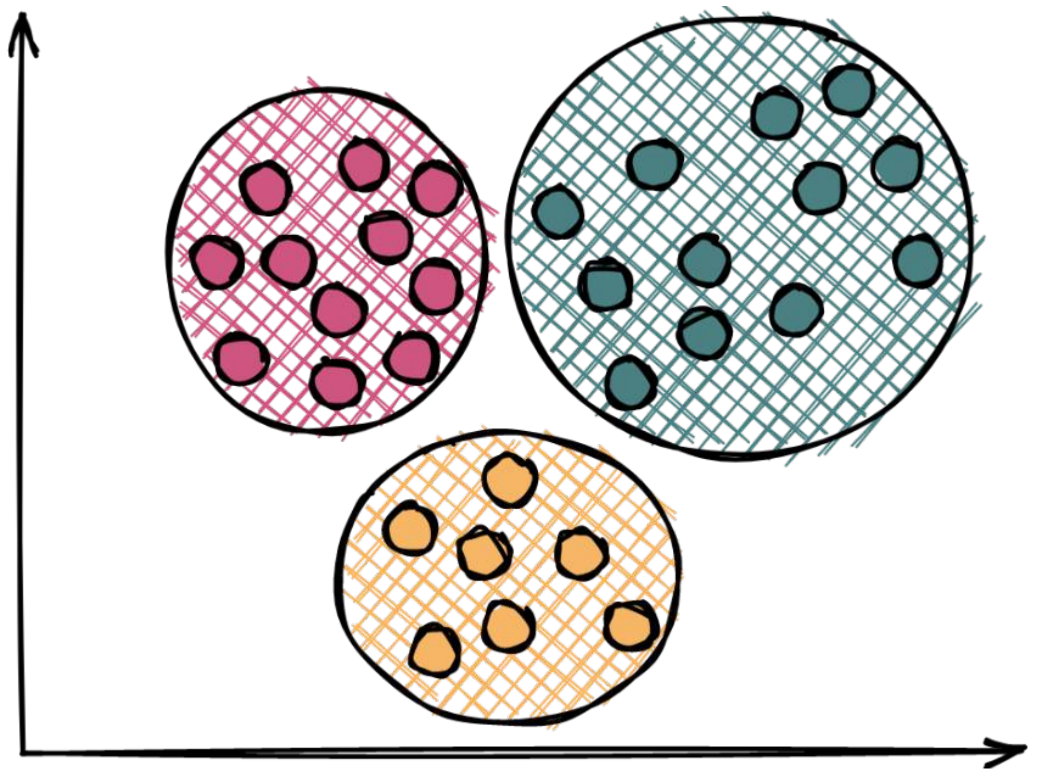
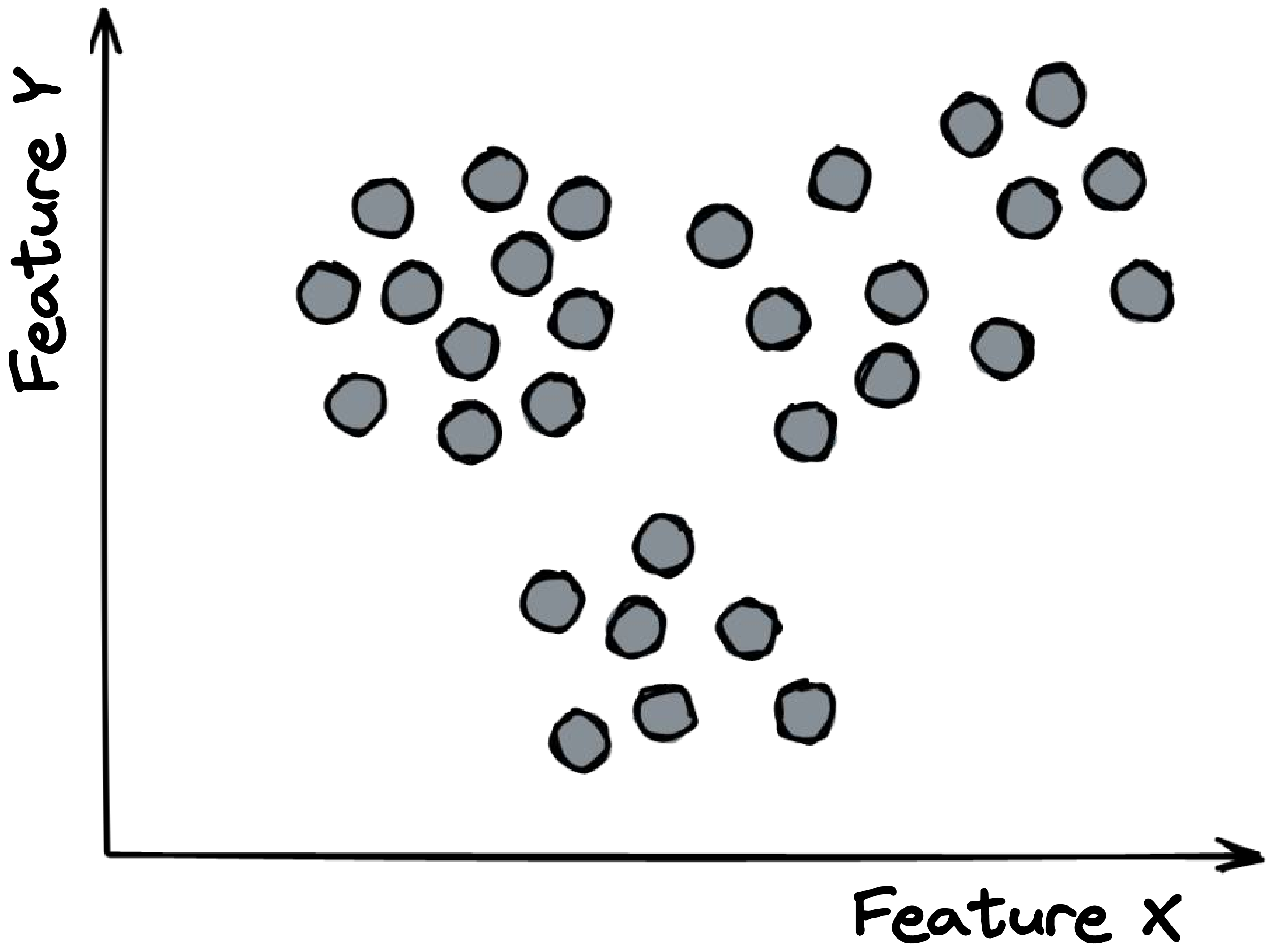
Clustering



