Sanjivani Rural Education Society's

Sanjivani College of Engineering, Kopargaon-423603

(An Autonomous Institute Affiliated to Savitribai Phule Pune University, Pune)

(NAAC 'A' Grade Accredited)

Department of Information Technology

Foundation of Data Science (IT305B)

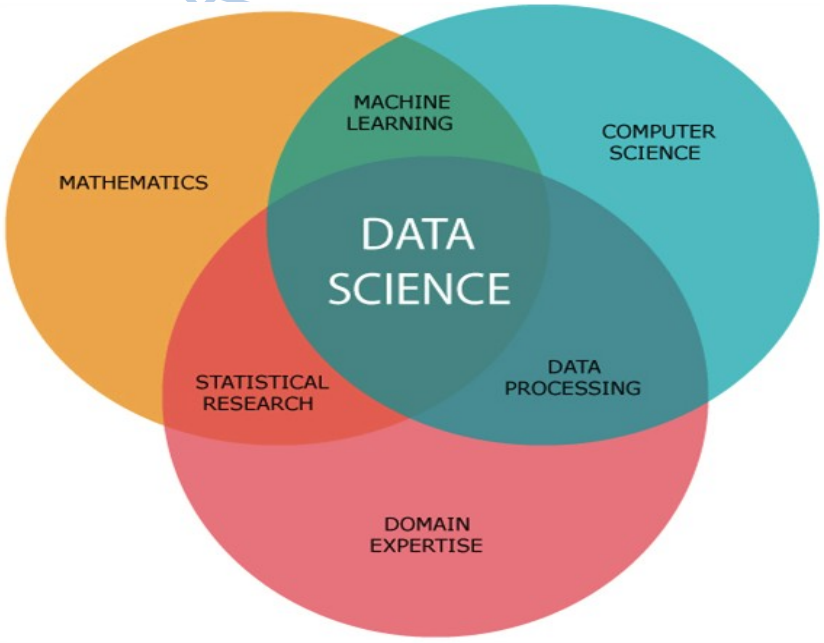
Unit-I

INTRODUCTION TO DATA SCIENCE

Data Science Definition:

" *Data science is the field of applying advanced analytics techniques and scientific principles to extract valuable information from data for business decision-making,strategic planning and other uses*"

"Data science is the study of data. It involves developing methods of recording,storing, and analyzing data to effectively extract useful information. The goal of data science is to gain insights and knowledge from any type of data i.e. both structured and unstructured"



- ***Benefits and Uses of Data Science:***

- ***Uses of Data Science:***

1) In the healthcare industry, physicians use Data Science to analyze data from wearable trackers to ensure their patients' well-being and make vital decisions. Data Science also enables hospital managers to reduce waiting time and enhance care.
2) Retailers use Data Science to enhance customer experience and retention.
3) Data Science is widely used in the banking and finance sectors for fraud detection and personalized financial advice.
4) Transportation providers use Data Science to enhance the transportation journeys of their customers. For instance, Transport for London maps customer journeys offering personalized transportation details, and manages unexpected circumstances using statistical data.
5) Construction companies use Data Science for better decision making by tracking activities, including average time for completing tasks, materials-based expenses, and more.
6) Data Science enables trapping and analyzing massive data from manufacturing processes, which has gone untapped so far.
7) With Data Science, one can analyze massive graphical data, temporal data, and geospatial data to draw insights. It also helps in seismic interpretation and reservoir characterization.
8) Data Science facilitates firms to leverage social media content to obtain real-time media content usage patterns. This enables the firms to create target audience-specific content, measure content performance, and recommend on-demand content.
9) Data Science helps study utility consumption in the energy and utility domain. This study allows for better control of utility use and enhanced consumer feedback.
10) Data Science applications in the public service field include health-related research, financial market analysis, fraud detection, energy exploration, environmental protection, and more.

- ***Benefits of Data Science:***

***1) Increases business predictability*:**
When a company invests in structuring its data, it can work with what we call predictive analysis. With the help of the data scientist, it is possible to use technologies such as Machine Learning and Artificial Intelligence to work with the

data that the company has and, in this way, carry out more precise analyses of what is to come. Thus, you increase the predictability of the business and can make decisions today that will positively impact the future of your business.

***2) Ensures real-time intelligence*:**
The data scientist can work with RPA(Robotic Process Automation) professionals to identify the different data sources of their business and create automated dashboards, which search all this data in real-time in an integrated manner.
This intelligence is essential for the managers of your company to make more accurate and faster decisions.

***3) Favors the marketing and sales area*:**
Data scientists can integrate data from different sources, bringing even more accurate insights to their team. Can you imagine obtaining the entire customer journey map considering all the touch points your customer had with your brand? This is possible with Data Science.

