VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANASANGAMA, BELGAVI-590018

SE ASSIGNMENT REPORT ON
"DATA SCIENCE"

Submitted in partial fulfillment of the requirements for the 3rd Semester

INFORMATION SCIENCE AND ENGINEERING

Submitted by

NITESH NAG	1BI20IS059
POOJA HARIHAR	1BI20IS061
PRANAV SUDHAKAR	1BI20IS062
PRASANNAGOUDA S PATIL	1BI20IS063

Under the guidance of

Dr. Hema Jagadish
Associate Professor
Dept of ISE, BIT, Bangalore-04

2022

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING
BANGALORE INSTITUTE OF TECHNOLOGY
K. R. Road, V. V. Puram, Bengaluru-560004

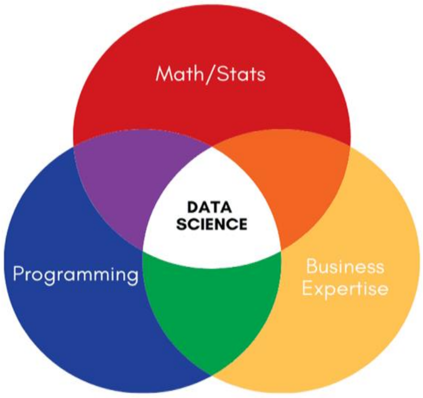
1. Abstract:

Data science is a set of principles, problem definitions, algorithms, and processes for extracting useful patterns from large data sets. Many of the elements of data science have been developed in related fields such as machine learning and data mining. In fact, the terms data science, machine learning, and data mining are often used interchangeably. The commonality across these disciplines is a focus on improving decision making through the analysis of data. However, although data science borrows from these other fields, it is broader in scope. Machine learning (ML) focuses on the design and evaluation of algorithms for extracting patterns from data. Data mining generally deals with the analysis of structured data and often implies an emphasis on commercial applications. Data science takes all of these considerations into account but also takes up other challenges, such as the capturing, cleaning, and transforming of unstructured social media and web data; the use of big-data technologies to store and process big, unstructured data sets; and questions related to data ethics and regulation.

2. Introduction:

The key objective of Data Science is to extract valuable information for use in strategic decision making, product development, trend analysis. Data Science has become one of the most demanded jobs of the 21st century. In this presentation, we will demystify Data Science, the role of a Data Scientist, Data Engineer and Data analyst and have a look at the tools required to master Data Science.

Daniel Keys Moran as quoted that "*You can have data without information, but you cannot have information without data*".



Data Science is about data gathering, analysis and decision-making. Data Science is about finding patterns in data, through analysis, and make future predictions.

While the buzzword of Data Science has been circulating for a while, very few people know about the real purpose of being a Data Science.

SO, what exactly is data science?

It's the domain of study that deals with vast volumes of data using modern tools and techniques to find unseen patterns, derive meaningful information, and make business decisions.

It actually makes use of machine learning algorithms and then is used in modelling purposes. So, we can say that data science is about extraction, preparation, analysis, visualization, and maintenance of information." With the emergence of new technologies, there has been an exponential increase in data. This has created an opportunity to analyze and derive meaningful insights from data.

now the question arises why is it important?

Data creates magic. Industries need data to help them make careful decisions. Data Science churns raw data into meaningful insights. Therefore, industries need data science. A Data Scientist is a wizard who knows how to create magic using data.

3. Pre-requisites:

We can understand Data Science as a field that deals with data processing, analysis, and extraction of insights from the data using various statistical methods and computer algorithms. It is a multidisciplinary field that combines mathematics, statistics, and computer science.

MACHINE LEARNING

Machine learning is the backbone of data science. Machine learning and data science can work hand in hand. It basically automates the process of Data and makes data-informed predictions in real-time without any human intervention. Data Scientists need to have a solid grasp of ML in addition to basic knowledge of statistics

STATISTICS

Models enable to make quick decisions. It is a part of machine learning and involves identifying the best algorithms. Statistics for Data Science is essential because these disciples form the basic foundation of all the Machine Learning Algorithms.

PROGRAMMING

R is a scripting language that is specifically tailored for statistical computing. R is mostly used for statistical operations. ***Python*** is an interpreter based high-level programming language. It is mostly used for Data Science and Software Development.

MODELLING

It's used for better understanding of data and do the analysis. It is a part of machine learning and involves identifying the best algorithm. Helps us by Identifying patterns, trends, and anomalies. Identifying new data sources and know the value of data and how to utilize it. Build a business model using data rules to optimize targets.