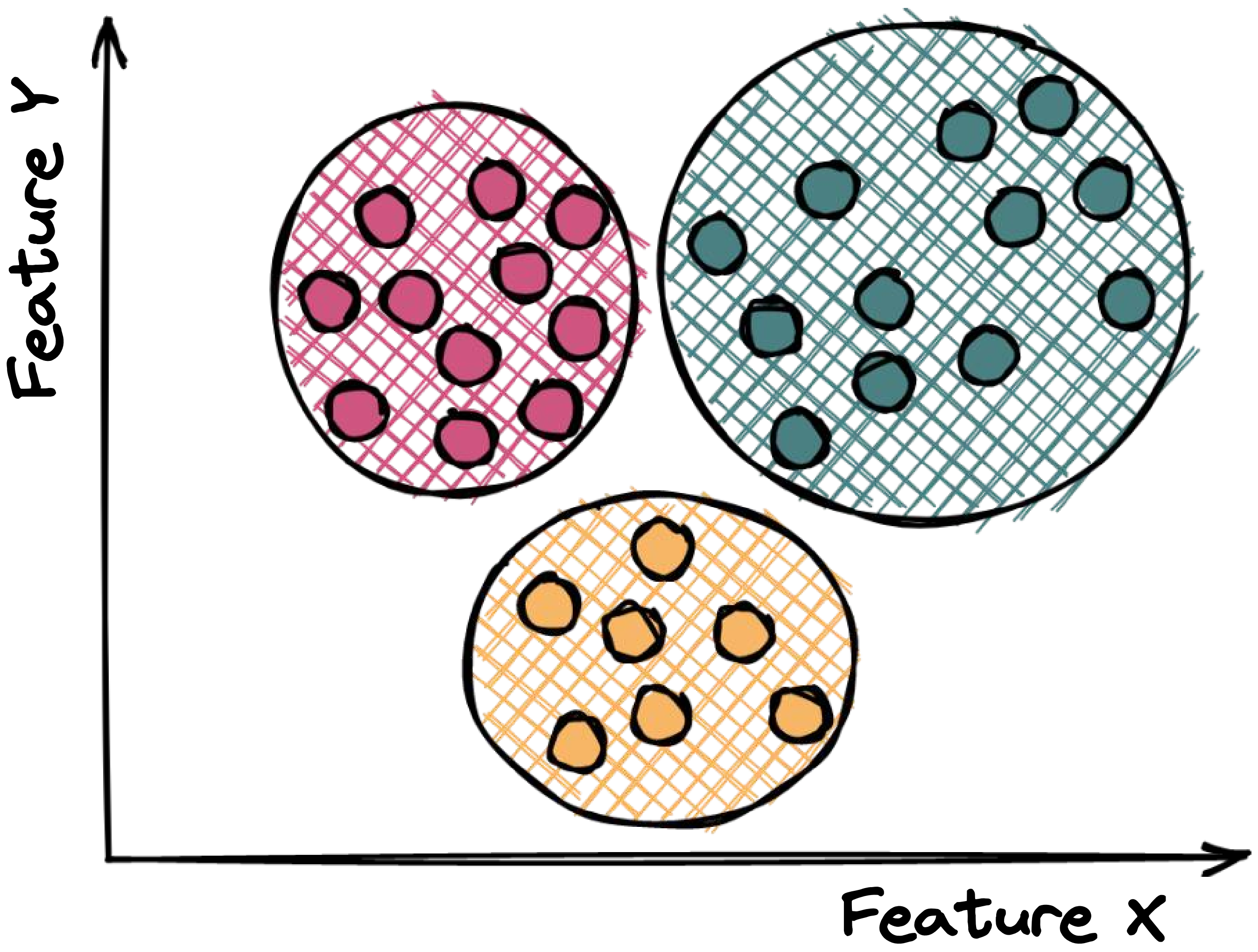
Additional materials

Middlesex University Dubai; CST4050;
Instructor: Ivan Reznikov

Unlabeled Data

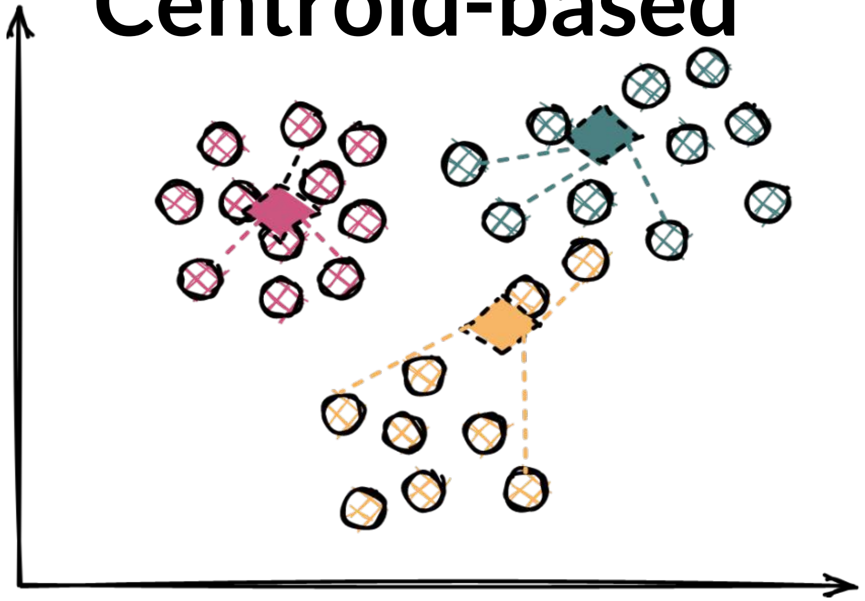


Labeled Clusters

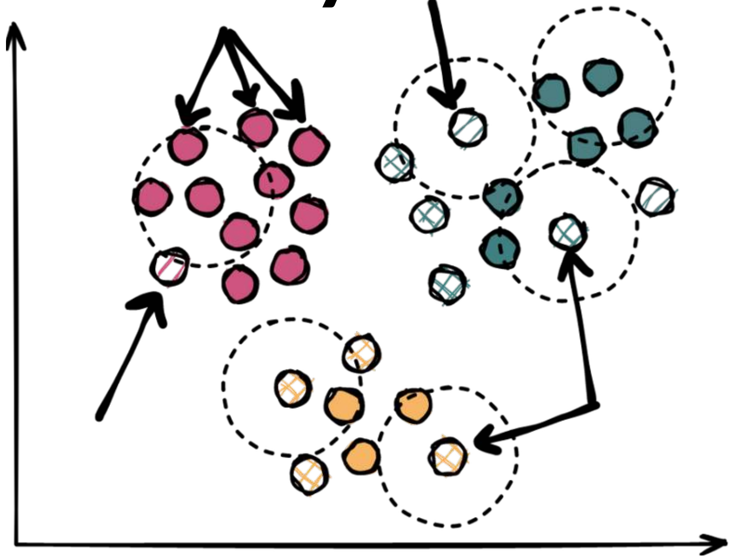


Types of clustering:

