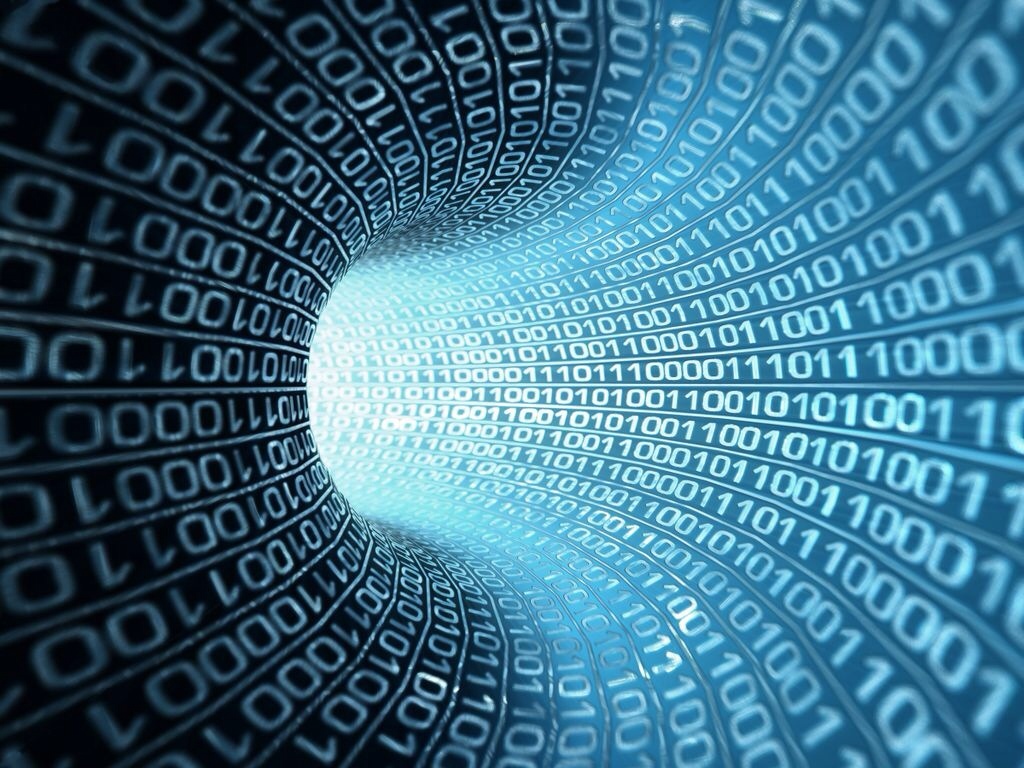


***Module 1***

Statistical Fundamental-I

# Introduction to Data and Database:

1. **What is Data**?

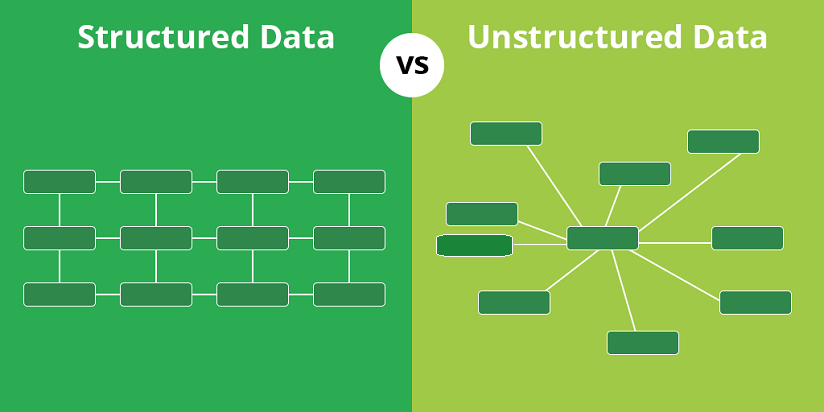
Data is information that can be processed by a computer or other electronic system. It can be in the form of numbers, text, images, audio, or any other form of digital information.

1. **Types of data**
2. **Structured Data**- Data that is organized in a well-defined format, such as a spreadsheet or database.

