# DATA SCIENCE AND DATA ANALYTICS

Your Name

Course Name and Number

Name of Professor

Institutional Affiliation

Date Submitted

# Table of Contents

## LIST OF FIGURES



## LIST OF TABLES

## ABSTRACT

The 21st century has witnessed a steep rise in data production in every corner of the world. Governments and organizations have collected and analyzed vast quantities of data in the past decades. There has been an ever-rising need to develop and conduct research activities around data science and data analytics. This rising demand has led to the development of curriculums that seek to create more knowledge and understanding of this field. A study on scholarly sources of information on the topic to arrive at an overview of current knowledge, to identify relevant theories, methods, and gaps in the existing research is necessary. The paper contributes knowledge to this area by presenting findings from reviewing relevant literature that explains the concepts of data science and data analysis and explains the differences between the two areas. The vocabulary that will address this problem is the differences between data science and data analytics?

# LITERATURE REVIEW

## Introduction

There has been an acute shortfall in the supply of skilled workforce who can handle data analysis tasks in governments and institutions to meet the rising data analysis needs. This deficit has led to a substantial global need to educate and train more people in this area. The training programs all seek to create proper knowledge of data science and data analysis by equipping individual learners with skills that will enable them to stay relevant and abreast with the evolving data needs of the 21$^{st}$ century. There exist multiple academic literature sources that define the concepts of data science and data analysis. Apart from that, experts and professionals in this field use practitioner literature to teach, learn and share knowledge with other professionals. This literature source provides a strong foundation for research activities that seek to further the study and contribute immense knowledge to it.

## Big Data

Big data is simply large data sets. It is a term used to define vast quantities of structured, semi-structured, or unstructured data. The vast volumes make them unable to be handled by traditional methods of processing data and traditional software. The volume of data keeps continually growing as the business carries out its activities. Big data is beneficial to businesses because businesses analyze it for insights.

Present-day advancements in technology have made it possible to create large quantities of data and store all this data in enormous volumes. The internet has created an explosion in the quantity of data available for individuals and businesses. As more and more people use the World Wide Web and social media sites, the quantity of data produced in that sector alone is immense. For example, Facebook is estimated to receive photo uploads from its users, reaching up to ten million per hour.

## Data Science

Data science can be defined as the study of data to extract useful insights from it. Data science deals with extracting value from vast volumes of data by bringing together several fields, including scientific methods, statistics, and data analysis. Data scientists analyze the data that is collected using a wide range of skills. This data can be collected from the internet, computers and smartphones, business customers, and many other sources.

Data science has gained greater significant importance over the years. This has been fuelled by recent advancements in technology that have defined the technological revolution, a time when humans have resorted to digital means of communicating and carrying out day-to-day activities (Aasheim et al., 2015). Data science is a necessary discipline that utilizes various mechanisms to manipulate data for individuals, organizations, and society. The technological era has created the need for organizations to develop systems through which they collect various data types to facilitate their operations.

Data science has enabled businesses to colonize the global market by providing product development trends to produce more innovative products and services.

## Data Analytics

Data analytics deals with the technologies that allow for a deep understanding and the extraction of insights from data (Cao, L., 2017). This also includes processes and tools that facilitate the process. Data analytics is made up of descriptive analytics, predictive analytics, prescriptive analytics, and diagnostic analytics. During data analytics, raw data and data sets are examined by data analysts to uncover patterns and draw conclusions from the information they hold. Once data has been collected, it has to go through a series of processes to deliver useful information. The data has to, first of all, be inspected. This is important in ensuring that the data held is the correct one needed by the system. The data is then cleansed to get rid of any unwanted bits that might compromise its integrity. After that, the data is then transformed into a form that is usable for the system. It can then be modeled for practical organizational use, including organization decision making.

**Types of Data Analysis**

Data analysis is crucial for both individuals and institutions. It helps them convert the data into usable forms to suit their objectives and needs. Data analysis can be broken down into four main categories.

I. Descriptive Analysis. This is a type of data analytics that describes the data using statistics to gain information from it.

II. Predictive Analysis. This is the type of data analytics that uses available information to predict possible future occurrences that are not known. It also explains the reasons behind these occurrences.

