**What is Data Science**

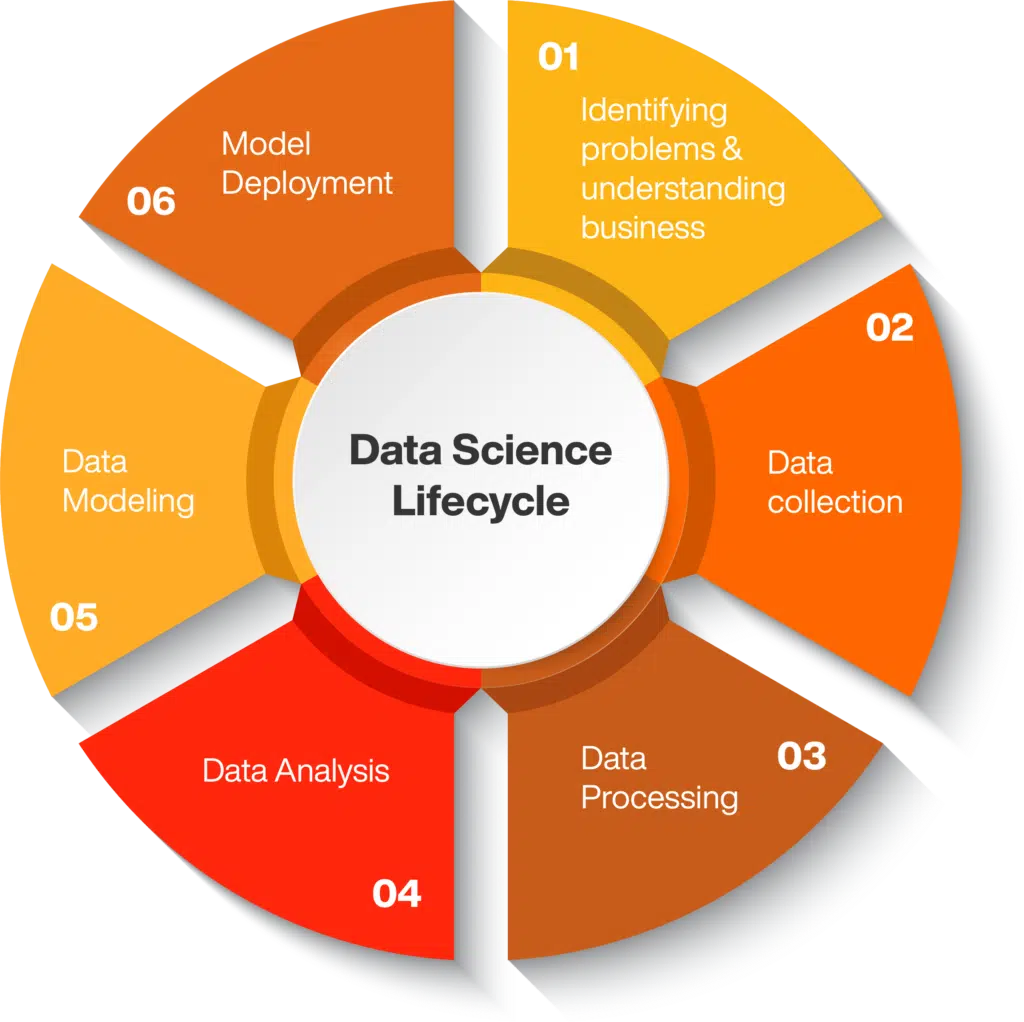
Data science is defined as an interdisciplinary field that involves extracting knowledge and insights from data using scientific methods, algorithms, and systems. Data Science combines elements of mathematics, statistics, computer science, and domain expertise to analyze large volumes of structured and unstructured data. The goal of data science is to uncover patterns, trends, and relationships within the data to make informed decisions, solve complex problems, and create predictive models.

The term “data science” combines two key elements: “data” and “science.”

**Data:** It refers to the raw information that is collected, stored, and processed. In today’s digital age, enormous amounts of data are generated from various sources such as sensors, social media, transactions, and more. This data can come in structured formats (e.g., databases) or unstructured formats (e.g., text, images, videos).