In Excel (arranged in Rows& Columns), Easy to locate

1. **Unstructured Data**- Data that is not organized in a well-defined format, such as emails or social media posts.

Difficult to locate, audio video, mail, Images



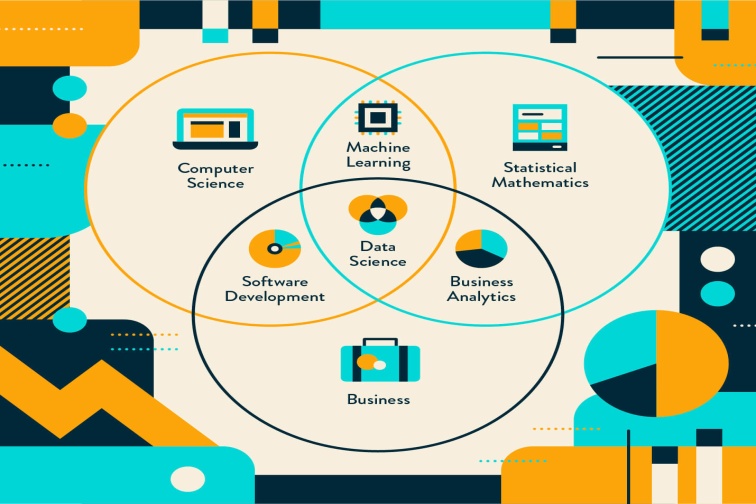
1. **Formats in which we can store data**
2. Text files
3. Spreadsheets
4. CSV (comma-separated values) files
5. JSON (JavaScript Object Notation) files
6. XML (eXtensible Markup Language) files
7. Relational databases
8. **What is database?**

Data can be stored in a database, which is a structured collection of data that is organized for easy access and manipulation. Databases can be used to store a wide variety of data, including customer information, product inventories, financial records, and more. There are many different types of databases, including relational databases, NoSQL databases, and more, each with its specific features and capabilities.



# Introduction to Data science and Analytics:

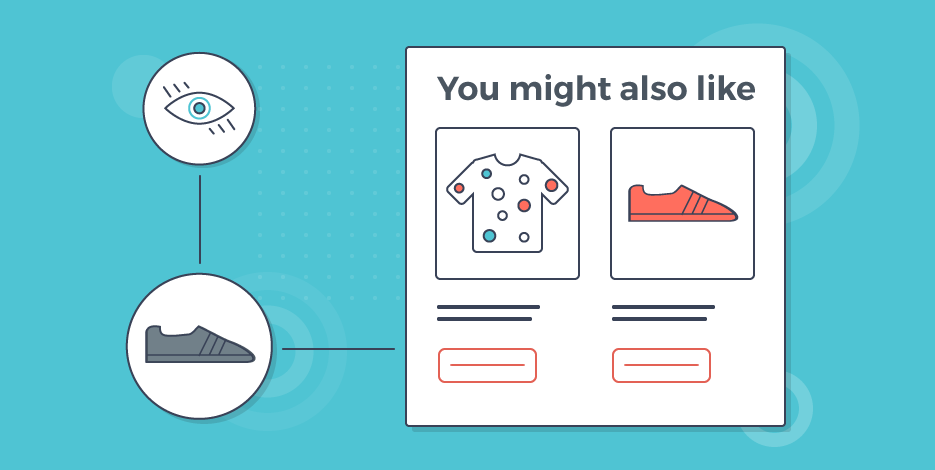
1. **What are data science and analytics?**

***Data science*** is a field that uses scientific methods, processes, and systems to extract knowledge and insights from structured and unstructured data. 