III. Prescriptive Analysis. This type of data analytics uses the available data to provide the best decisions for the business.

IV. Diagnostic Analysis. The diagnostic analysis uses available data to find out why an event or phenomenon occurred.

**Evolution of Data Science and Data Analytics**

Modern organizations have developed the need to extract useful information from data to help them make better decisions, rather than merely collecting the data to fulfill traditional organizational functions like keeping records and processing business transactions. The methods employed for achieving this new data target have evolved over the years. Traditional methods of statistics evolved mainly during periods when acquiring data was very expensive and consumed a lot of time (Viswanathan, 2014). Statistics became crucial in handling the data problem. With time, however, data turned into a default product of several processes in business activities. This means that businesses began accumulating large data sets because the cost of collecting data dropped. This created a problem of handling the enormous volumes of data. Data analytics emerged as a possible solution to this problem. Data analytics pushed beyond the limits of statistical analysis used in the past to cover and analyze larger quantities of data. This data was collected from several sources. The analytics incorporates machine learning and data mining techniques to solve organizational issues.

Data science approaches the big data problem in a multidisciplinary way. Data scientists are required to possess a variety of skills and not be limited to just data analytics. These skills include coding skills to develop computer programs and business skills to evaluate the value of insights generated and ensure that the programs address fundamental business problems.

**Differences between Data Science and Data Analytics.**

Data science shares very many similarities with data analytics. However, this paper will focus on the differences between these two fields as far as academic literature is concerned. The first noticeable difference is that data analytics programs are developed in business programs and data science programs in specialties like computer science. The new programs that emerge seem to borrow from the disciplines' heritage that they emerged from and emulate the faculty's expertise in those disciplines. Courses on data mining in data analytics programs major in the application of data mining techniques. On the other hand, programs of data science are built around the algorithms that the course teaches. Data analytics and modeling courses make a similar observation.

Data analytics programs can utilize case studies in the analytical courses they offer in institutions, while data science programs do not make use of such studies. Data analytics programs emphasize evaluating techniques and tools. Contrary to this, data science programs instead major in implementing these techniques. The science programs also differ from those of data analytics because they involve more mathematics courses. They are taught in linear algebra, Calculus, and discrete mathematics. Moreover, they all require a significant level of programming and statistics courses.

# DATA ANALYTICS PROBLEMS

A vocabulary is introduced to bring out the contrasting characteristics of data science and those of data analytics. The vocabulary is *principles and elements of data analysis*. The vocabulary helps used to describe the fundamental concepts of data science and data analysis.

## Elements of Data Analysis

The analysis comprises individual essential components that the analyst puts together to make up the complete analysis. These components are known as elements of the analysis.

I. Narrative text. These are sentences that explain the data analysis in a way that is simple and easy to understand.

II. Code. These are computer instructions that are fed to a computer to execute a specific programming language.

III. Code comment. This is an explanatory text that accompanies the main code. It explains the main code and is not executed with the program.

IV. Data Visualisation. This is the graphical representation of data using visual elements like graphs and charts.

V. Narrative diagram. This is a diagram that explains data but doesn't contain the data.

VI. Summary Statistics. A simple description of the data. Derived using mean, mode, standard deviation, etc.

VII. Table. Use of rows and columns to arrange data.

VIII. Statistical model. Sample data generated from mathematical models and statistical assumptions.

**Principles Of Data Analysis**

The prioritized features that hold relevance to the data analysis, either in entirety or as single parts, are referred to as the analysis principles. These principles are able to be observed and measured. They include components like:

- Data Matching principle.

Data is ready or available to analysts who match data for investigating a phenomenon with analytic elements of data in analyses with high data matching. If a question exists on quantities that are not at the disposal of the analyst or those that cannot be measured directly, data that is matched to the question at hand may be an alternative to the data phenomena under study. Since an analyst chooses the analytic data elements utilized to solve the question, this is therefore considered a principle of the analysis. Elements used here rely on how well the data are matched. The major question and the study data are taken to be inputs to the analysis. In case the data are not matched properly, the analyst has to find out the major question with one set of analytic data elements. The analyst will also have to employ other elements to explain how well the surrogate data is related to the main data phenomena.