***4) Improves data security*:**
One of the benefits of Data Science is the work done in the area of data security. In that sense, there is a world of possibilities.
The data scientists work on fraud prevention systems, for example, to keep your company's customers safer. On the other hand, he can also study recurring patterns of behaviour in a company's systems to identify possible architectural flaws.

***5) Helps interpret complex data*:**
Data Science is a great solution when we want to cross different data to understand the business and the market better. Depending on the tools we use to collect data, we can mix data from "physical" and virtual sources for better visualization.

***6) Facilitates the decision-making process*:**
One of the benefits of Data Science is improving the decision-making process. This is because we can create tools to view data in real-time, allowing more agility for business managers. This is done both by dashboards and by the projections that are possible with the data scientist's treatment of data.

- ***Data Science and Big Data:***

"Data is the new science. Big data holds the answers."

- Big data includes:

Structured data: transaction data, OLTP, RDBMS, and other structured formats
Semi-Structured: text files, system log files, XML files, etc.
Unstructured data: web pages, sensor data, mobile data, online data, sources, digital audio, and video feeds, digital images, tweets, blogs, emails, social networks, and other sources etc.

Meaning:

Big Data: Large volumes of data that can't be handled using a normal database program. Characterized by velocity, volume, and variety.

Data Science: Data focused scientific activity.Similar in nature to data mining. Harnesses the potential of big data to support business decisions. Includes approaches to process big data.

Concept:

Big Data: Includes all formats and types of data. Diverse data types are generated from several different sources.

Data Science: Helps organizations make decisions. Provides techniques to help extract insights and information to create large datasets. A specialized approach that involves scientific programming tools, techniques, and models to process big data.

Basis of Formation:

Big Data: Data is generated from system logs. Data is created in organizations – emails, spreadsheets, DB, transactions, Online discussion forums. Video and audio streams that include live feeds. Electronic devices – RFID, sensors, and so on. Internet traffic and users.

Data Science: Working apps are made by programming developed models. It captures complex patterns from big data and developed models. It is related to data analysis, preparation, and filtering. Applies scientific methods to find the knowledge in big data.

Approach:

Big Data: To understand the market and to gain new customers. To find sustainability. To establish realistic ROI and metrics. To leverage datasets for the advantage of the business. To gain competitiveness. To develop business agility.

Data Science: Data Visualization and prediction. Data destroy, preserve, publishing, processing, preparation, or acquisition. Programming skills, like NoSQL, SQL, and Hadoop platforms. State-of-the-art algorithms and techniques for data mining. Involves the extensive use of statistics, mathematics,and other tools.

Application Areas:

Big Data: Security and law enforcement.
Research and development.
Commerce, Sports and health.
Performance optimization
Optimizing business processes.

Telecommunications and Financial services.

Data Science: Web development.
Fraud and risk detection.
Image and speech recognition.
Search recommenders etc.

Different types of Data (in Data science process):

In data science and big data we will come across many different types of data, and each of them tends to require different tools and techniques.

The main categories of data are these:

1) Structured
2) Unstructured
3) Natural language
4) Machine-generated
5) Audio, video, and images
6) Streaming

1) Structured Data:

Structured data is the data which conforms to a data model, has a well define structure, follows a consistent order and can be easily accessed and used by a person or a computer program.

Structured data is usually stored in well-defined schemas such as Databases. It is generally tabular with column and rows that clearly define its attributes.

SQL (Structured Query language) is often used to manage structured data stored in databases.

Example: Examples of structured data include **names, dates, addresses, credit card numbers, stock information, geolocation**, and more

2) Unstructured Data:

Unstructured data is the data which does not conforms to a data model and has no easily identifiable structure such that it can not be used by a computer program easily. Unstructured data is not organised in a pre-defined manner or does not have a pre-defined data model, thus it is not a good fit for a mainstream relational database.

Example: Videos, texts, images, document files, audio materials, email contents.

3) Natural Language Data:

Natural language processing (also known as Computational linguistics) is the scientific study of language from a computational perspective, with a focus on the interactions between natural (human) languages and computers.

Example: Natural Language Processing, or NLP for short, is broadly defined as the automatic manipulation of natural language, like speech and text, by software.

Example: personal voice assistants like Siri, Cortana, Alexa, etc.

4) Machine Generated Data(MGD):

Machine-generated data (MGD) is information that is produced by mechanical or digital devices.

The term is often used to describe the data that is generated by an organization's industrial control systems as well as mechanical devices that are designed to carry out a single function.

Example: Web server logs, Call detail records, Financial instrument trades, Network event logs, Security information and event management (SIEM) logs, Telemetry collected by the government.

5) Audio Video and Images Data:

Audio, image, and video are data types that pose specific challenges to a data scientist. Audio and video are used for enhancing the experience with Web pages (e.g. audio background) to serving music, family videos, presentations, etc.