***Data analytics*** is a field that focuses on using data to understand and inform decision-making.



1. **Why data science?**
2. With the right tools, technologies, and algorithms, we can use data and convert it into a distinct business advantage
3. Data Science can help you to detect fraud using advanced machine-learning algorithms
4. It helps you to prevent any significant monetary losses
5. Allows to build intelligence ability in machines
6. You can perform sentiment analysis to gauge customer brand loyalty
7. It enables you to take better and faster decisions
8. It helps you to recommend the right product to the right customer to enhance your business



1. **How does data science work?**

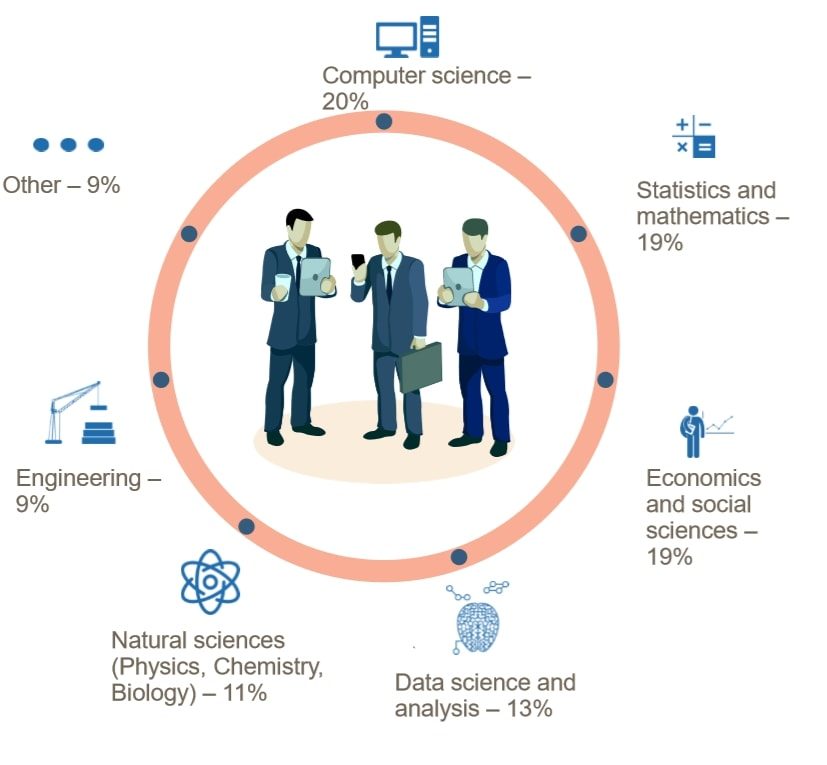
Data science involves using techniques from a range of disciplines to collect, process, and analyze data, to find patterns and insights that can inform decisions or understanding. This may involve cleaning and organizing data, building and training models, and visualizing and communicating results.

1. **What are the different jobs roles in Data Science?**
2. ***Data Scientist:*** Analyzes and interprets large datasets to solve complex problems using statistical techniques and machine learning algorithms.
3. ***Data Analyst:*** Examines and interprets data to draw insights and inform business decisions.
4. ***Statistician:*** Designs and conducts statistical analysis to solve problems and inform decision-making.
5. ***Data Architect:*** Designs and builds data models and systems to support an organization's data needs.
6. ***Data Admin:*** Maintains and manages an organization's data systems and infrastructure.
7. ***Machine Learning Engineer***: Designs, develops and deploys machine learning models in production.
8. ***Data Engineer:*** Builds and maintains data pipelines and infrastructure to store, process, and analyze data.
9. ***Business Intelligence Analyst***: Uses data and analytics to inform business decisions through the creation of reports and dashboards.
10. ***Big Data Engineer***: Builds and maintains systems for handling large volumes of structured and unstructured data.
11. ***Data Visualization Designer:*** Creates visual representations of data and insights to effectively communicate complex ideas to a wide audience.

Top of Form

1. **How does a data scientist work?**
2. Ask the right questions - To understand the business problem.
3. Explore and collect data - From the database, web logs, customer feedback, etc.
4. Extract the data - Transform the data to a standardized format.
5. Clean the data - Remove erroneous values from the data.
6. Find and replace missing values - Check for missing values and replace them with a suitable value (e.g. an average value).
7. Normalize data - Scale the values in a practical range (e.g. 140 cm is smaller than 1,8 m. However, the number 140 is larger than 1,8. - so scaling is important).
8. Analyze data, find patterns and make future predictions.
9. Represent the result - Present the result with useful insights in a way the "company" can understand.