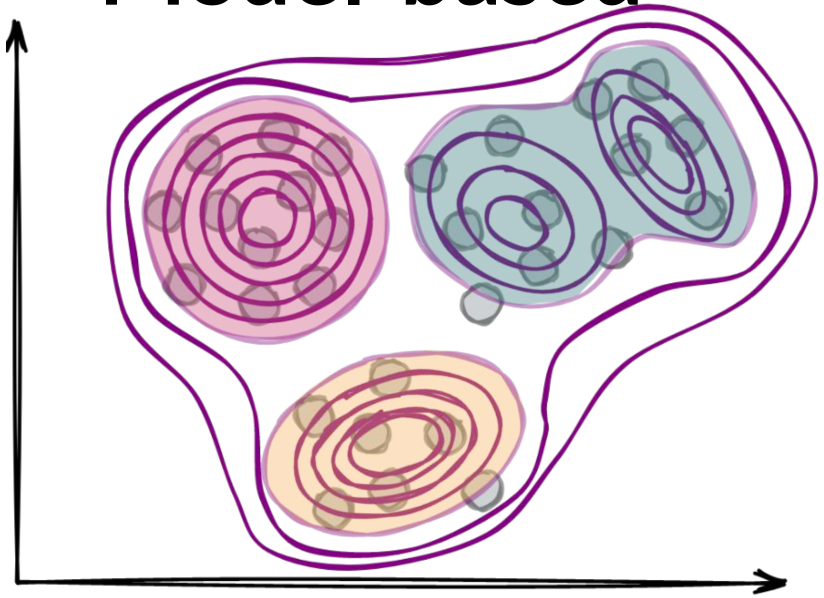
Centroid-based



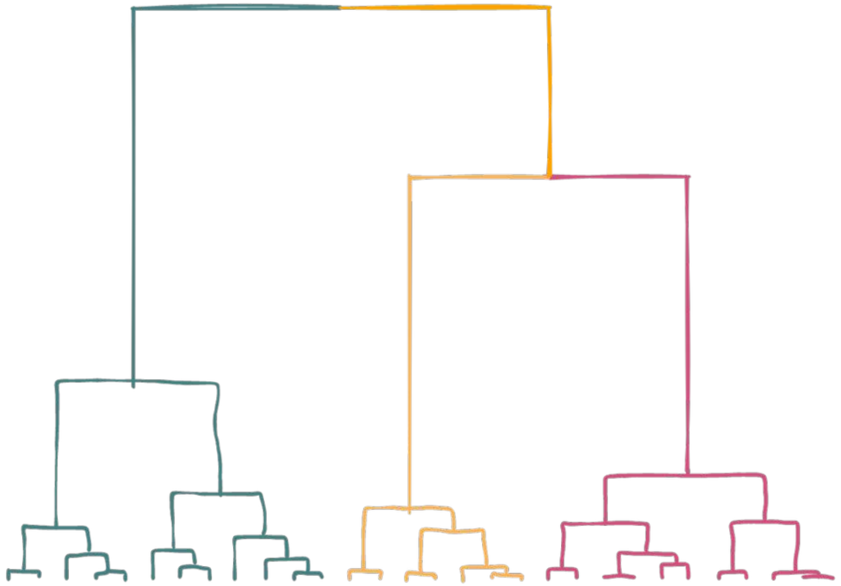
Density-based



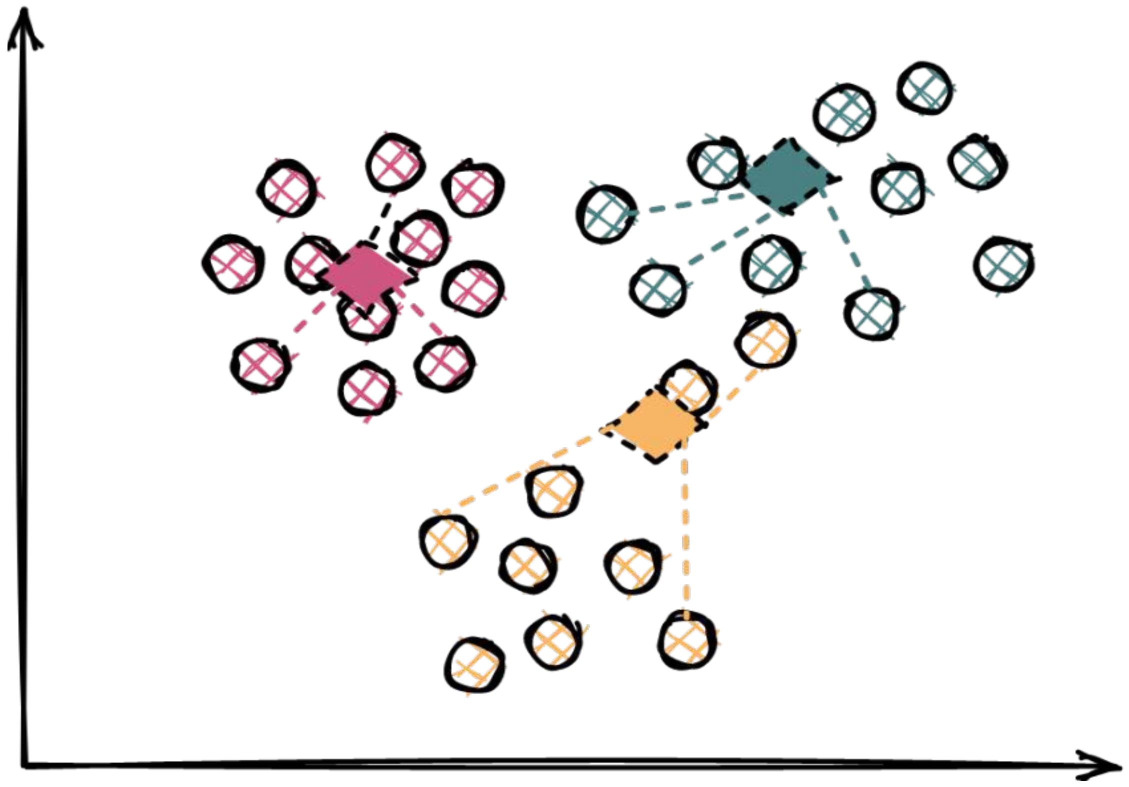
Model-based



