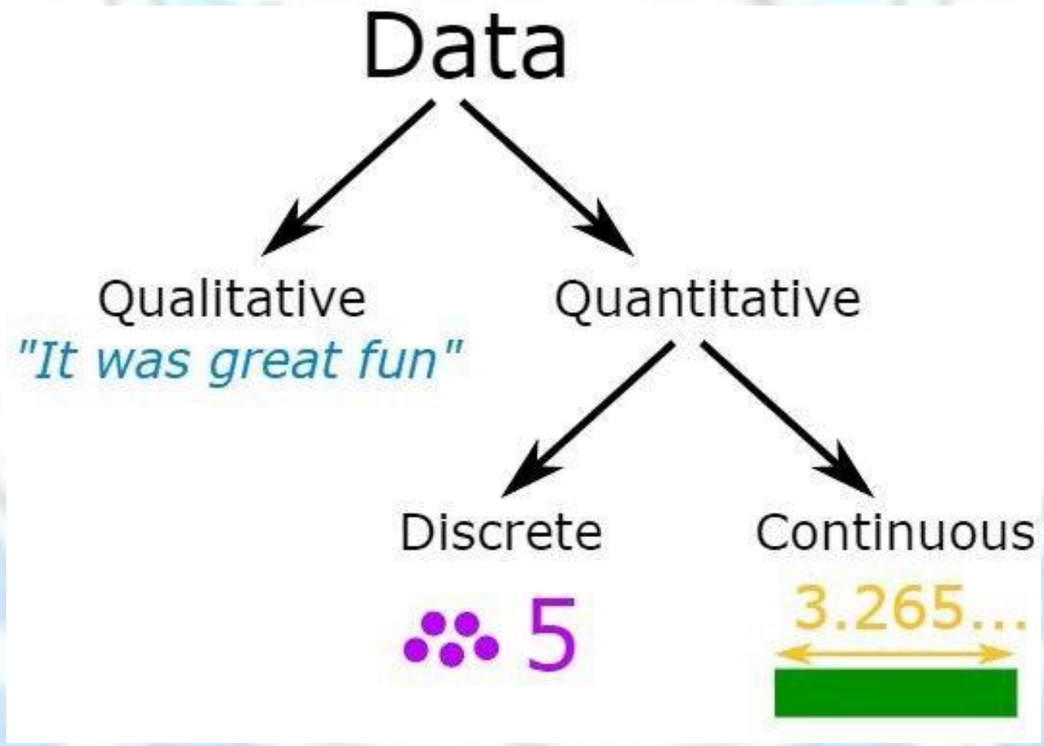
Data representation and preprocessing: Normalization, encoding, missing values

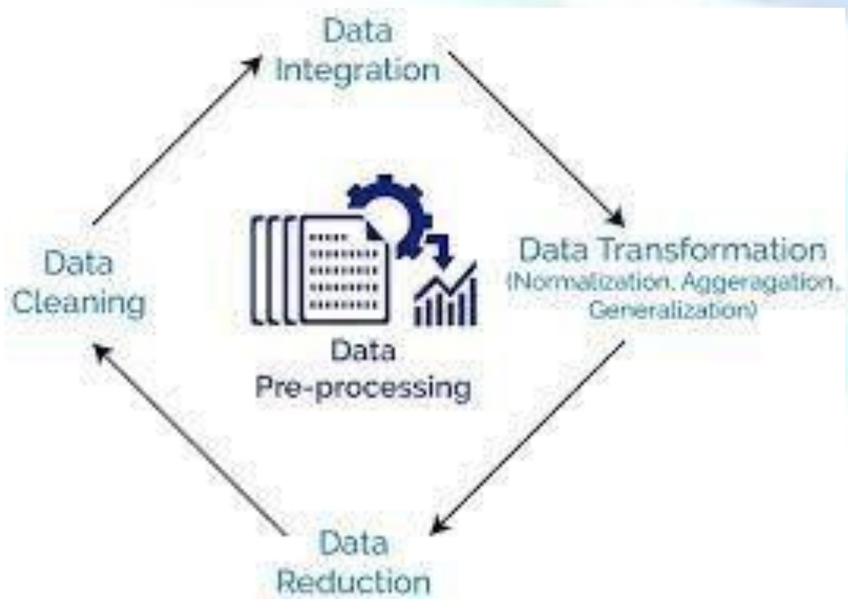
Data preprocessing is the process of preparing raw data for analysis by cleaning and transforming it into a usable format.

Data preprocessing involves several techniques such as cleaning the data to handle missing values, removing outliers, scaling features, encoding categorical variables, and splitting the data into training and testing sets. These techniques are key for ensuring the data is in a consistent and usable format for the ML algorithms.

Data is a collection of facts, such as numbers, words, measurements, observations, or just descriptions of things.