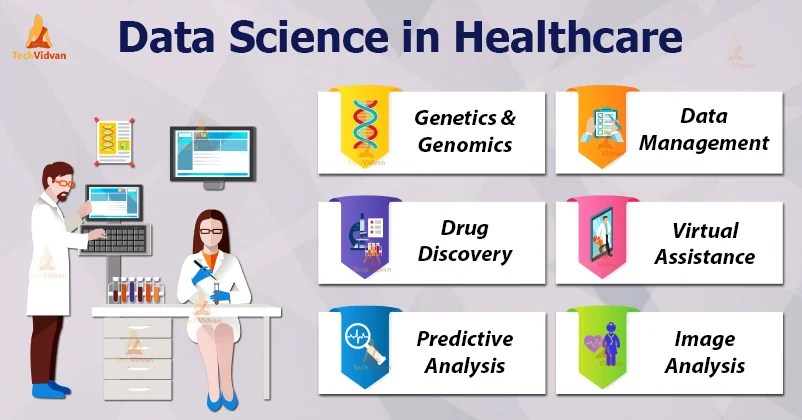


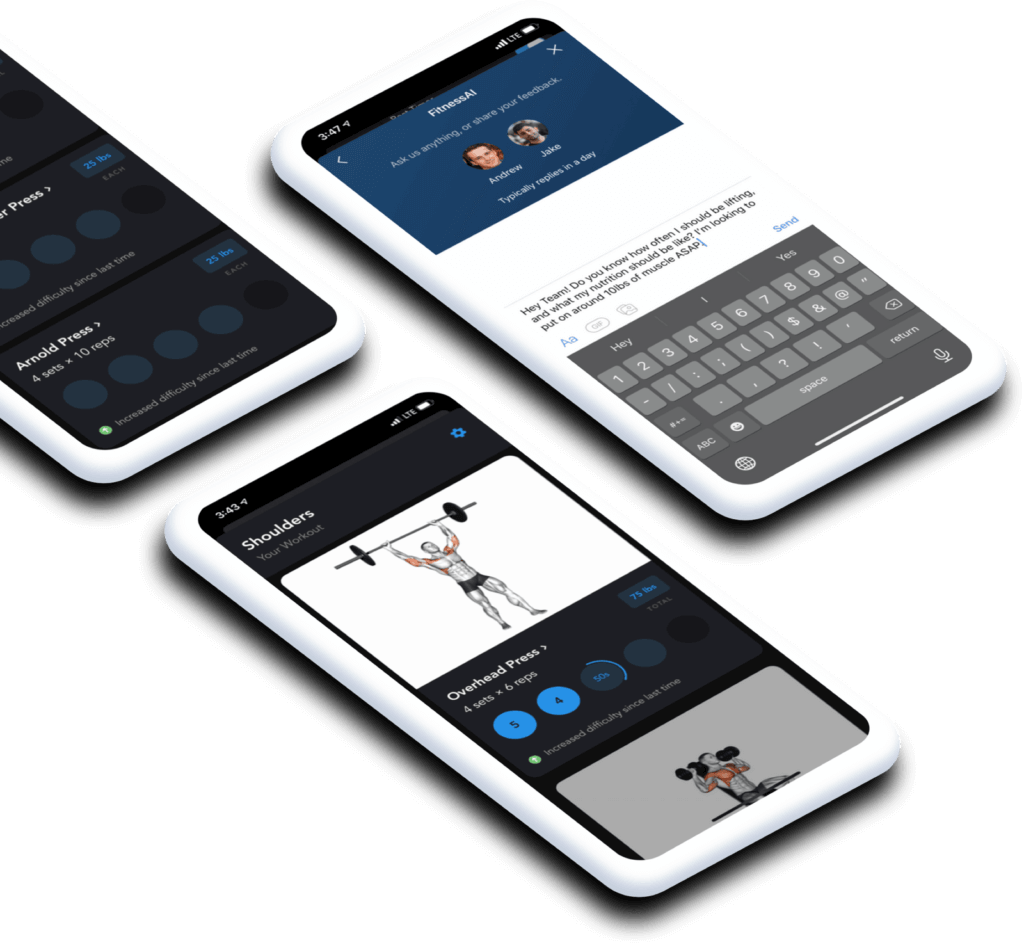
1. **What background is needed to be a data scientist?**
2. Computer Science: Data scientists often have a strong foundation in computer science, with experience in programming languages such as Python, R, and SQL.
3. Statistics: Data scientists should have a strong understanding of statistical concepts and methods, including probability, hypothesis testing, and regression analysis.
4. Mathematics: A background in mathematics, particularly linear algebra and calculus, can be helpful for data scientists as they work with large, complex datasets.
5. Domain Expertise: Data scientists should have a deep understanding of the domain in which they are working, as they will be using data to solve problems and inform decisions in that area. 