Distance-based



Centroid-based clustering

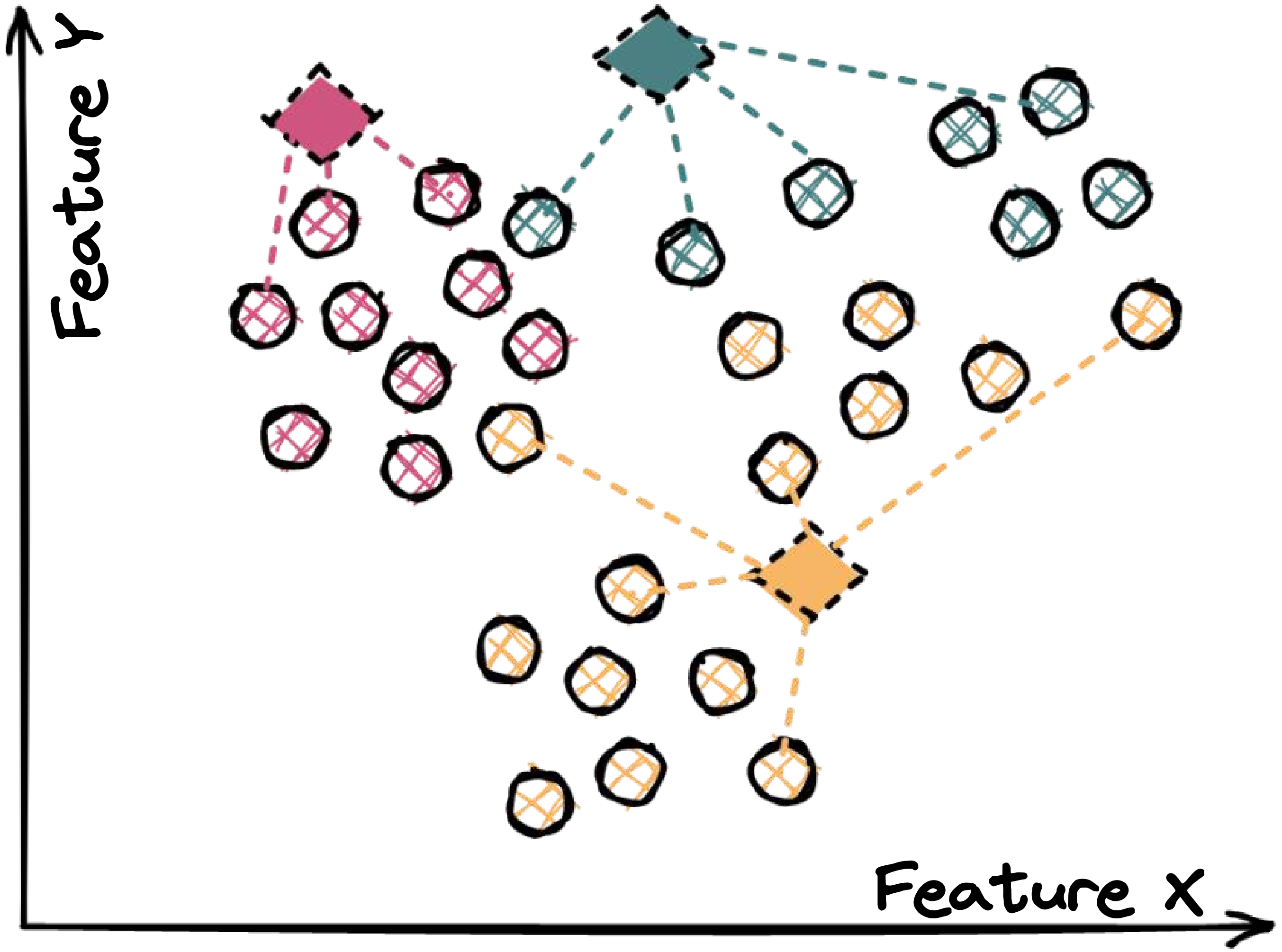


Main idea:
Minimize the squared distances of all points in the cluster to cluster centroids.