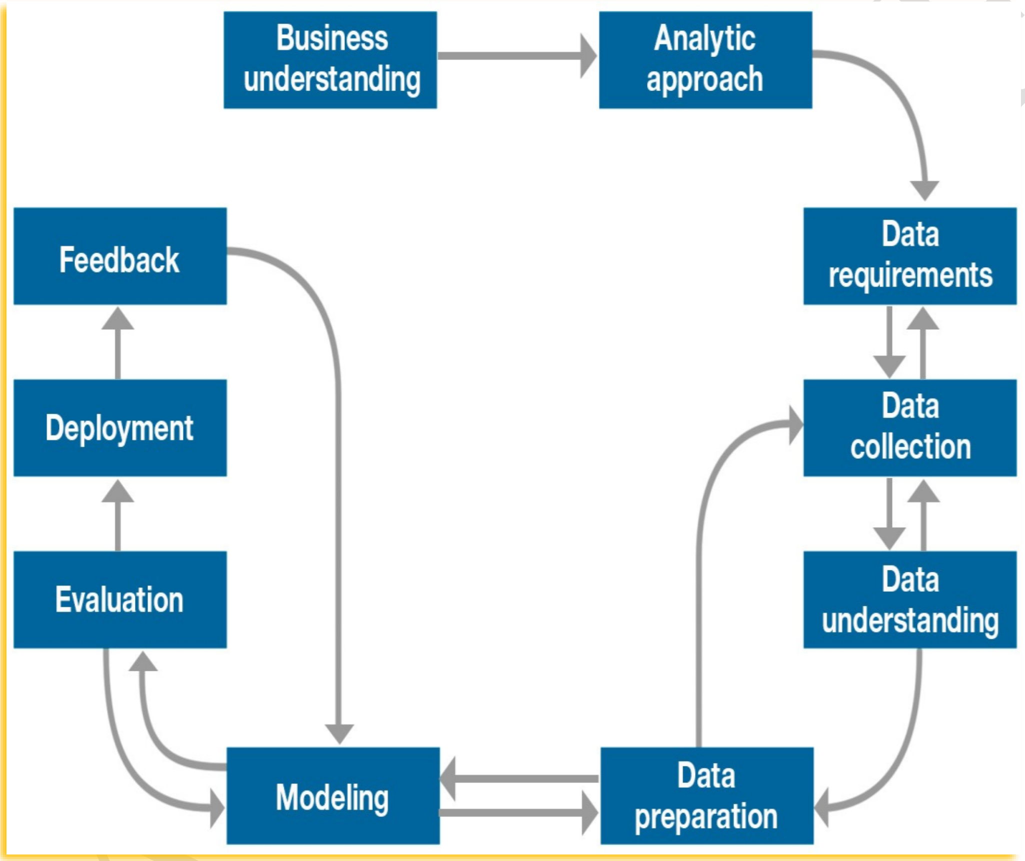
4. Life Cycle Process

- **Capture**: It includes processes like Data Acquisition, Data Entry, Signal Reception, Data Extraction. This stage involves gathering raw structured and unstructured data all together.
- **Maintain**: Here maintain involves, Data Warehousing, Data Cleansing, Data Staging, Data Processing, Data Architecture. This stage covers taking the raw data and putting it in a form that can be used.
- **Process**: This step includes Data Mining, Clustering/Classification, Data Modeling, Data Summarization. Data scientists take the prepared data and examine its patterns, ranges, and biases to determine how useful it will be in predictive analysis.
- **Analyze**: Exploratory/Confirmatory, Predictive Analysis, Regression, Text Mining, Qualitative Analysis. Very important stage. This stage involves performing the various analyses on the data.
- **Communicate**: It mainly involves with Data Reporting, Data Visualization, Business Intelligence, Decision Making. In this final step, analysts prepare the analyses in easily readable forms such as charts, graphs, and reports. Any changes required can be done as per requirements.

Now the question arises, what does a data scientist do?
A data scientist analyze business data to extract meaningful insights. In other words,

a data scientist solves business problems through a series of steps, including:

- Before tackling the data collection and analysis, the data scientist determines the problem by asking the right questions and gaining understanding.
- The data scientist then determines the correct set of variables and data sets.
- The data scientist gathers structured and unstructured data from many disparate sources—enterprise data, public data, etc.



- Once the data is collected, the data scientist processes the raw data and converts it into a format suitable for analysis. This involves cleaning and validating the data to guarantee uniformity, completeness, and accuracy.
- After the data has been rendered into a usable form, it's fed into the analytic system—ML algorithm or a statistical model. This is where the data scientists analyze and identify patterns and trends.
- When the data has been completely rendered, the data scientist interprets the data to find opportunities and solutions.
- The data scientists finish the task by preparing the results and insights to share with the appropriate stakeholders and communicating the results.