The Web content accessibility guidelines recommend to always provide alternatives for time-based media, such as captions, descriptions, or sign language.

6) Streaming Data:

Streaming data is data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously, and in small sizes (order of Kilobytes).

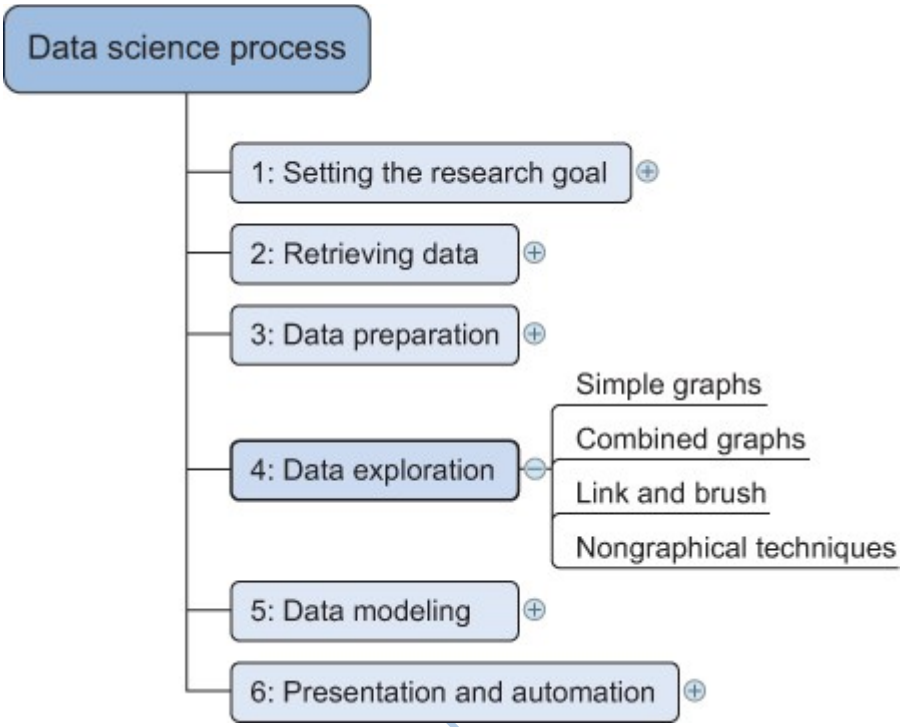
Streaming data includes a wide variety of data such as log files generated by customers using your mobile or web applications, e-commerce purchases, in-game player activity, information from social networks, financial trading floors, or geospatial services, and telemetry from connected devices or instrumentation in data centres.

Example: Location data.
Fraud detection.
Real-time stock trades.
Marketing, sales, and business analytics.
Customer/user activity.
Monitoring and reporting on internal IT systems.
Log Monitoring: Troubleshooting systems, servers, devices, and more.

<u>Data Science Process:</u>

The data science process typically consists of six steps:

1) Setting the research goal
2) Retrieving data
3) Data preparation
4) Data exploration
5) Data modelling or model building
6) Presentation and automation



1) Setting the research goal:

Data science is mostly applied in the con- text of an organization.

When the business asks you to perform a data science project, you'll first prepare a project charter.

This charter contains information such as what you're going to research, how the company benefits from that, what data and resources you need, a timetable, and deliverables.

2) Retrieving data:

The second step is to collect data.

You've stated in the project charter which data you need and where you can find it.

In this step you ensure that you can use the data in your program, which means checking the existence of, quality, and access to the data.

Data can also be delivered by third-party companies and takes many forms ranging from Excel spreadsheets to different types of databases.

3) Data preparation:

Data collection is an error-prone process; in this phase you enhance the quality of the data and prepare it for use in subsequent steps.

This phase consists of three sub-phases:

data cleansing removes false values from a data source and inconsistencies across data sources,

data integration enriches data sources by combining information from multiple data sources,
data transformation ensures that the data is in a suit- able format for use in your models.

4) Data exploration:

Data exploration is concerned with building a deeper understanding of your data.
You try to understand how variables interact with each other, the distribution of the data, and whether there are outliers.
To achieve this you mainly use descriptive statistics, visual techniques, and simple modelling.
This step often goes by the abbreviation EDA, for Exploratory Data Analysis.

5) Data modeling or model building:

In this phase you use models, domain knowledge, and insights about the data you found in the previous steps to answer the research question.
You select a technique from the fields of statistics, machine learning, operations research, and so on.
Building a model is an iterative process that involves selecting the variables for the model, executing the model, and model diagnostics.

6) Presentation and automation:

Finally, you present the results to your business.
These results can take many forms, ranging from presentations to research reports.
Sometimes you'll need to automate the execution of the process because the business will want to use the insights you gained in another project or enable an operational process to use the outcome from your model.