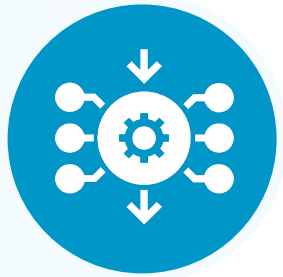
Learning from a dataset without labels: unsupervised learning

AI | BUSINESS SCHOOL



How does unsupervised learning work?

- **Unsupervised learning** uses ML algorithms to discover hidden patterns in the data without **human supervision.** It is an effective tool for data analysis to find similarities and contrasts.
- For example, grouping customers according to the types of products they look up & the amount of time they spend online, to analyze which group spends more.

One of the most common unsupervised learning techniques: clustering

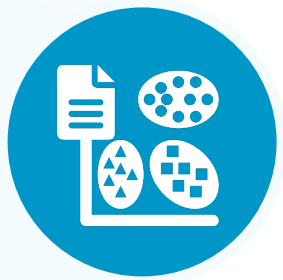
AI BUSINESS SCHOOL

Clustering is the process of grouping different data points based on similarities in their features, instead of labels.

- For example, fruits with similar shape, size and color, are part of the **same cluster** & hence they are the same type of fruit, say melons or blueberries.

To make computers perform clustering, we use K-Means algorithm

AI BUSINESS SCHOOL



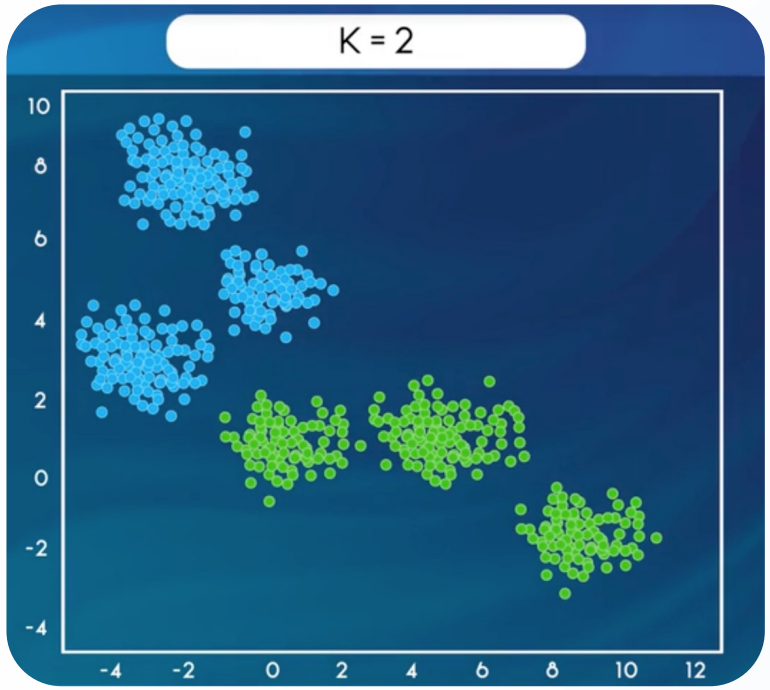
What is K-means algorithm?

K-means is a clustering approach in which data points are divided into a “K” number of groups. K-means algorithm groups data points that are close to a specific center together.