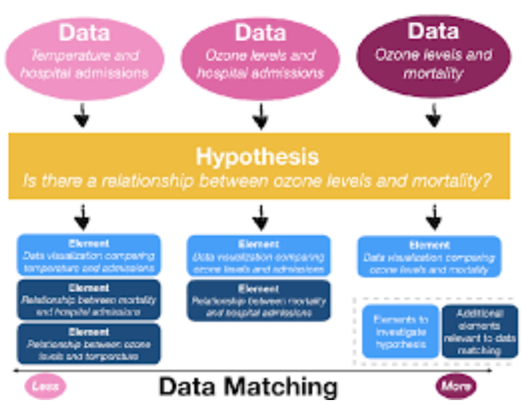


Figure 1. Data matching principle.

- Exhaustive principle.

An exhaustive analysis is one that addresses specific questions and addresses them using multiple, complementary elements. To be exhaustive in how they approach a question, analysts will solve the same question using tools or methods that complement each other. The analysts are aware that each tool will reveal certain information and features of the data but hides others. Consequently, the combination of tools or elements employed might give a complete picture of the evidence in the data than when an individual element is used.

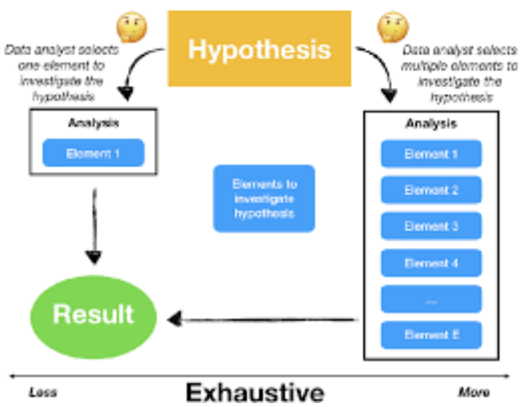


Figure 2. Exhaustive principle.

- Skeptical principle

A skeptical analysis is one that uses the same data to consider multiple and related questions. Analyses will most of the time look at alternative explanations of observed phenomena and evaluate the data's consistency using these explanations. Those that do not use alternate explanations lack skepticism. Contextual information that is outside the data is used to establish whether or not the configuration of alternative explanations is relevant to the problem at hand.

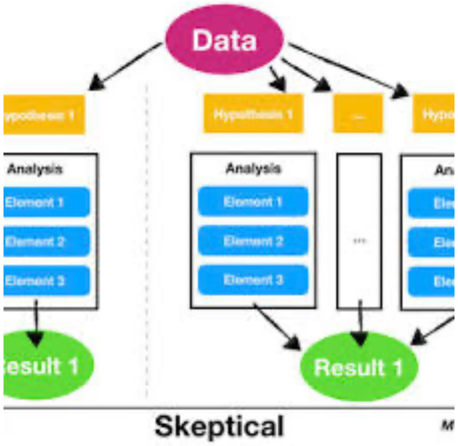


Figure 3. Skeptical principle.

- Second-Order Principle

A second-order analysis is made up of elements that do not address the primary question directly. Instead, the elements provide crucial supporting information to the analysis. Any given analysis will have elements that contribute directly to the results or conclusions of the analysis. The analysis will also have elements that give context or are needed for some other reasons. An example is when the data are less well matched to the investigation of the question. Second-order analyses will always have several of these background or contextual elements in their analysis.

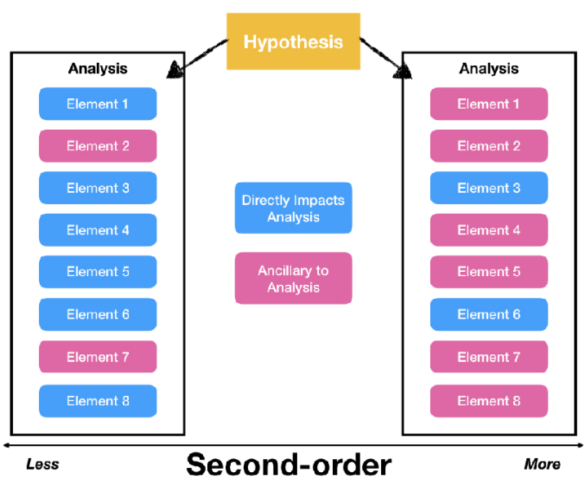


Figure 4. Second-order principle.