**Science:** It refers to the systematic study and investigation of phenomena using scientific methods and principles. Science involves forming hypotheses, conducting experiments, analyzing data, and drawing conclusions based on evidence.

**Data Science Life cycle**



**Identifying problems and understanding business** : Discovering the answers for basic questions including requirements, priorities and budget of the project.

**Data Collection :** Collecting data from relevant sources either in structured or unstructured form.

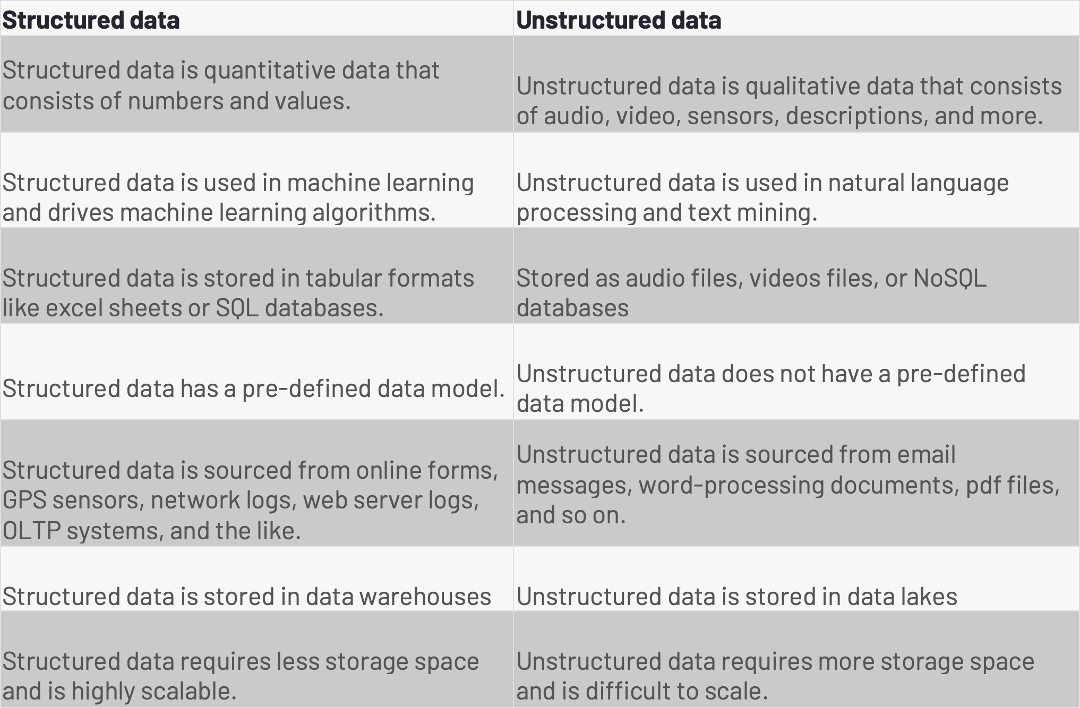
**Data processing** : Processing and fine-tuning the raw data, critical for the goodness of the overall project.

**Data analysis** : Capturing ideas about solutions and factors that influence the data life cycle.

**Data modelling** : Preparing the appropriate model to achieve desired performance.

**Model deployment** : Executing the analysed model in desired format and channel.

**Difference between structured and unstructured data**



**What is Semi-structured data?**

Semi-structured data is a type of data that is not purely structured, but also not completely unstructured. It contains some level of organization or structure, but does not conform to a rigid schema or data model, and may contain elements that are not easily categorized or classified.

Semi-structured data is typically characterized by the use of metadata or tags that provide additional information about the data elements. For example, an XML.

Other examples of semi-structured data include JSON, which is commonly used for exchanging data between web applications, and log files, which often contain a mix of structured and unstructured data.

**Roadmap of Data Science**



**Applications of Data Science**

There are various applications of data science, including:

**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 is one of the most commonly known applications of data science. in images and detecting objects in an image is one of the most popular data science applications.