It is not a requirement to have a background in technology to pursue a career in data science.

1. **Where is data science needed?**

Data science is needed in a wide variety of industries and organizations, as data is being generated in increasing quantities in nearly every field. Some common examples include healthcare, finance, marketing, e-commerce, and government.



1. **Real-world examples that use data science**
2. Analyzing data from fitness apps to identify trends and suggest personalized workout plans
3. Supply chain optimization in the logistics industry
4. Recommendation systems in marketing & advertising
5. Weather predictions in the agriculture sector
6. Analyzing data from social media and other online platforms to track the spread of misinformation and predict the likelihood of it going viral

# Tools and Technologies used in Data science:

1. ***Programming languages:***

* **Python:** A popular general-purpose programming language that is widely used in data science and machine learning.
* **R:** A programming language specifically designed for statistical analysis and data visualization.
* **SQL:** A programming language used for interacting with relational databases.

1. ***Data storage and processing technologies:***

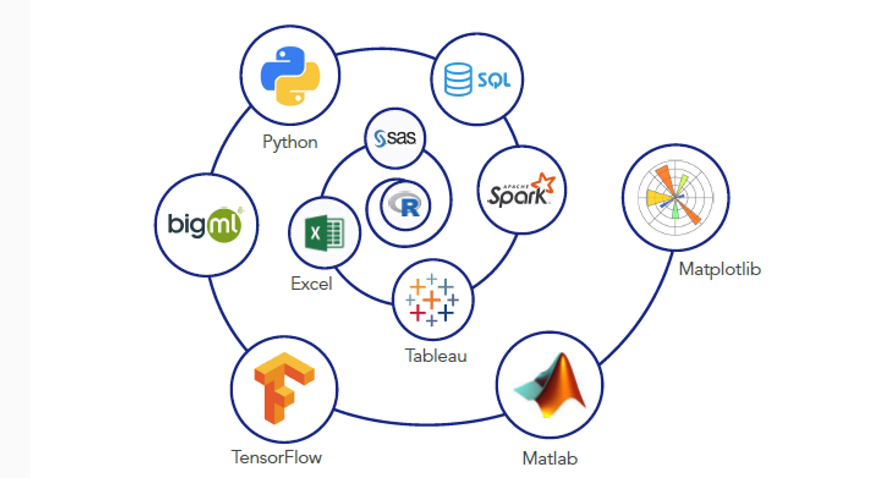
* **Relational databases:** A type of database that stores data in tables and is designed for easy data manipulation and querying. Examples include MySQL and PostgreSQL.
* **NoSQL databases:** A type of database that is designed to handle large amounts of unstructured data and is often used in big data and real-time web applications. Examples include MongoDB and Cassandra.
* **Hadoop:** An open-source big data processing framework that allows for distributed storage and processing of large data sets.
* **Spark:** An open-source big data processing framework that is used for fast data processing and analytics.

1. ***Machine learning frameworks:***

* **TensorFlow:** An open-source machine learning framework developed by Google that is widely used for training and deploying machine learning models.
* **PyTorch:** An open-source machine learning framework developed by Facebook that is popular for its flexibility and ease of use.
* **Scikit-learn:** An open-source machine learning library for Python that is easy to use and has a wide range of machine learning algorithms.

1. ***Data visualization tools:***

* **Matplotlib:** A popular data visualization library for Python.
* **ggplot2:** A data visualization library for R that is particularly well-suited for producing high-quality graphs and charts.
* **Tableau:** A commercial data visualization software that allows users to easily create interactive charts and dashboards.
* **Power BI:** It is particularly well-suited for creating interactive dashboards that can be shared with others and updated in real time.