- Transparency Principle.

Transparent analyses are those that present an element or subset of elements that summarize or visualize data that influence the explanation of how the underlying data phenomena or data generation process links to any major products, results, or conclusions. While the entirety of analysis may be complicated and involve numerous procedures, transparent analyses identify and extract a single or a few elements from the analysis that summarizes or visualize important bits of evidence in the data that are most influential in understanding the key results. One way of being transparent is showing the approximate mechanism by which the data informs the results or conclusion.

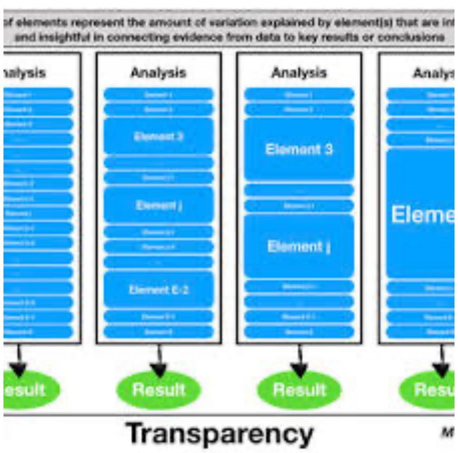


Figure 5. Transparency principle.

- Reproducible Principle.

A reproducible analysis is an analysis in which one who is not the original analyst can take the published code and data and produce the exact outcome just like the original analyst. Provided with similar elements of the data analysis, we can produce the exact same results of the analysis. The availability of the analytic container to other parties who may want to examine the results again is important to the principle of reproducibility. It is difficult for businesses, such as those in the finance industry, to make entire analytic containers available for proprietary or financial reasons. On the contrary, analytic containers that are integrated as part of the analytic product or analytic presentation are more reproducible. Reproducibility, or lack of it, is usually easy to verify, and it does not depend on the nature of the audience viewing the analysis. The reproducibility principle speaks to the coherence of the workflow in the analysis in that the workflow should display how the data are converted to results.

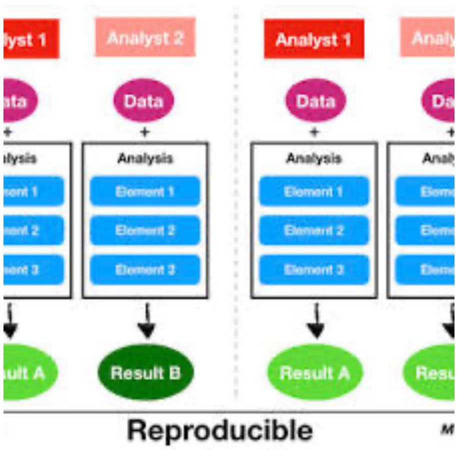


Figure 6. Reproducible principle.

**Data Analysis Problem**

To effectively carry out an analysis, certain mechanisms have to be put in place to ensure that data is processed and accurate results are produced. This might not be the case all the time. An analysis can run into several problems that might delay, affect or even completely stop the process from happening. The following problems arose as possible data analysis problems.

- Poor quality data. When data has errors or is inaccurate, it can have negative impacts on the data analysis process.

- Incomplete data. Certain pieces of data may be unavailable. Either because it was not collected or it got lost somewhere in between, and therefore analysis cannot proceed smoothly.
- Visual representation of data. This occurs when the tools for representing the data in the form of graphs and charts are unavailable or unreliable.
- Inaccessible data. When the data analysts are unable to access all the data, they need to perform an analysis.
- Shortage of skills. Lack of the skills necessary to properly analyze data and produce important insights from it might cause a major problem to the analysis process.
- Budget. When the data analysis incurs expenses that fall outside the budget constraints, or if there lacks a budget entirely.

**Data Analysis Question**

The following questions were arrived at from the elements of analytics and also after analyzing the principles of analytics.