**4. Recommendation Systems**

Next up in the data science and its applications list comes Recommendation Systems. Netflix and Amazon give movie and product recommendations based on what you like to watch, purchase, or browse on their platforms..

**5. Fraud Detection**

Fraud detection comes the next in the list of applications of data science. Banking and financial institutions use data science and related algorithms to detect fraudulent transactions.

**6. Speech recognition**

Speech recognition is one of the most commonly known applications of data science. It is a technology that enables a computer to recognize and transcribe spoken language into text.

**What is Statistic & its type**

Statistics is a method of interpreting, analysing and summarising the data. Hence, the types of statistics are categorised based on these features: Descriptive and inferential statistics. Based on the representation of data such as using,pie charts, bar graphs, or tables, we analyse and interpret it.

**Types of Statistics**

Statistics have majorly categorised into two types:

Descriptive statistics

Inferential statistics

Statistics have majorly categorised into two types:

**Descriptive Statistics**

In this type of statistics, the data is summarised through the given observations. The summarisation is one from a sample of population using parameters such as the mean or [standard deviation](https://byjus.com/maths/standard-deviation/).

Descriptive statistics is a way to organise, represent and describe a collection of data using tables, graphs, and summary measures. For example, the collection of people in a city using the internet or using Television.

Descriptive statistics are also categorised into four different categories:

Measure of frequency

Measure of dispersion

Measure of central tendency

Measure of position

**Inferential Statistics**

This type of statistics is used to interpret the meaning of Descriptive statistics. That means once the data has been collected, analysed and summarised then we use these stats to describe the meaning of the collected data. Or we can say, it is used to draw conclusions from the data that depends on random variations such as observational errors, sampling variation, etc.

**Data wrangling**

Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data Wrangling is also known as Data Munging.

Data Wrangling is a crucial topic for Data Science and Data Analysis. Pandas Framework of Python is used for Data Wrangling. [Pandas](https://www.geeksforgeeks.org/python-pandas-dataframe/) is an open-source library in [Python](https://www.geeksforgeeks.org/python-programming-language/) specifically developed for Data Analysis and Data Science. It is used for processes like data sorting or filtration, Data grouping, etc.

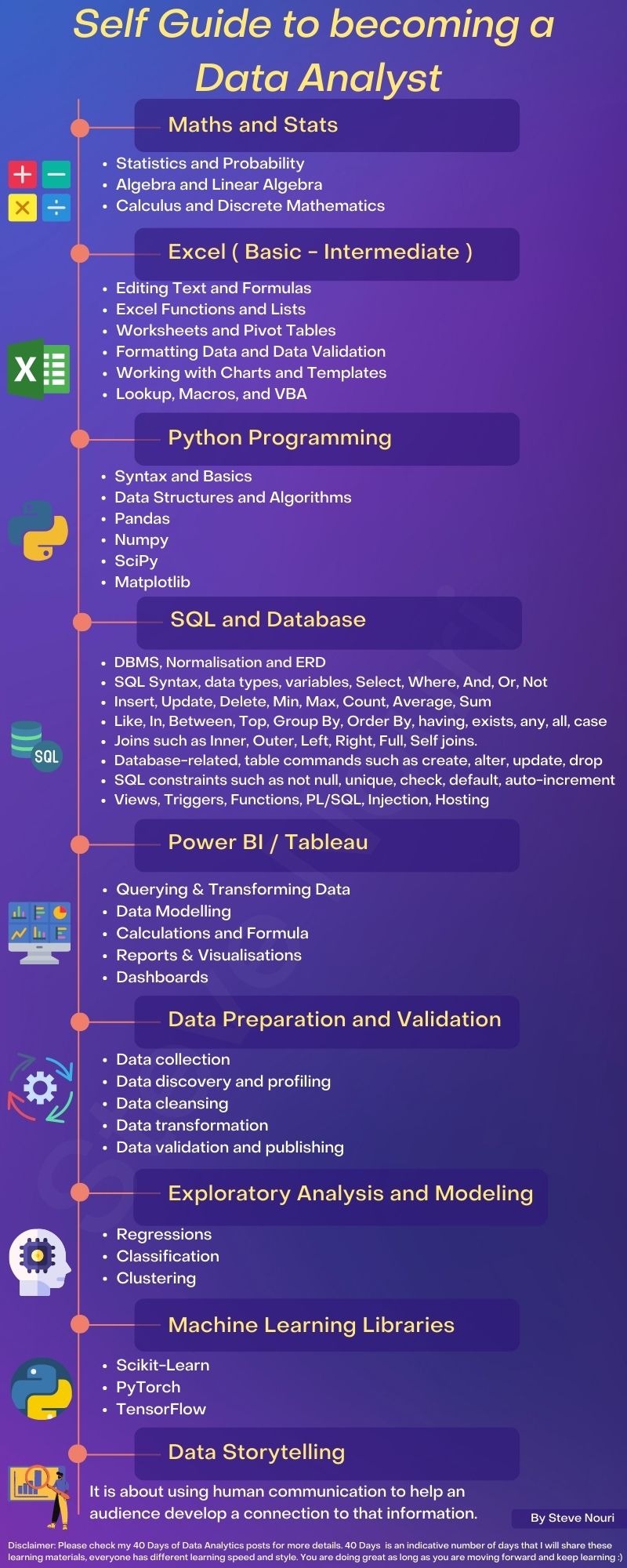
Data wrangling in Python deals with the below functionalities:  
1. Data exploration