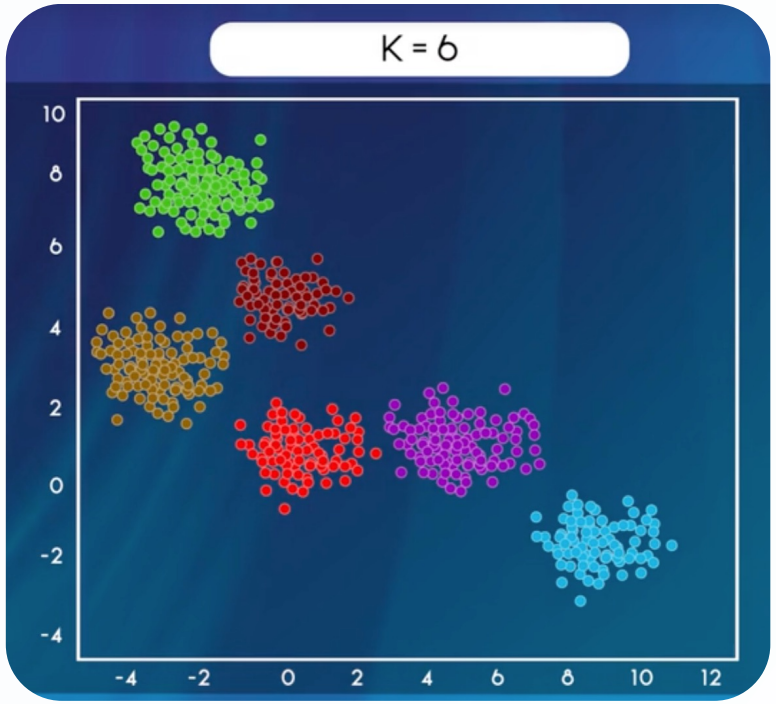
K-means algorithm

AI BUSINESS SCHOOL

A small number of K = larger clusters



A big number of K = smaller clusters

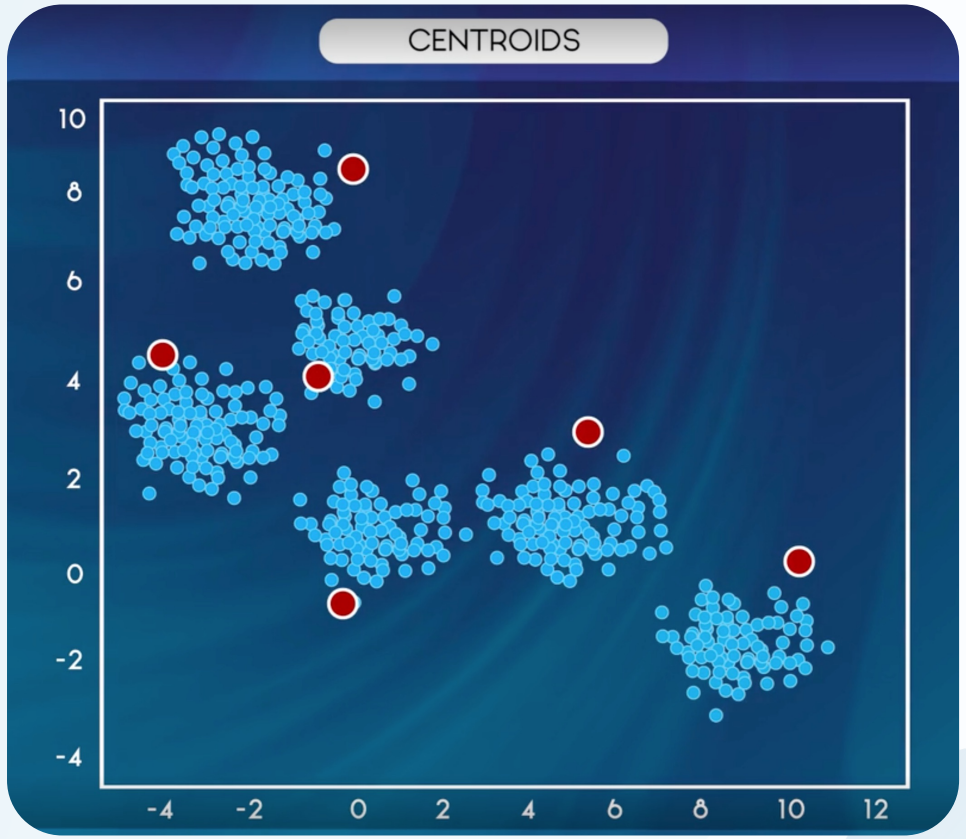


Optimizing the positions of the center points in K-means: centeroids

AI BUSINESS SCHOOL

First, a group of randomly selected cluster center points, **centeroids**, are used as the beginning points for every cluster.

Then iterative calculations are performed to optimize the positions of the centroids.



Performance evaluation metrics for clusters: silhouette coefficient

AI | BUSINESS SCHOOL

What is silhouette coefficient?

- **The silhouette coefficient** or **silhouette score** is a metric used to calculate the goodness of a clustering technique.
- It compares the average distance of a point to other points in the same cluster with the average distance of the same point to the points in the nearest cluster.

$$s = \frac{b-a}{max\,(a,b)}$$

a: The mean distance between a sample and all other points in the same cluster.

b: The mean distance between a sample and all other points in the next nearest cluster.