- What is data science?
- What is data analytics?
- What are the major differences between data science and data analytics?
- What is the data industry?
- What are the impacts of data science and data analytics in the data industry

# PRESENTATION OF FINDINGS

## Introduction

The findings of the study were able to establish the distinct differences between data science and data analytics. Although the study established the differences between data science and data analytics, it also found the two to share several functions that make the two work together in several ways as part of a big system. Knowing the differences that exist between the two is, however, very important for any business that seeks to operate on optimum efficiency. At the same time, knowing what data science and data analytics offer at an individual level is generous so as to understand and appreciate their collective input to the overall organizational goals.

## What is Data Science?

Data science is defined as the study and practice of getting information and meaningful insights from large volumes of data using different methods. Data scientists are responsible for attempting to understand the objectives set by the stakeholders of a business and then figuring out how to use data to meet these objectives. From the findings, data science was found to have four components. These are:

- Statistics. This is the use of mathematical methods to collect data, analyze the data, interpret the data and then present the data
- Data visualization. This is the presentation of analysis results in graphical forms that are easy to understand. These include graphs, charts, and diagrams. They provide a highlight on the key points and thereby ease the interpretation.

- Machine Learning. Machine learning is the use of computer intelligence through algorithms that learn trends and patterns over time to be able to make future forecasts and predict human behavior.

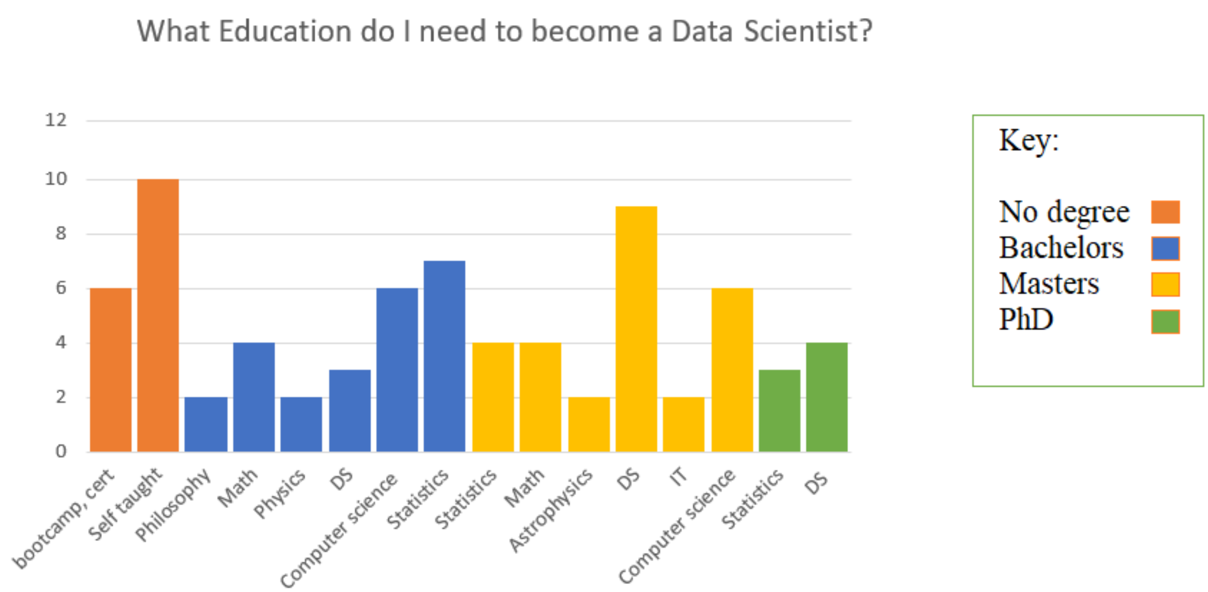


Figure 7. Education levels are needed to become a data scientist.

**What is Data Analytics?**

Data analytics is the study and also the practice of analyzing raw data that is collected by businesses and converting it into forms that can benefit the business. There are four major steps involved in data analytics. The first step involves identifying the requirements of the analysis. This is arrived at based on the business problem or the group that is targeted. The group can be segmented into different demographics. The second step is data collection. Here, actual data collection is done. The data is collected from the relevant sources, which could include internet sources or physical customer surveys. After that, the data is organized in a presentable way for analysis. Data cleaning is then done to get rid of duplicate data, incomplete records, or any other unnecessary data. This stage prepares the data for the actual analysis.

