:≡ **Course Content**

# Unsupervised Learning

Data science and machine learning algorithmic techniques can be divided into three categories:

- **Supervised Learning:** The algorithm has prior data (labeled data) to learn from.
- **Unsupervised Learning:** The algorithm has no prior labeled data to learn from i.e., it learns from unlabeled data by finding patterns among examples and grouping them accordingly.
- **Reinforcement Learning:** The algorithm learns patterns or behaviors in a recursive manner by taking actions to maximize a certain reward, and uses feedback from the environment to learn the best action to take under the right circumstances in order to maximize that reward.

In the coming lecture, our focus will be on unsupervised learning.

**Unsupervised Learning** is a category of techniques that train computers to use a set of unlabeled / unseen data and learn by themselves. The algorithm is provided with a large volume of data and expected to identify hidden patterns. Since there's no prior information / label available for each data instance, there is no defined / correct outcome. The machines only need to determine if there are any patterns in the given data.



Unsupervised Learning is utilised in the following use cases:

1. Clustering
2. Association
3. Dimensionality Reduction

**Clustering:**

In clustering, the key idea is to divide the data into groups such that *each group shares similar properties and each group is as dissimilar as possible to the other groups.*

Let us now try to understand clustering using a simple example:

Let us now try to understand clustering using a simple example:

An organization wants to make a marketing strategy for its new product. They want to segment the customers into different groups in order to target the right audience. They have made segments:

1. based on customer demographics like occupation, age and gender

2. based on income and spending habits

3. based on geographic locations of the customers
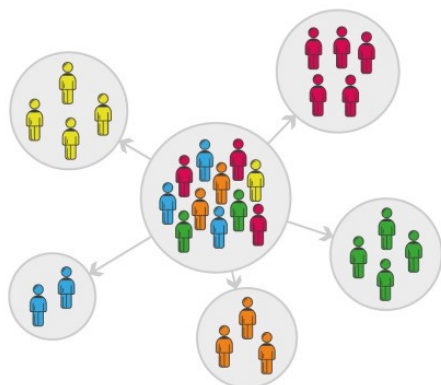


Image source

This is **unsupervised learning.** There was no labeled information about the people, just various features describing their characteristics such as age, gender, income, etc. The algorithm simply found groups of people that may be similar to each other based on these features and organized them into these groups. Due to the absence of labels, we don't know what exactly these groups represent, just that the algorithm found the people in each group similar to the others in that group for some reason, and dissimilar to people from other groups for another reason.

**Association:**

Association learning is a rule-based machine learning and data mining technique that finds important relations between variables or features in a data set. It checks for the dependency of one data item on another data item and maps accordingly so that it can be more profitable.

**Dimensionality Reduction:**

Dimensionality Reduction is the transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its intrinsic dimension.