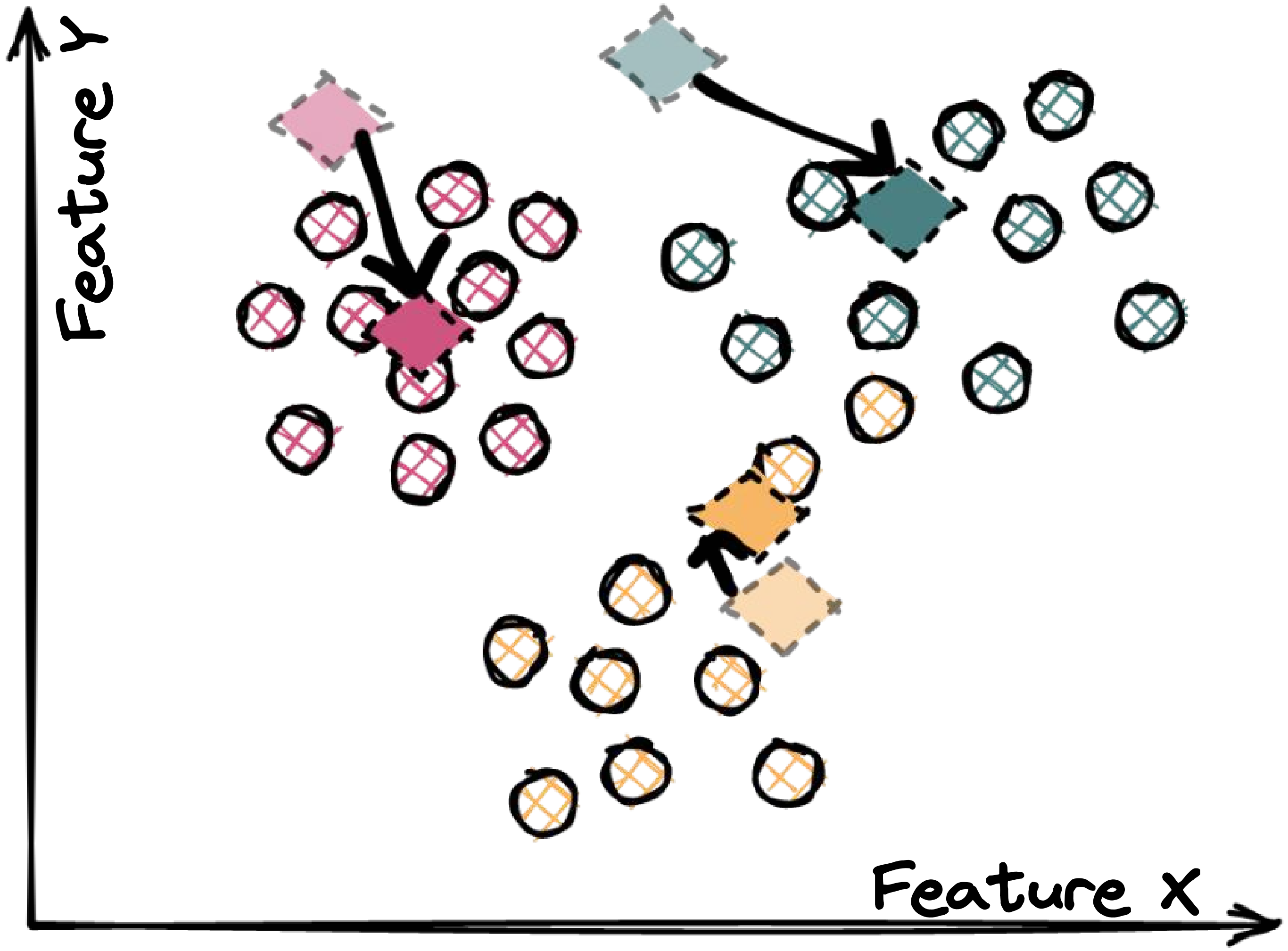
Most used method:
k-Means

Centroid-based clustering

Step1a. Calculate distances



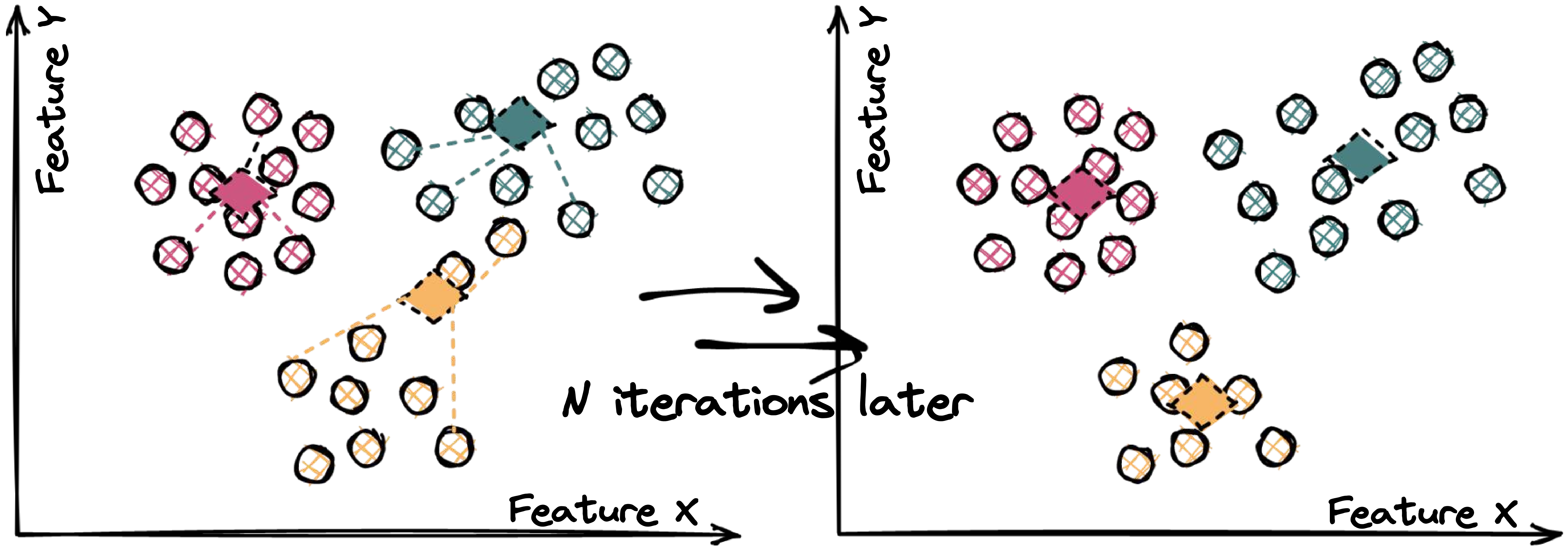
Step1b. Relocate centroids



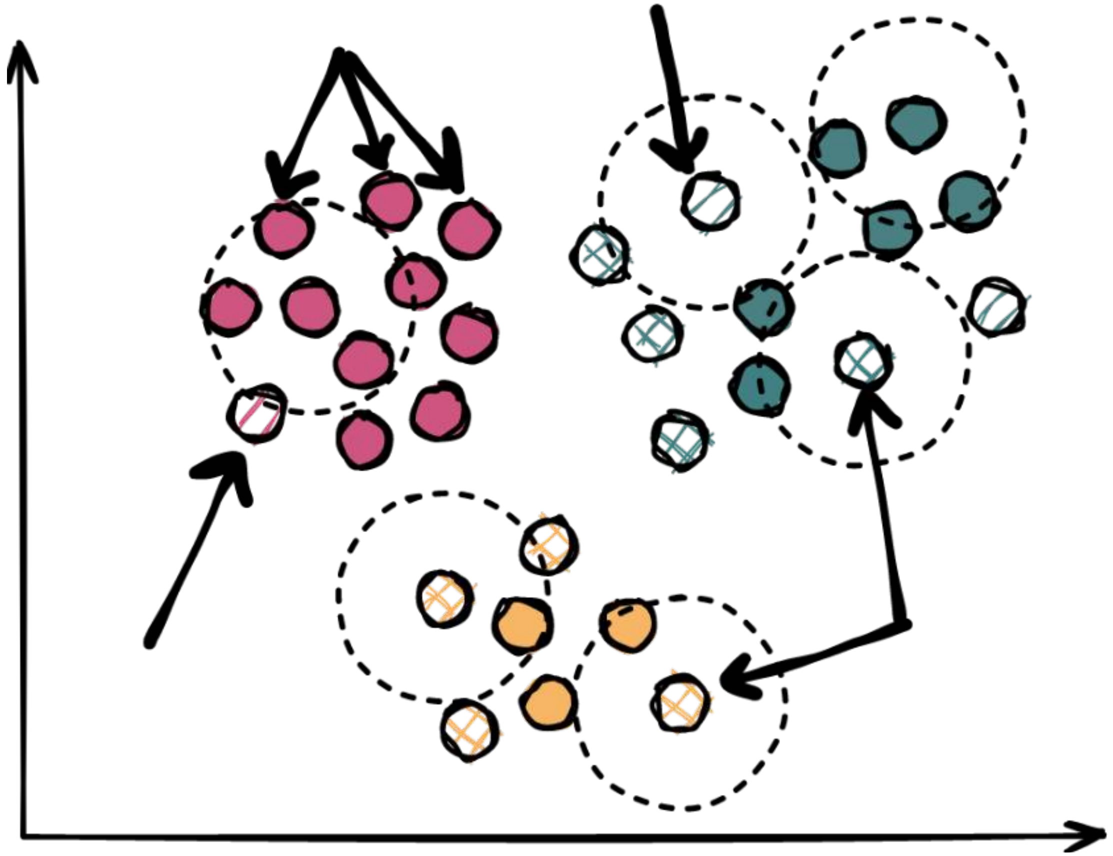
Centroid-based clustering