2. Dealing with missing values

3. Reshaping data

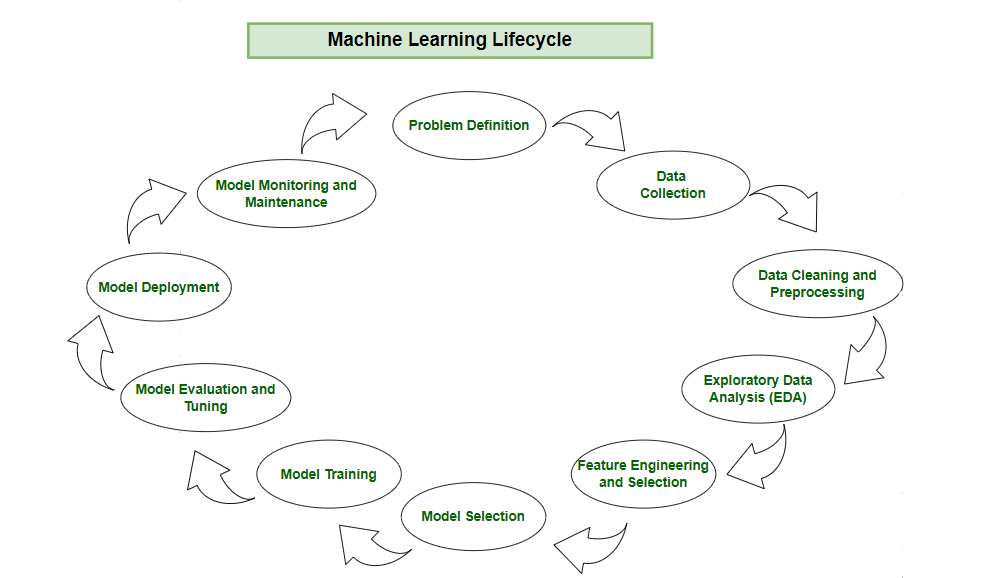
4. Filtering data

**Roadmap of Data Analytics**



**Machine Learning Lifecycle**

Machine learning, a subset of[artificial intelligence,](https://www.geeksforgeeks.org/artificial-intelligence-an-introduction/) teaches computers to mimic human thinking by training them with real-world data. To train a machine with specific data we have to follow predefined steps and this whole process is known as a machine learning lifecycle.



**Step 1: Problem Definition**

Embarking on the machine learning journey involves a well-defined lifecycle, starting with the crucial step of problem definition.

**Step 2:**[**Data Collection**](https://www.geeksforgeeks.org/data-collection-its-methods/)

This phase involves the systematic gathering of datasets that will serve as the raw material for model development.