4. Specializations

1)DATA SCIENCTIST

- **Job role:** Determine what the problem is, what questions need answers, and where to find the data. Also, they mine, clean, and present the relevant data.
- **Skills needed:** Programming skills (SAS, R, Python), storytelling and data visualization, statistical and mathematical skills, knowledge of Hadoop, SQL, and Machine Learning.

2)DATA ANALYST

- **Job role:** Analysts bridge the gap between the data scientists and the business analysts, organizing and analyzing data to answer the questions the organization poses.
- **Skills needed:** Statistical and mathematical skills, programming skills (SAS, R, Python), plus experience in data wrangling and data visualization.

3)DATA ENGINEER

- **Job role:** Data engineers focus on developing, deploying, managing, and optimizing the organization’s data infrastructure and data pipelines. Engineers support data scientists by helping to transfer and transform data for queries.
- **Skills needed:** NoSQL databases (e.g., MongoDB, Cassandra DB), programming languages such as Java and Scala, and frameworks (Apache Hadoop).

Basic requirements:

- Mathematical Expertise
- Technological skills
- Project management
- Communication
- Basic programming languages
- Business Acumen

4. Applications

Data science has found its applications in almost every industry. Let’s look into them briefly

1. **Healthcare**

Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases.

2. **Gaming**

Video and computer games are now being created with the help of data science and that has taken the gaming experience to the next level.

3. **Image Recognition**

Identifying patterns in images and detecting objects in an image is one of the most popular data science applications.

4. **Recommendation Systems**

Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase, or browse on their platforms.

5. **Logistics**

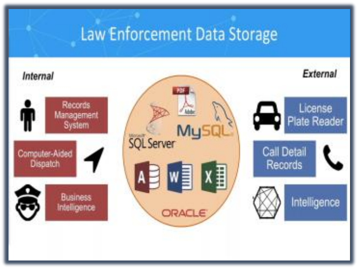
Data Science is used by logistics companies to optimize routes to ensure faster delivery of products and increase operational efficiency.

6. **Fraud Detection**

Banking and financial institutions use data science and related algorithms to detect fraudulent transactions.

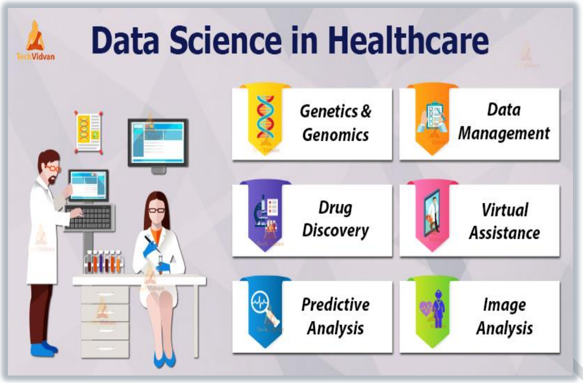
4. Use-cases

1) LAW ENFORCEMENT



In this scenario, data science is used to help police in Belgium to better understand where and when to deploy personnel to prevent crime. With only limited resources and a large area to cover data science used dashboards and reports to increase the officers' situational awareness, allowing a police force that's spread thin to maintain order and anticipate criminal activity.

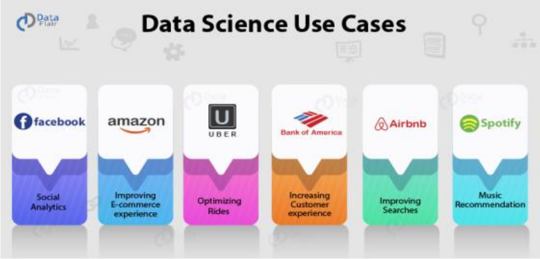
2) **MEDICAL INDUSTRY**



It is one of those fields in healthcare that helps to figure out better treatment strategies. IBM estimates that the medical images contain around 90% of the overall medical data. Doctors use the medical imaging technique to effectively visualize the interior parts of the body.

Also, to analyze the function of some of the organs to diagnose and treat any disorder or disease. The insights gained from these images can make a difference in the patient's treatment.

- **Top companies using data science:**



- Amazon
- Facebook
- Netflix
- Uber
- Spotify

1)META (known as Facebook)

Using deep learning, Facebook makes use of facial recognition and text analysis. In facial recognition, Facebook uses powerful neural networks to classify faces in the photographs. It uses its own text understanding engine called “Deep-Text” to understand user sentences. It also uses Deep Text to understand people’s interest and aligning photographs with texts.