Step2a. Calculate distances

StepNb. Relocate centroids



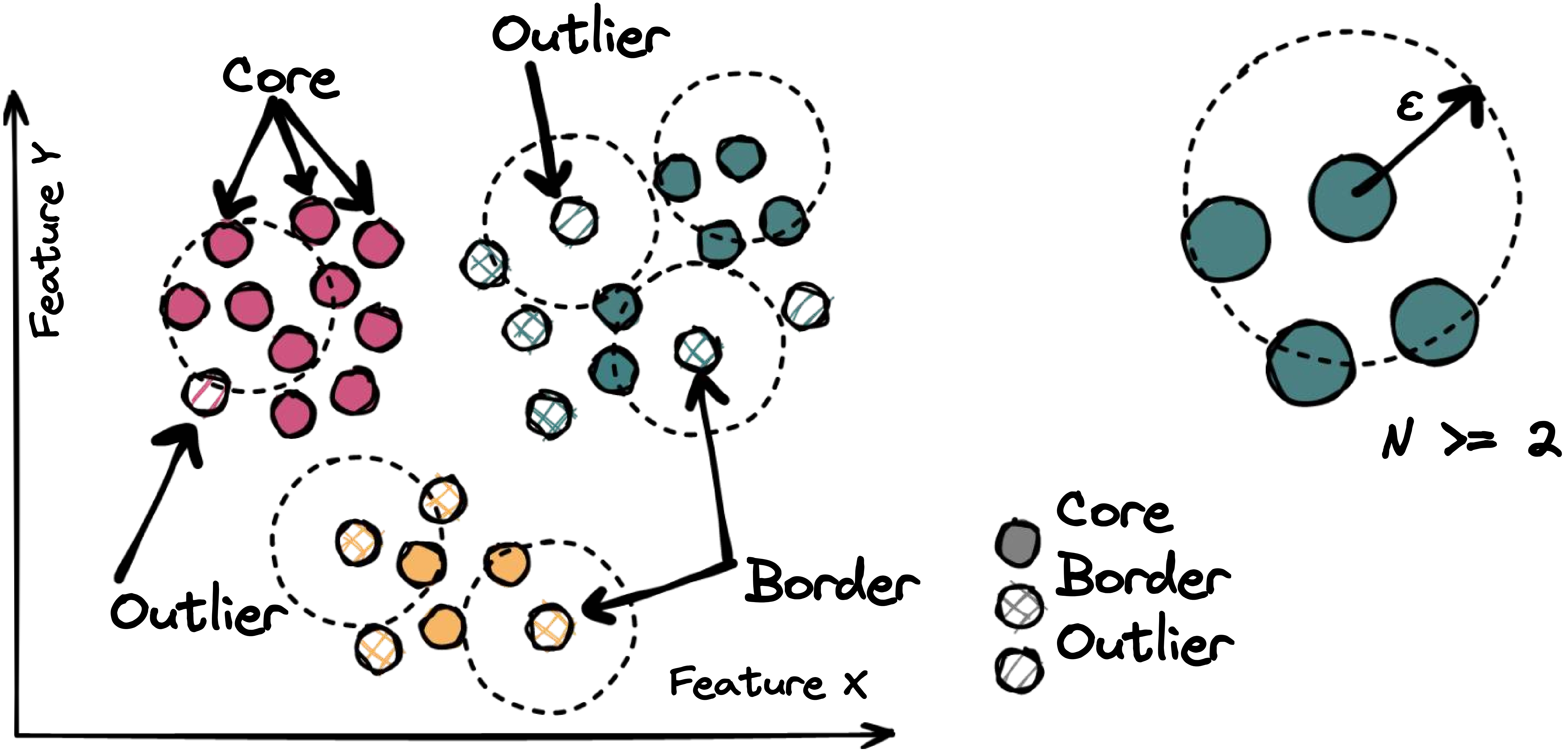
Density-based clustering



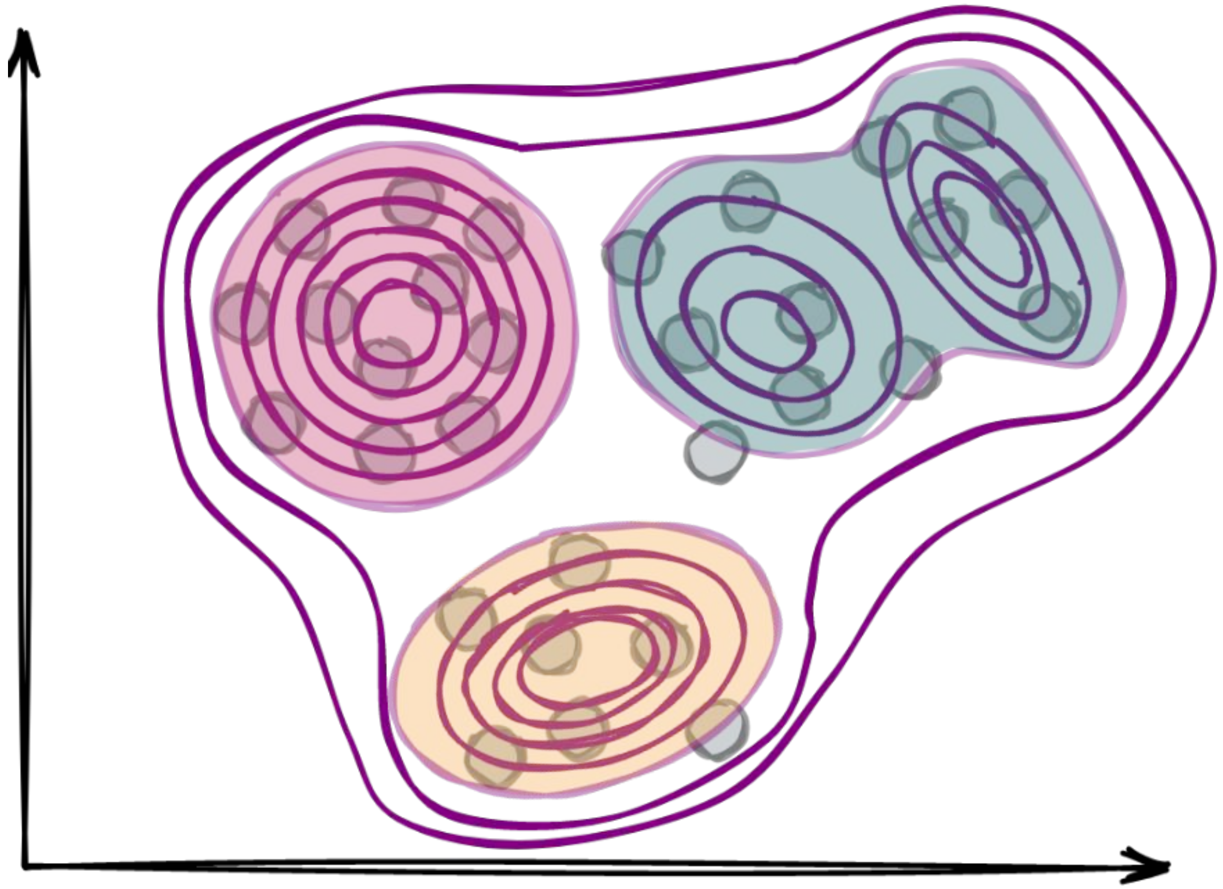
Main idea:
Clusters are defined based on identifying areas of higher density.