# Steps included in Data science Process:

1. **ACQUISITION** –

* Data acquisition has been understood as the process of gathering, filtering, and cleaning data before the data is put in a data warehouse or any other storage solution.
* Some sources from where we can acquire data.

1. Website data
2. Purchase data
3. Smart tv data
4. Social data
5. Offline data
6. Mobile data
7. 3 rd Party data
8. **EXPLORATION –**

* The second step is to understand the data that how will use it
* How much information you can get from the data
* To better explain the essence of the data, data exploration refers to the initial step of data processing in which data analysts use data visualization and mathematical methods to define dataset characterizations, such as scale, quantity, and precision.

1. **MUGGING, WRANGLING –**

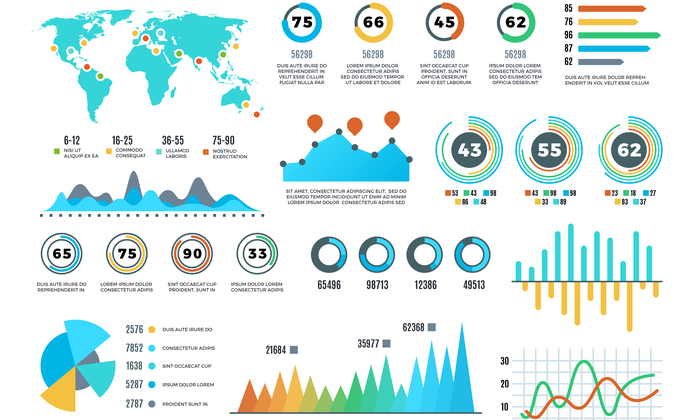
* Data is rarely in the needed form for the desired analysis.
* It is the initial step of preprocessing and refining raw data into content or formats better suited for analysis.

1. **Analysis and Modeling –**

* Data modeling is a collection of methods and tools used to explain and analyze how data can be processed, modified and maintained.
* Data analysis allows the analyst to extrapolate insights from huge bulks of data, it is done with the help of statistical and machine-learning techniques.

1. **COMMUNICATION (Data Visualization) –**

* At the end of the pipeline, we need to give the data in a compelling form and structure, sometimes to ourselves to inform the next iteration, and sometimes to a completely different audience.
* The data products produced can be a simple one-off report or a scalable web product that will be used interactively by millions.



# Capstone Projects:

1. **Identify a topic:** The first step is to identify a topic of interest. This might be based on the area of study or a specific problem or issue that they want to explore.

For example, a student interested in data science might choose a topic related to analyzing and predicting patterns in large datasets using machine learning techniques.

1. **Conduct research:** Once the topic has been identified, we need to conduct research to gather information and resources on their chosen topic. This might involve using online databases, visiting libraries, and conducting interviews or surveys.
2. **Gather and prepare data:** To perform EDA and ML on a dataset, students will need to gather and prepare the data for analysis. This might involve acquiring and cleaning the data, selecting relevant features, and splitting the data into training and testing sets.
3. **Conduct EDA:** The next step is to perform EDA on the dataset. This might involve exploring the data using statistical and visual techniques, such as generating summary statistics, creating plots and charts, and identifying trends and patterns.
4. **Select and apply ML techniques:** Based on the results of the EDA, students will need to select and apply appropriate ML techniques to the dataset. This might involve training and testing different models, tuning model hyper-parameters, and evaluating the performance of the models.
5. **Create a proposal:** After conducting the EDA and ML, we will need to create a proposal outlining the scope and objectives of the capstone project. The proposal should include a detailed description of the project, the research methods that were used, and the expected outcomes.
6. **Get approval:** The proposal will need to be reviewed and approved by the capstone project committee or other appropriate authority before we proceed.
7. **Conduct the project:** Once the proposal has been approved, we can begin working on project. This might involve collecting and analyzing additional data, creating visualizations or other materials to present the results, and writing a final report.
8. **Present the project:** The final step in completing a project is to present the results of the project to the appropriate audience, which might be a panel of experts, classmates, or the general public. This presentation is typically done in the form of a written report, oral presentation, or other format.