Data

According to Oxford *"Data is distinct pieces of information, usually formatted in a special way".* Data is measured, collected and reported, and analyzed, whereupon it is often visualized using graphs, images or other analysis tools. Raw data ("unprocessed data") may be a collection of numbers or characters before it's been "cleaned" and corrected by researchers. It must be corrected so that we can remove outliers, instrument or data entry errors. Data processing commonly occurs in stages, and therefore the "processed data" from one stage could also be considered the "raw data" of subsequent stages. Field data is data that's collected in an uncontrolled "in situ" environment. Experimental data is the data that is generated within the observation of scientific investigations.
Data can be generated by:

- Humans

  Human-generated data is data that is created by people through human action, as opposed to machine learning or other artificial means. This can include anything from text data to social media posts to pictures and videos. Even though machine generated data and technologies like generative AI have become more popular. human-generated data remains an important source of information for businesses and tech developers.

- Machines:

  Machine generated data (MGD) is information that is produced by mechanical or digital devices without human intervention. MGD can often be used to describe data which has been generated by an organization's industrial control systems and mechanical devices that are designed to carry out a single function. Manually entered data by an end user is not considered to be machine generated.
  Machine generated data can be found across all sectors of computing and a business. This type of data makes use of computers in any of their daily operations, and this type of data can be generated by users unknowingly.
  Sensor data is an example of machine generated data.
  Machine generated data is valuable because it contains a definitive, real time record of all user behavior and activity, as well as transactions, applications, servers, networks and mobile devices.

- Human-Machine combines.

It can often be generated anywhere where any information is generated and stored in structured or unstructured formats.
Why is data important ?

- Data helps in making better decisions.
- Data helps in solving problems by finding the reason for underperformance.
- Data helps one to evaluate the performance.
- Data helps one improve processes.

- Data helps one understand consumers and the market.

Types of Data:

Generally data can be classified into two parts:

1. **Categorical Data:**
   In categorical data we see the data which have a defined category, for example:
   a. Marital Status
   b. Political Party
   c. Eye color
2. **Numerical Data:**
   Numerical data can further be classified into two categories:
   a. **Discrete Data:**
      Discrete data contains the data which have discrete numerical values for example Number of Children, Defects per Hour etc.
   b. **Continuous Data:**
      Continuous data contains the data which have continuous numerical values for example Weight, Voltage etc.

At advanced level, we can further classify the data into four parts:

1. **Nominal Scale:**
   A nominal scale classifies data into several distinct categories in which no ranking criteria is implied. For example Gender, Marital Status.
2. **Ordinary Scale:**
   An ordinal scale classifies data into distinct categories during which ranking is implied For example:
   a. Faculty rank : Professor, Associate Professor, Assistant Professor
   b. Students grade : A, B, C, D.E.F
3. **Interval scale:**
   An interval scale may be an ordered scale during which the difference between measurements is a meaningful quantity but the measurements don't have a true zero point. For example:
   a. Temperature in Fahrenheit and Celsius.
   b. Years
4. **Ratio scale:**
   A ratio scale may be an ordered scale during which the difference between the measurements is a meaningful quantity and therefore the measurements have a true zero point. Hence, we can perform arithmetic operations on real scale data. For example : Weight, Age, Salary etc.

Data Science

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

expertise from various domains such as statistics, computer science, mathematics, domain knowledge, and data engineering to uncover hidden patterns, make predictions, and support decision-making.
Let's dive into each step of the data science process in more detail:

1. Problem Formulation:
- In this initial phase, the data scientist works closely with domain experts and stakeholders to clearly define the problem or question to be addressed using data.
- It involves understanding the business or research context, setting specific goals, and defining what success looks like.

2. Data Collection:
- Data can be collected from various sources such as databases, APIs, web scraping, files, sensors, surveys, or social media platforms.
- Data collection might involve selecting the relevant data sources, setting up data pipelines, and automating data retrieval processes.

3. Data Cleaning and Preprocessing:
-Raw data often contains errors, missing values, outliers, and inconsistencies. Data cleaning involves:
- Handling missing data through imputation or removal.
- Dealing with outliers, which may involve capping/extending data or removing extreme values.
- Correcting data format issues, such as data type conversions.

-Data preprocessing includes:
- Normalization or scaling to ensure all features have similar scales.
- Encoding categorical variables into numerical format.
- Feature selection and engineering to create relevant variables or transform existing ones.

4. Exploratory Data Analysis (EDA):
EDA is a crucial step for understanding the data and gaining insights. Common activities include:
- Creating summary statistics like mean, median, and standard deviation.
- Visualizing data using plots such as histograms, scatter plots, and heatmaps.
- Identifying patterns, correlations, and trends in the data.
- Detecting and addressing data anomalies.

5. Feature Selection and Engineering:
- Feature selection involves choosing the most important features that contribute significantly to the model's performance while reducing noise.
- Feature engineering involves creating new features based on domain knowledge or data patterns to improve model performance.

6. Model Selection:
- Select an appropriate machine learning or statistical model based on the problem type (classification, regression, clustering) and data characteristics.
- Consider factors like model complexity, interpretability, and scalability when choosing the model.

7. Data Splitting:
- Split the dataset into three parts: a training set, a validation set, and a test set. Common splits are 70% for training, 15% for validation, and 15% for testing.
- The training set is used to train the model, the validation set is used for hyperparameter tuning, and the test set is used for final model evaluation.

8. Model Training:
- Train the selected model on the training data by feeding it with input features and corresponding labels.
- The model iteratively learns from the data and updates its parameters to make accurate predictions.

9. Model Evaluation:
- Assess the model's performance using appropriate evaluation metrics that are relevant to the problem, such as accuracy, precision, recall, F1-score, or mean squared error (MSE).
- Cross-validation can be used to estimate the model's generalization performance.

10. Model Tuning:
- Adjust the model's hyperparameters (e.g., learning rate, regularization strength) based on the results of model evaluation to optimize performance.
- Revisit feature selection and engineering if necessary.

11. Deployment:
- If the model meets the desired performance criteria, deploy it into a production environment where it can make real-time predictions.
- Integration with existing systems and applications is a critical aspect of deployment.

12. Monitoring and Maintenance:
- Continuously monitor the model's performance in the production environment.
- Implement strategies to handle data drift (changes in the distribution of input data) and concept drift (changes in the relationship between inputs and outputs).
- Periodically retrain and update the model to maintain accuracy.

13. Interpretability and Communication:
- Interpret the model's predictions to understand how it arrived at its decisions. Techniques like feature importance analysis and model visualization can help.
- Communicate results effectively to stakeholders through reports, visualizations, and dashboards.

14. Feedback Loop:
- Collect feedback from users and stakeholders to improve the model and refine the data science process.
- Iterate on the model and the problem formulation as needed based on feedback and changing business needs.

The data science process is iterative, and it may involve going back and forth between these steps as you gain more insights, refine your models, and adapt to changing data and requirements. Collaboration, domain knowledge, and effective communication are essential components of successful data science projects.