Most used method:
DBSCAN

Density-based clustering



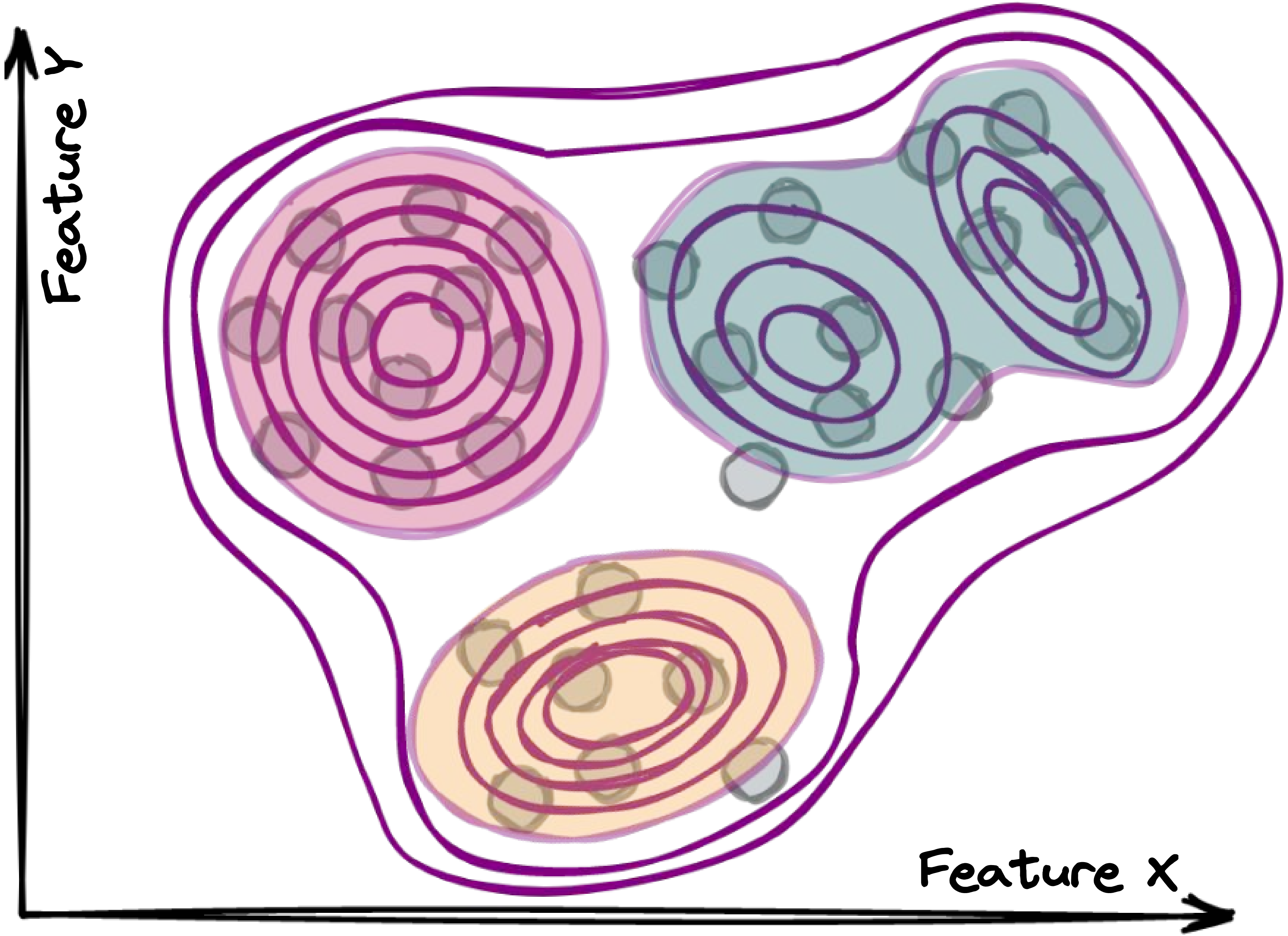
Model-based clustering



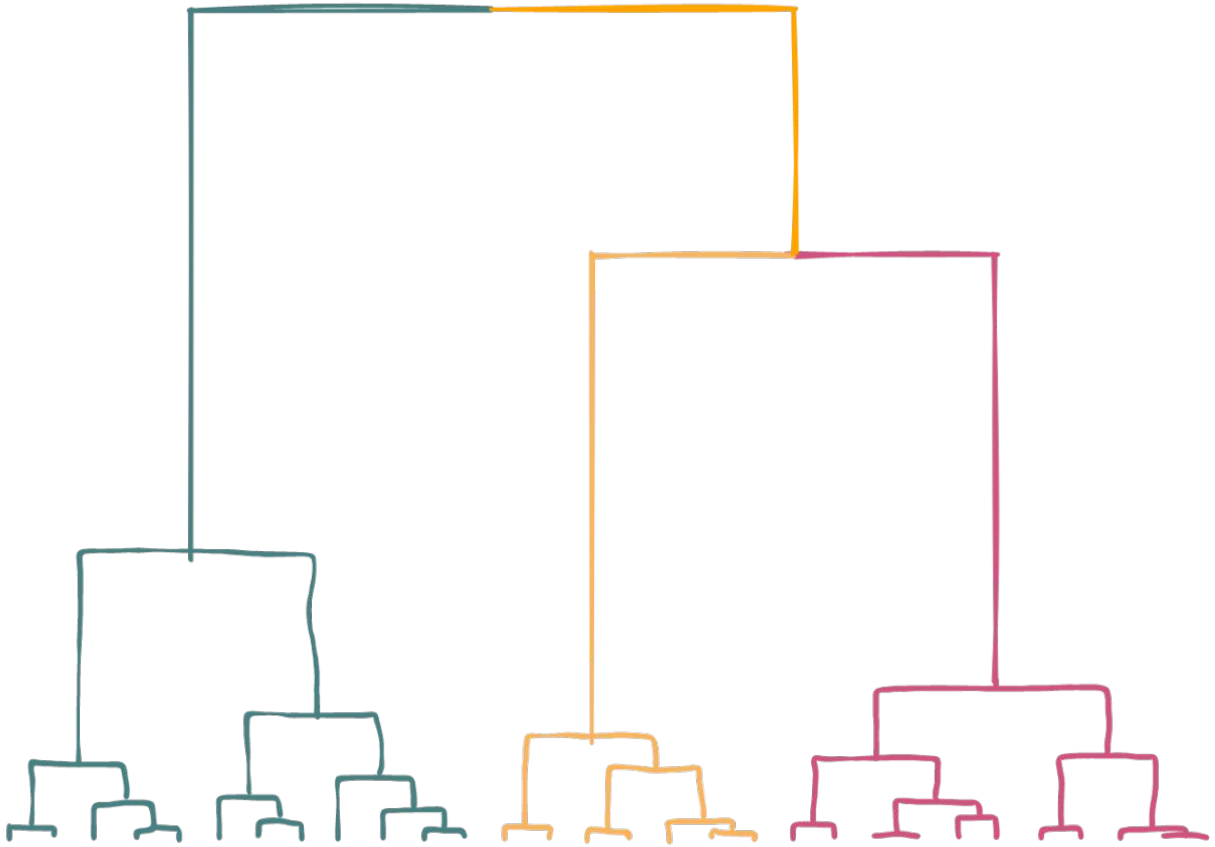
Main idea:
Clusters are defined based on how likely the objects included are likely to belong to the same distribution.