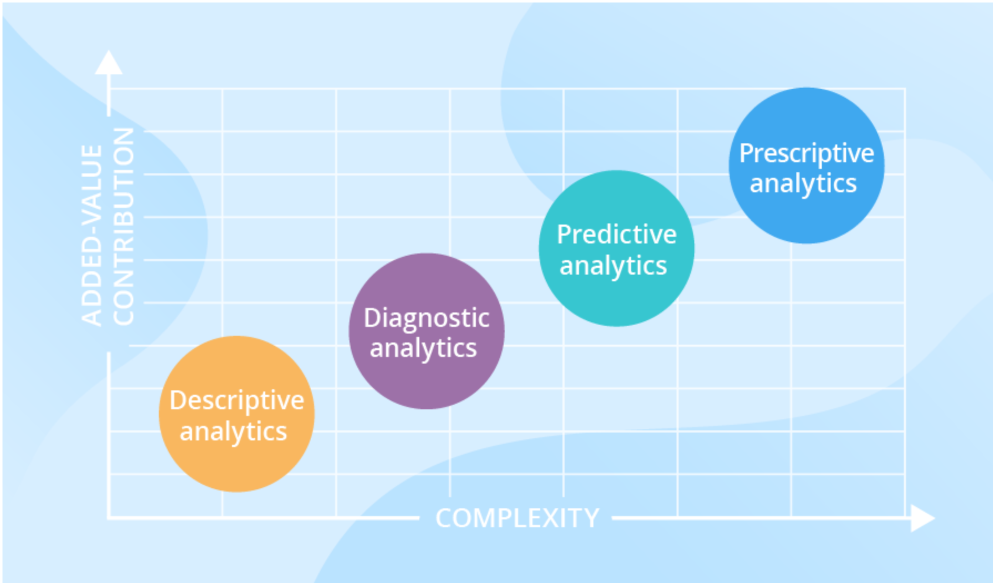


Figure 8. Types of Data Analytics

Table 1. The contrasting features of data analysis and a data science

| | **Data Analyst** | **Data Scientist** |
|---|---|---|
| Education | Four-year degree in mathematics, statistics with a focus on analytics. | Masters or doctorate in Mathematics, statistics, or computer science. |
| Experience and skills | Background in mathematics and statistics with a solid understanding of data mining techniques.<br>Strong written and verbal communication skills, as well as familiarity with data mining, data modeling, database management, statistical analysis, and database analysis. | Combination of mathematical and statistical knowledge, programming expertise, as well as analytical skills, including familiarity with machine learning, software development, data mining, python, and data analysis. |

| | | |
|---|---|---|
| Expectations | Designing and maintaining data systems and databases using statistical tools to interpret data, prepare reports that can communicate trends, patterns, and predictions based on data, as well as to conduct consumer data research and analytics.<br>Work with customer-centric algorithm models and be able to shape them to each customer as required. Extract insights from vast databases, help convert data into visualizations, metrics, and goals and have a general proactive approach. | Data scientists require the creation of algorithms and predictive models that extract the needed information to solve complex problems. Focus on designing and constructing new processes for data and modeling production.<br>Able to use predictive modeling to increase and optimize customer experiences, revenue generation, and targeting. Develop data models and algorithms; develop processes and tools to monitor and analyze the performance of models.<br>Ask unique questions and predict future trends. |

The demand for a skilled workforce in data science skills has increased in recent years, resulting in a re-evaluation of the practice and data analysis teaching. Data analysts can define their work with a formal set of elements and principles for describing data analyses without being limited to the confines of their specific application or area of study. Further exploration of data analysis and formalization of processes around it becomes possible through the existence of these elements and principles (Nadikattu. 2020). The development of elements and principles is vital in creating the foundation for a clear identity for the field of data science. This, in turn, helps develop an effective mechanism for teaching data science.

