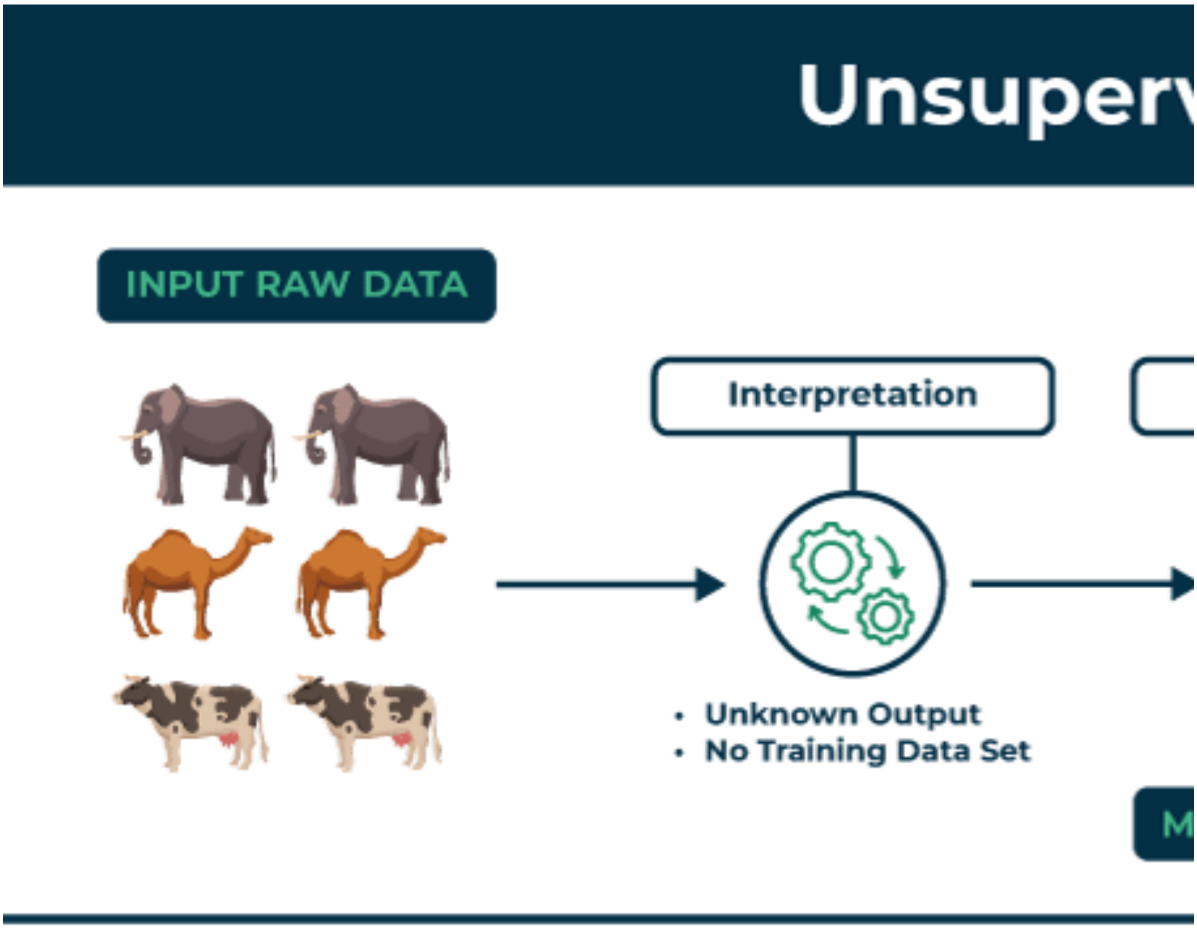
What is Unsupervised Learning?

Unsupervised learning is a branch of machine learning that deals with unlabeled data. Unlike supervised learning, where the data is labeled with a specific category or outcome, unsupervised learning algorithms **are tasked with finding patterns and relationships within the data without any prior knowledge of the data's meaning**. Unsupervised machine learning algorithms **find hidden patterns and data without any human intervention, i.e., we don't give output to our model. The training model has only input parameter values and discovers the groups or patterns on its own.**



https://www.geeksforgeeks.org/machine-learning/unsupervised-learning/

The image shows set of animals: elephants, camels, and cows that represents raw data that the unsupervised learning algorithm will process.

- The "Interpretation" stage signifies that the algorithm doesn't have predefined labels or categories for the data. It needs to figure out how to group or organize the data based on inherent patterns.
- **Algorithm** represents the core of unsupervised learning process using techniques like clustering, dimensionality reduction, or anomaly detection to identify patterns and structures in the data.
- **Processing** stage shows the algorithm working on the data.

The output shows the results of the unsupervised learning process. In this case, the algorithm might have grouped the animals into clusters based on their species (elephants, camels, cows).

How does unsupervised learning work?

Unsupervised learning works by analyzing unlabeled data to identify patterns and relationships. The data is not labeled with any predefined categories or outcomes, so the algorithm must find these patterns and relationships on its own. This can be a challenging task, but it can also be very rewarding, as it can reveal insights into the data that would not be apparent from a labeled dataset.

Data-set in Figure A is Mall data that contains information about its clients that subscribe to them.

