Clustering refers to a broad set of techniques for finding subgroups in a dataset. Data points are partitioned into distinct **clusters** such that they are

- Similar to each other (within the same cluster)
- Different to data points in other clusters

The criteria for similarity and dissimilarity are an ambiguous, *domain-specific* consideration.

Partitional Clustering

Partition the data points into k distinct groups such that each object belongs to exactly *one* cluster. Achieved through K-means, or the more mature K-means++ algorithm.

Cost Function

Cost functions use a distance function to measure the variance between solutions.

- Output is *minimized* for good solutions
- Bigger output for poor solutions

$$\sum_{i}^{k} \sum_{x \in C_i} d(x, \mu_i)^2$$

Hierarchical Clustering

A set of nested clusters organized into a *tree-based* representation called a **dendrogram**. There are two types of hierarchical clustering:

- Agglomerative (more common)
- Divisive

    1. Start with each point in the same cluster
    2. Split the cluster at each step until every point is in its own cluster

At every step of a hierarchical clustering, a record is kept of which clusters were *merged* to produce the dendrogram.

- The resulting dendrogram can be "cut" at any threshold to produce any number of clusters
- Finding the threshold with which to cut the dendrogram requires exploration and tuning

Density-Based Clustering

Data points are clustered together based on their local density. Given a fixed radius ϵ around a point, that area is considered **dense** if there are a minimum of some `min_pts` points in that area.

To distinguish between points at the core of a dense region and those at the borders, we define the following:

Point type	Description
Core	if its ϵ-neighborhood contains at least `min_pts` points
Border	if it is in the ϵ-neighborhood of a core point
Noise	if it is neither a core nor border point

Density-based clustering can be achieved through the DBSCAN algorithm.

Soft Clustering

Each point is assigned to every cluster with a certain probability.