**Step 3:**[**Data Cleaning and Preprocessing**](https://www.geeksforgeeks.org/data-cleansing-introduction/)

the machine learning journey advances to the critical stages of[data cleaning](https://www.geeksforgeeks.org/data-cleansing-introduction/) and preprocessing. Raw data, is often messy and unstructured. Data cleaning involves addressing issues such as missing values, outliers, and inconsistencies

**Step 4:**[**Exploratory Data Analysis (EDA)**](https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/)

focus turns to understanding the underlying patterns and characteristics of the collected data. Exploratory Data Analysis (EDA) emerges as a pivotal phase, where practitioners leverage various statistical and visual tools to gain insights into the dataset’s structure.

**Step 5: Feature Engineering and Selection**

Feature engineering takes center stage as a transformative process that elevates raw data into meaningful predictors.

**Step 6: Model Selection**

Model selection is a pivotal decision that determines the algorithmic framework guiding the predictive capabilities of the machine learning solution.

**Step 7: Model Training**

This process involves exposing the model to historical data, allowing it to learn patterns, relationships, and dependencies within the dataset.

**Step 8: Model Evaluation and Tuning**

[Model evaluation](https://www.geeksforgeeks.org/machine-learning-model-evaluation/) involves rigorous testing against validation datasets, employing metrics such as accuracy, precision, recall, and F1 score to gauge its effectiveness.

**Step 9: Model Deployment**

pon successful evaluation, the machine learning model transitions from development to real-world application through the deployment phase.

**Measures of Central Tendency in Statistics**

Central Tendencies in Statistics are the numerical values that are used to represent mid-value or central value a large collection of numerical data. These obtained numerical values are called central or average values in [Statistics](https://www.geeksforgeeks.org/statistics/).

**Mean**

Arithmetic mean is defined as the sum of the individual observations (xi) divided by the total number of observations N.

xˉ=N∑xi​​

OR

Mean = Sum of all Observations ÷ Total number of Observations

Example: If there are 5 observations, which are 27, 11, 17, 19, and 21 then the mean (u) is given by

u = (27 + 11 + 17 + 19 + 21) ÷ 5

⇒ u = 95 ÷ 5

⇒  u  = 19

**Median**

Median of any distribution is that value that divides the distribution into two equal parts such that the number of observations above it is equal to the number of observations below it. Thus, the median is called the central value of any given data either grouped or ungrouped.

Example 1: If the observations are 25, 36, 31, 23, 22, 26, 38, 28, 20, 32 then the Median is given by

*Arranging the data in ascending order: 20, 22, 23, 25, 26, 28, 31, 32, 36, 38*

*N = 10 which is even then*

*Median = Arithmetic mean of values at (10 ÷ 2)th and [(10 ÷ 2) + 1]th position*

*⇒ Median = (Value at 5th position + Value at 6th position) ÷ 2*

*⇒ Median = (26 + 28) ÷ 2*

*⇒ Median = 27*

**Mode**

Mode is the value of that observation which has a maximum frequency corresponding to it. In other, that observation of the data occurs the maximum number of times in a dataset.

**Encoding**

Hand written note

**Data Normalization**

Normalization is an essential step in the [preprocessing](https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/) of data for machine learning models, and it is a feature scaling technique. Normalization is especially crucial for data manipulation, scaling down, or up the range of data before it is utilized for subsequent stages in the fields of soft computing, cloud computing, etc. Min-max scaling and [Z-Score](https://www.geeksforgeeks.org/z-score-in-statistics/) Normalisation [(Standardisation](https://www.geeksforgeeks.org/normalization-vs-standardization/)) are the two methods most frequently used for normalization in feature scaling.

**Min-Max normalization:**This method of normalising data involves transforming the original data linearly. The data’s minimum and maximum values are obtained, and each value is then changed using the formula that follows.



**Data standardization**

In Machine Learning we train our data to predict or classify things in such a manner that isn’t hardcoded in the machine. So for the first, we have the Dataset or the input data to be pre-processed and manipulated for our desired outcomes.