The new name reflects the company's growing ambitions beyond social media. Facebook, now known as Meta, has adopted the new moniker, based on the sci-fi term metaverse, to describe its vision for working and playing in a virtual world.

2) AMAZON

Here, Amazon heavily relies on predictive analysis to increase customer satisfaction. It does so through a personalized recommendation system.

Amazon has an anticipatory shipping model that uses big data for predicting the products that are most likely to be purchased by its users. It analyzes the pattern of your purchases and sends products to your nearest warehouse which you may utilize in the future.

3)UBER

In this scenario, Uber contains a database of drivers. Therefore, whenever you hail for a cab, Uber matches your profile with the most suitable driver. What differentiates Uber from other cab companies is that Uber charges you based on the time it takes to cover the distance and not the distance itself.

It calculates the time taken through various algorithms that also make use of data related to traffic density and weather conditions.

5)Advantages of Data Science

The various benefits of Data Science are as follows:

1. It's in Demand

Data Science is greatly in demand. Prospective job seekers have numerous opportunities. It is the fastest growing job on LinkedIn and is predicted to create 11.5 million jobs by 2026. This makes Data Science a highly employable job sector.

2. Abundance of Positions

There are very few people who have the required skill-set to become a complete Data Scientist. This makes Data Science less saturated as compared with other IT sectors.

Therefore, Data Science is a vastly abundant field and has a lot of opportunities. The field of Data Science is high in demand but low in supply of Data Scientists.

3. A Highly Paid Career

Data Science is one of the most highly paid jobs. According to Glassdoor, Data Scientists make an average of $116,100 per year. This makes Data Science a highly lucrative career option.

4. Data Science is Versatile

There are numerous applications of Data Science. It is widely used in health-care, banking, consultancy services, and e-commerce industries. Data Science is a very versatile field. Therefore, you will have the opportunity to work in various fields.

5. Data Science Makes Data Better

Companies require skilled Data Scientists to process and analyze their data. They not only analyze the data but also improve its quality. Therefore, Data Science deals with enriching data and making it better for their company.

6. Data Scientists are Highly Prestigious

Data Scientists allow companies to make smarter business decisions. Companies rely on Data Scientists and use their expertise to provide better results to their clients. This gives Data Scientists an important position in the company.

7. No More Boring Tasks

Data Science has helped various industries to automate redundant tasks. Companies are using historical data to train machines in order to perform repetitive tasks. This has simplified the arduous jobs undertaken by humans before.

6)Disdvantages of Data Science

1. Mastering Data Science is near to impossible

Being a mixture of many fields, Data Science stems from Statistics, Computer Science and Mathematics. It is far from possible to master each field and be equivalently expert in all of them.

While many *online courses* have been trying to fill the skill-gap that the data science industry is facing, it is still not possible to be proficient at it considering the immensity of the field.

A person with a background in Statistics may not be able to master Computer Science on short notice in order to become a proficient Data Scientist. Therefore, it is an ever-changing, dynamic field that requires the person to keep learning the various avenues of Data Science.

2. Large Amount of Domain Knowledge Required

Another disadvantage of Data Science is its dependency on Domain Knowledge. A person with a considerable background in Statistics and Computer Science will find it difficult to solve Data Science problem without its background knowledge.
The same holds true for its vice-versa. For example, A health-care industry working on an analysis of genomic sequences will require a suitable employee with some knowledge of genetics and molecular biology.

3. Arbitrary Data May Yield Unexpected Results

A Data Scientist analyzes the data and makes careful predictions in order to facilitate the decision-making process. Many times, the data provided is arbitrary and does not yield expected results. This can also fail due to weak management and poor utilization of resources.

4. Problem of Data Privacy

For many industries, data is their fuel. Data Scientists help companies make data-driven decisions. However, the data utilized in the process may breach the privacy of customers.

The personal data of clients are visible to the parent company and may at times cause data leaks due to lapse in security. The ethical issues regarding preservation of data-privacy and its usage have been a concern for many industries.

7)Summary

After weighing the pros and cons of Data Science we are able to envision the full picture of this field. While **Data Science is a field with many lucrative advantages**, it also suffers from its disadvantages.
Being a less-saturated, high paying field that has revolutionized several walks of life, it also has its own backdrops when considering the immensity of the field and its cross-disciplinary nature.

Data Science is an ever-evolving field that will take years to gain proficiency. In the end, it is up to you to decide whether the pros of Data Science motivate you to take this up as your future career or the cons that help you take a careful decision!