Data Preprocessing in Machine Learning

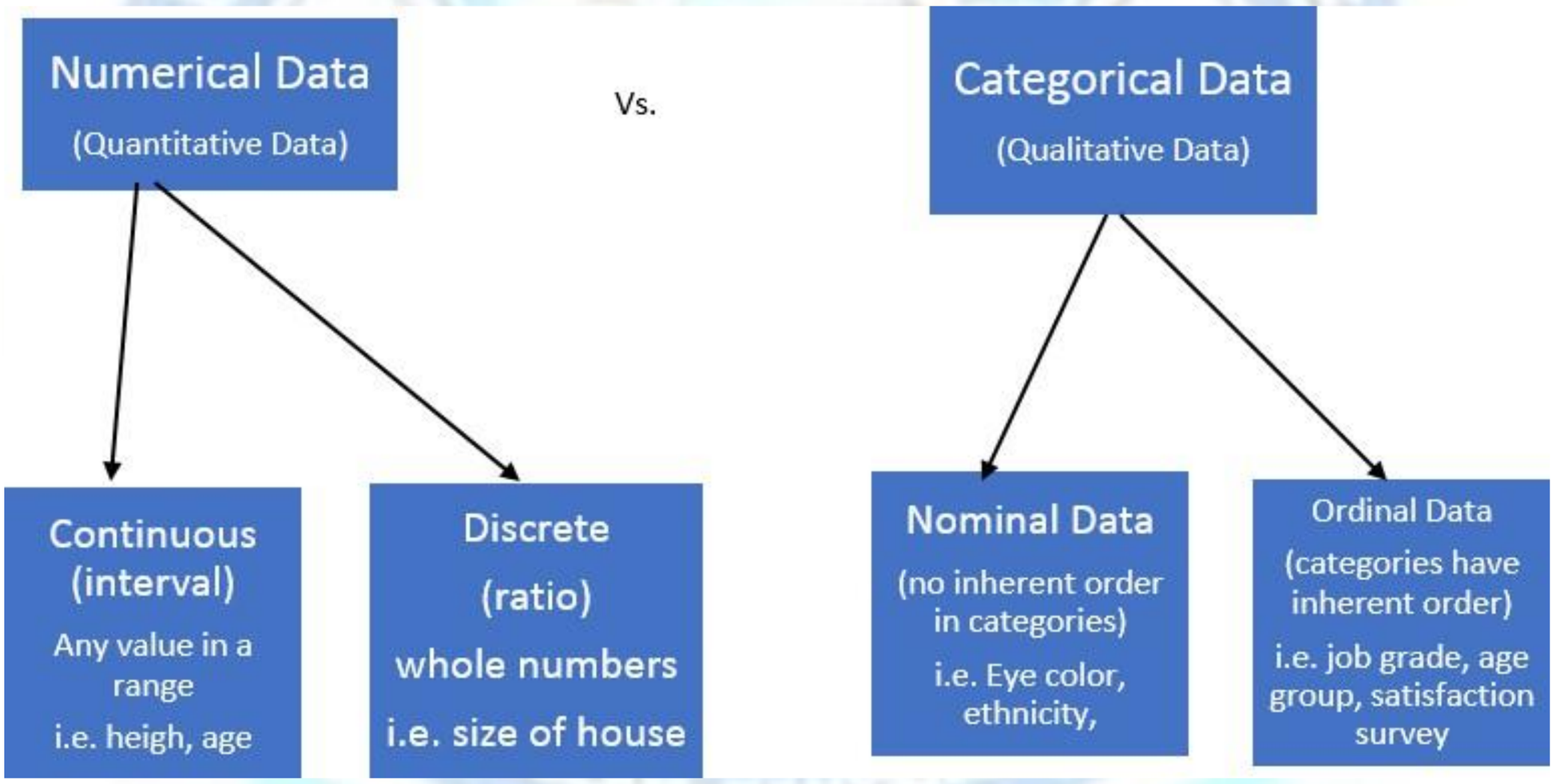


Data preprocessing in machine learning is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. It is a data mining technique that involves the transformation of raw data into an insightful and organized format. A real-world data generally contains noises, missing values, and maybe in an unusable format that cannot be directly used for machine learning models. For resolving such issues it prepares raw data for further processing. Data preprocessing is a required task for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

Various Types of Data

Numerical & Categorical Data

Numerical data is a data type expressed in numbers, rather than natural language description. Sometimes called quantitative data, numerical data is always collected in number form. Categorical variables represent types of data that may be divided into groups. Examples of categorical variables are race, sex, age group, and educational level.



Handling Missing Values

Missing values occur due to:

- **Data entry errors**
- **Sensor failures**
- **Incomplete surveys**

Methods to Handle Missing Values

(a) **Deletion**

(b) **Mean / Median / Mode Imputation**

(c) **Forward / Backward Fill**

(d) **Model-based Imputation**

Handling the missing values is one of the greatest challenges faced by analysts because making the right decision on how to handle it generates robust data models. Let us look at different ways of imputing the missing values.

Determine the impact of missing values on your analysis or model. Consider the percentage of missing values in each column and their importance to the overall data set.

If the percentage of missing values is small and those rows or columns are not critical, you can choose to remove them

For numerical features, you can impute missing values using techniques like mean, median, or mode imputation