the test user's News Feed in India was filled with inflammatory material containing violent and graphic images related to India's border skirmishes with Pakistan. [2]

Artificial intelligence introduces another set of ethical concerns. As AI systems become more sophisticated, there is a risk of biased decision-making, lack of transparency, and loss of human control. Biases in training data can perpetuate discrimination and reinforce societal inequalities. The potential for AI to replace human workers also raises concerns about job displacement and economic inequality. Furthermore, the growing demand for data storage and processing power has led to a surge in data centers, which consume vast amounts of energy and contribute to carbon emissions. The carbon footprint of data centers, combined with the energy requirements for data storage and transmission, presents ecological challenges. According to Mckinsey & Company "in the United States alone, data center energy demand is expected to grow by some 10% a yer until 2030." [1] The extraction of raw materials for manufacturing data storage devices and the disposal of electronic waste further contribute to environmental degradation. Sustainable practices, energy-efficient technologies, and responsible data management strategies are needed to minimize the environmental impact of data science activities.

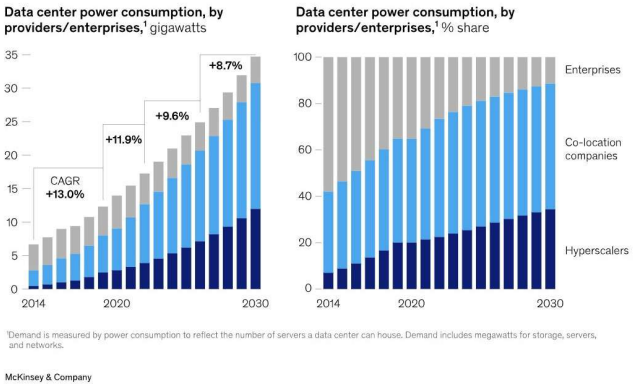

Figure 2: A visualization of the 2014–2030 forecast data of US data center demand. The image is from [1].

8 Conclusion

In conclusion, data science has become an essential discipline in today's data-driven world, empowering organizations, transforming industries, and creating new opportunities. As data science continues to evolve, it will further revolutionize decision-making processes, drive advancements in technology, and unlock new possibilities for solving complex problems across various domains. While achieving the full potential of data science poses many significant challenges, these are being addressed through the development of scalable solutions, data cleansing and preprocessing techniques, AI models and visualization methods. Undoubtedly the job market in virtually every industry has changed and will continue to evolve and expand with the emergence of new technologies and fields of study in data science, but it is also important to recognize and address the issues that come with them. By addressing all ethical and environmental concerns through responsible practices and meaningful collaboration, humanity can harness the power of data science while minimizing its negative impact and fostering a more equitable and sustainable future.

References

[1] S. Bangalore, A. Bhan, A. D. Miglio, P. Sachdeva, V. Sarma, R. Sharma, and B. Srivathsan. Investing in the rising data center economy. McKinseys Technology, Media and Telecommunications, January 2023.

[2] M. Bidar. Facebook researchers saw how its algorithms led to misinformation. CBS News, October 2021.

[3] K. Hawkins and M. Restivo. Data centers: expensive to build, but worth every penny. Jones Lang LaSalle Incorporated, April 2022.

[4] U. News. 100 best jobs. U.S. News and World Report, 2023.

[5] O. of Occupational Statistics and E. Projections. Occupational outlook handbook - data scientists. U.S. BUREAU OF LABOR STATISTICS, May 2023.

[6] T. Robinson. How janssen is leveraging data science to improve the trajectory of human health. Domino Data Science Blog, 2022.

[7] R. University. Data science vs. artificial intelligence and machine learning: What's the difference? Rice University - Deparment of Computer Science, April 2023.