**New data economy and industry transformation.**

The recognition of data science and data analytics's potentialities and its rapid growth has been driven by the evolution of a new data economy and transformations in the industry. An example is large private data enterprises. Advancement in data science, big data, and data analytics significantly influence and drives the development of a new data economy. This wave of industry transformation brings with it the revolution of advanced artificial intelligence-enabled technologies and businesses. The growth of AI and the AI-driven data economy is largely driven by data science and analytics. Large private enterprises like Facebook, Google, and Alibaba lead this new economy. This new economy's main product is data and is technology-driven (Cao, L., 2017).

Many more organizations are recognizing the value of data as an asset. These businesses then invest in building infrastructure, create resources, pool talent and teams to support enterprise innovation. Today, Internet-based and data-driven businesses like Google, Facebook, and Alibaba, have overtaken traditional enterprise giants.

**Data Industry**

Data has been looked at from the same perspective as oil was in the past. It has been viewed as the new international currency. Data industrialization has created a new business. Organizations and countries have competed on how to come up with new data products. Business in the retail and manufacturing industry is rapidly helping create a new economy around the data industry and a digital economy. The domination can see this change of data-driven firms emerging top in the best companies worldwide. For example, the biggest data company was Google Inc., while the largest IPO, Alibaba (Cao, L., 2017). The industry is rapidly shaping up and becoming recognized as a major force in the emerging global economy.

The main forces that drive the data industry can be found in the areas listed below:

- Data design and analytics design
- Data content and analytics content
- Data software and analytics software
- Data infrastructure and analytics infrastructure
- Data services and analytics services
- Data education and analytics education.

## LIMITATIONS OF THE ANALYSIS

This paper analyzed the differences in features of data analytics and the features of data science. However, the analysis faced several limitations, and given more time, these would have been overcome and improvements made on the analysis.

The elements and principles discussed in this paper were insufficient to provide an exhaustive analysis of the differences between the two fields of science and analytics. Data science is broad, and so is data analytics. These two fields continue evolving with each passing day. New concerns arise that either create new similarities or widen the existing differences between the two topics. Big data has brought its own set of issues that businesses have to handle. The more data businesses collect, the faster they encounter new issues that affect the methods they use to collect data, data storage, and data analysis. The consequence of this rapid evolution means that research has to be done and continuously updated for it to stay abreast with the current trends and latest developments in the industry and avoid losing relevance. The curricula that offer courses on data science and data analytics will also have to be updated more often to equip learners with skills relevant to the market.

Apart from that, the analysis was limited in its scope and could not look into data science and data analytics programs' effectiveness. Given time, I would do further research to go beyond the mere comparison of differences between the two areas. I would do a study to establish and evaluate the effectiveness of the programs. Are there enough institutions offering data science programs? Do these institutions produce an adequate to fill the enormous global demand for data scientists? It would also evaluate the availability of opportunities for data science and data analytics graduates in the job market. This is an essential area because the technological world and big data are evolving very rapidly. Academic programs will have to factor in this rate of change and evolving industry expectations to effectively ready graduates to enter the job market and stay relevant to adapt to future changes. An in-depth analysis of data science and data analytics curricula by analyzing course syllabi would provide insights into courses covered and provide potential future research topics.

Also, the study did not explore the area of ethics in data science and data analytics. There exists a need to carry out a comprehensive study that will look into the methods businesses use to collect consumer data and the policies they put in place to secure this data. Big data has necessitated the need to create awareness of data privacy and data protection. This study will also see the general consumer data protection laws that currently exist and evaluate their effectiveness. More than once, data has been described as the 'oil of the $21^{st}$ century.' Data mining has in the past been used by businesses for dishonest gains at the mercy of consumers. Businesses have realized the value of data in the $21^{st}$ century and are very likely to use it for reasons other than organizational decision making or product development.

**REFERENCE LIST**

Aasheim, C.L., Williams, S., Rutner, P. and Gardiner, A., 2015. Data analytics vs. data science: A study of similarities and differences in undergraduate programs based on course descriptions. *Journal of Information Systems Education*, *26*(2), p.103.

Cao, L., 2017. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), pp.1-42.

Nadikattu, R.R., 2020. Research On Data Science, Data Analytics And Big Data. *INTERNATIONAL JOURNAL OF ENGINEERING, SCIENCE AND*, *9*(5), pp.99-105.

Viswanathan, V. (2014). Data Analytics with R: A Hands-on Approach (1e). Infivista Inc