Once subscribed they are provided a membership card and the mall has complete information about the customer and his/her every purchase. Now using this data and unsupervised learning techniques, the mall can easily group clients based on the parameters we are feeding in.

The input to the unsupervised learning models is as follows:

- **Unstructured data**: May contain noisy(meaningless) data, missing values, or unknown data
- **Unlabeled data**: Data only contains a value for input parameters, there is no targeted value(output). It is easy to collect as compared to the labeled one in the Supervised approach.

Unsupervised Learning Algorithms

There are mainly 3 types of Algorithms which are used for Unsupervised dataset.

- **Clustering**
- **Association Rule Learning**
- **Dimensionality Reduction**

1. Clustering Algorithms

Clustering in unsupervised machine learning is the process of grouping unlabeled data into clusters based on their similarities. The goal of clustering is to identify patterns and relationships in the data without any prior knowledge of the data's meaning.

Broadly this technique is applied to group data based on different patterns, such as similarities or differences, our machine model finds. These algorithms are used to process raw, unclassified data objects into groups. For example, in the above figure, we have not given output parameter values, so this technique will be used to group clients based on the input parameters provided by our data.

Some common clustering algorithms:

- ***K-means Clustering**: Groups data into K clusters based on how close the points are to each other.*
- ***Hierarchical Clustering**: Creates clusters by building a tree step-by-step, either merging or splitting groups.*
- ***Density-Based Clustering (DBSCAN)**: Finds clusters in dense areas and treats scattered points as noise.*
- ***Mean-Shift Clustering**: Discovers clusters by moving points toward the most crowded areas.*
- ***Spectral Clustering**: Groups data by analyzing connections between points using graphs.*

2. Association Rule Learning

Association rule learning is also known as association rule mining is a common technique used to discover associations in unsupervised machine learning. This technique is a rule-based ML technique that finds out some very useful relations between parameters of a large data set. This technique is basically used for market basket analysis that helps to better understand the relationship between different products.

For e.g. shopping stores use algorithms based on this technique to find out the relationship between the sale of one product w.r.t to another's sales based on customer behavior. **Like if a customer buys milk, then he may also buy bread, eggs, or butter**. Once trained well, such models can be used to increase their sales by planning different offers.

Some common Association Rule Learning algorithms:

- ***Apriori Algorithm****: Finds patterns by exploring frequent item combinations step-by-step.*
- ***FP-Growth Algorithm****: An Efficient Alternative to Apriori. It quickly identifies frequent patterns without generating candidate sets.*
- ***Eclat Algorithm****: Uses intersections of itemsets to efficiently find frequent patterns.*
- ***Efficient Tree-based Algorithms****: Scales to handle large datasets by organizing data in tree structures.*

3. Dimensionality Reduction

Dimensionality reduction is the process of reducing the number of features in a dataset while preserving as much information as possible. This technique is useful for improving the performance of machine learning algorithms and for data visualization.

Imagine a dataset of 100 features about students (height, weight, grades, etc.). To focus on key traits, you reduce it to just 2 features: height and grades, making it easier to visualize or analyze the data.

Here are some popular ***Dimensionality Reduction algorithms****:*

- ***Principal Component Analysis (PCA):*** *Reduces dimensions by transforming data into uncorrelated principal components.*
- ***Linear Discriminant Analysis (LDA):*** *Reduces dimensions while maximizing class separability for classification tasks.*
- ***Non-negative Matrix Factorization (NMF):*** *Breaks data into non-negative parts to simplify representation.*
- ***Locally Linear Embedding (LLE):*** *Reduces dimensions while preserving the relationships between nearby points.*
- ***Isomap:*** *Captures global data structure by preserving distances along a manifold.*

Challenges of Unsupervised Learning

Here are the key challenges of unsupervised learning:

- **Noisy Data**: Outliers and noise can distort patterns and reduce the effectiveness of algorithms.
- **Assumption Dependence**: Algorithms often rely on assumptions (e.g., cluster shapes), which may not match the actual data structure.
- **Overfitting Risk**: Overfitting can occur when models capture noise instead of meaningful patterns in the data.
- **Limited Guidance**: The absence of labels restricts the ability to guide the algorithm toward specific outcomes.
- **Cluster Interpretability**: Results, such as clusters, may lack clear meaning or alignment with real-world

categories.

- **Sensitivity to Parameters**: Many algorithms require careful tuning of hyperparameters, such as the number of clusters in k-means.
- **Lack of Ground Truth**: Unsupervised learning lacks labeled data, making it difficult to evaluate the accuracy of results.

Applications of Unsupervised learning

Unsupervised learning has diverse applications across industries and domains. Key applications include:

- **Customer Segmentation:** Algorithms cluster customers based on purchasing behavior or demographics, enabling targeted marketing strategies.
- **Anomaly Detection:** Identifies unusual patterns in data, aiding fraud detection, cybersecurity, and equipment failure prevention.
- **Recommendation Systems**: Suggests products, movies, or music by analyzing user behavior and preferences.
- **Image and Text Clustering**: Groups similar images or documents for tasks like organization, classification, or content recommendation.
- **Social Network Analysis**: Detects communities or trends in user interactions on social media platforms.
- **Astronomy and Climate Science:** Classifies galaxies or groups weather patterns to support scientific research