Most used method:
GMM – Gaussian Mixture Models

Types of clustering: Model-based



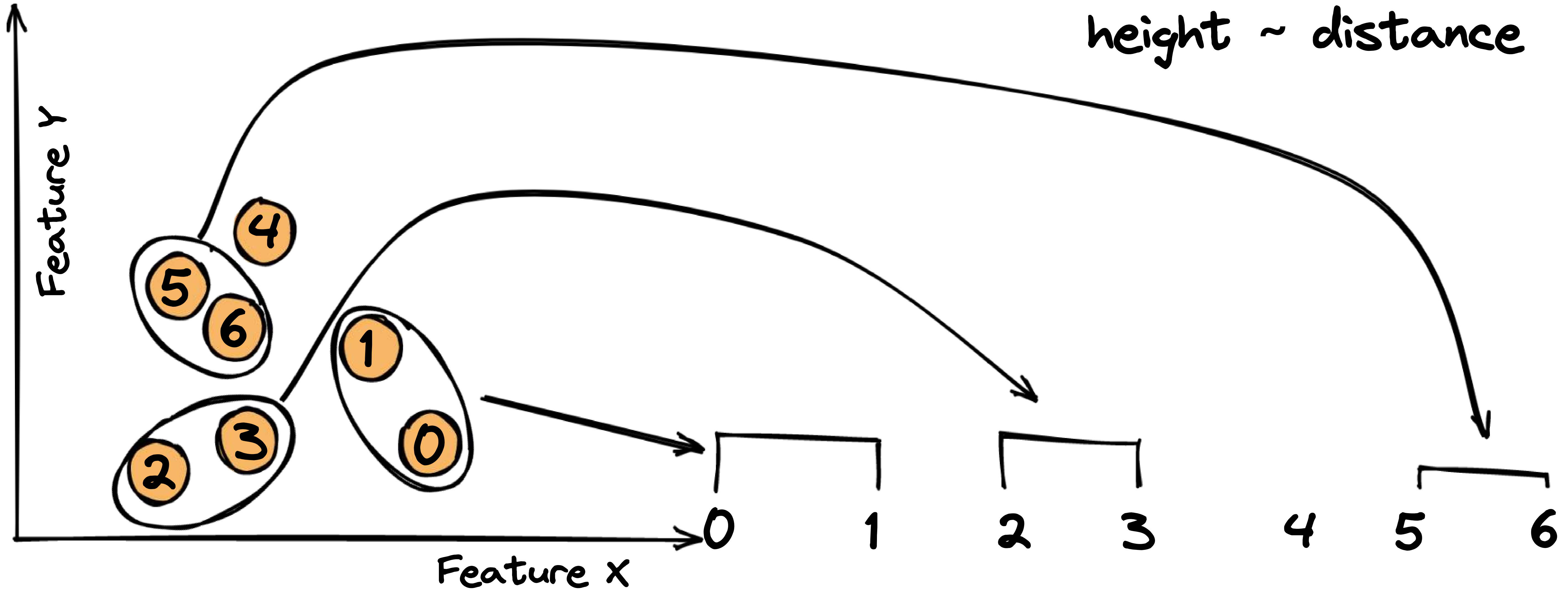
Distance-based clustering



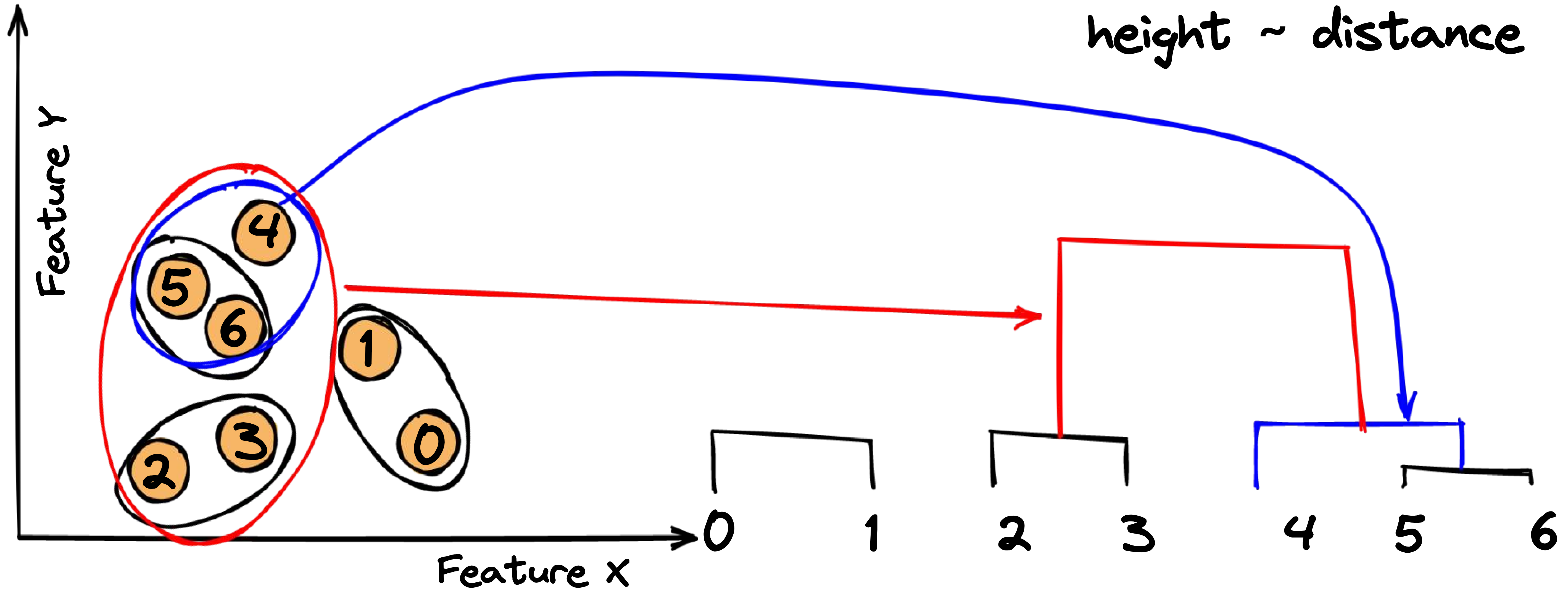
Main idea:
Clusters are developed based on distance between objects, as closer means more related.

Most used method:
AHC – Agglomerative hierarchical clustering

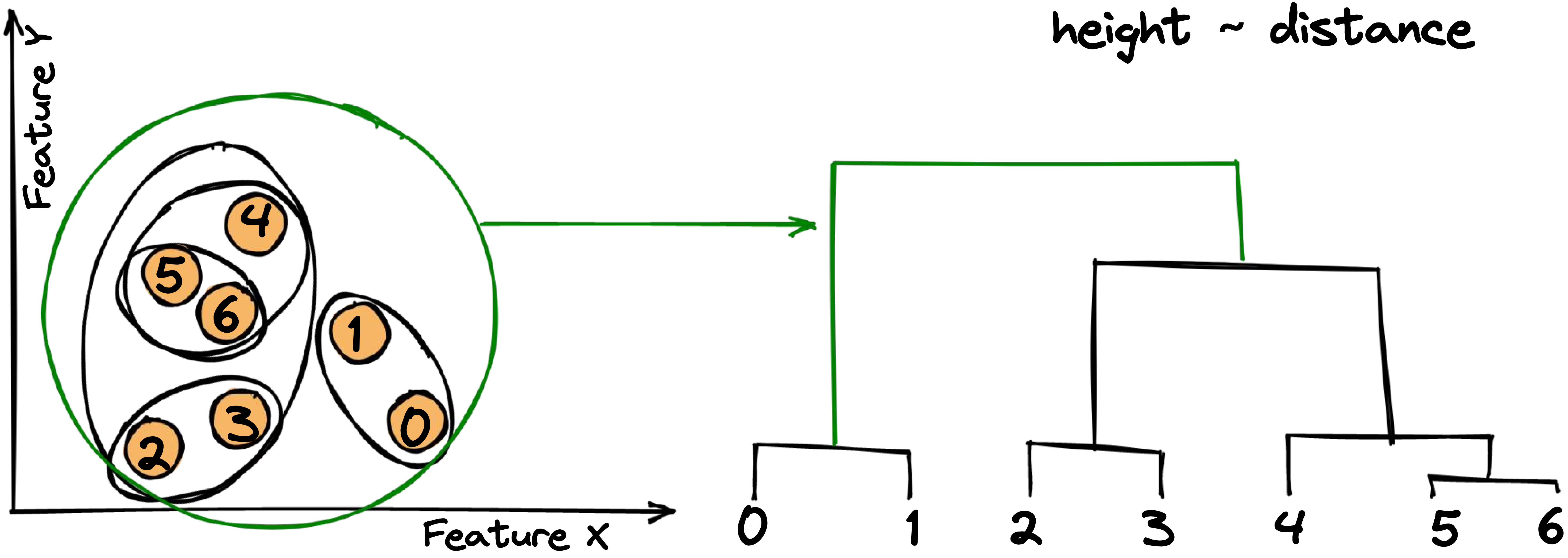
Distance-based clustering



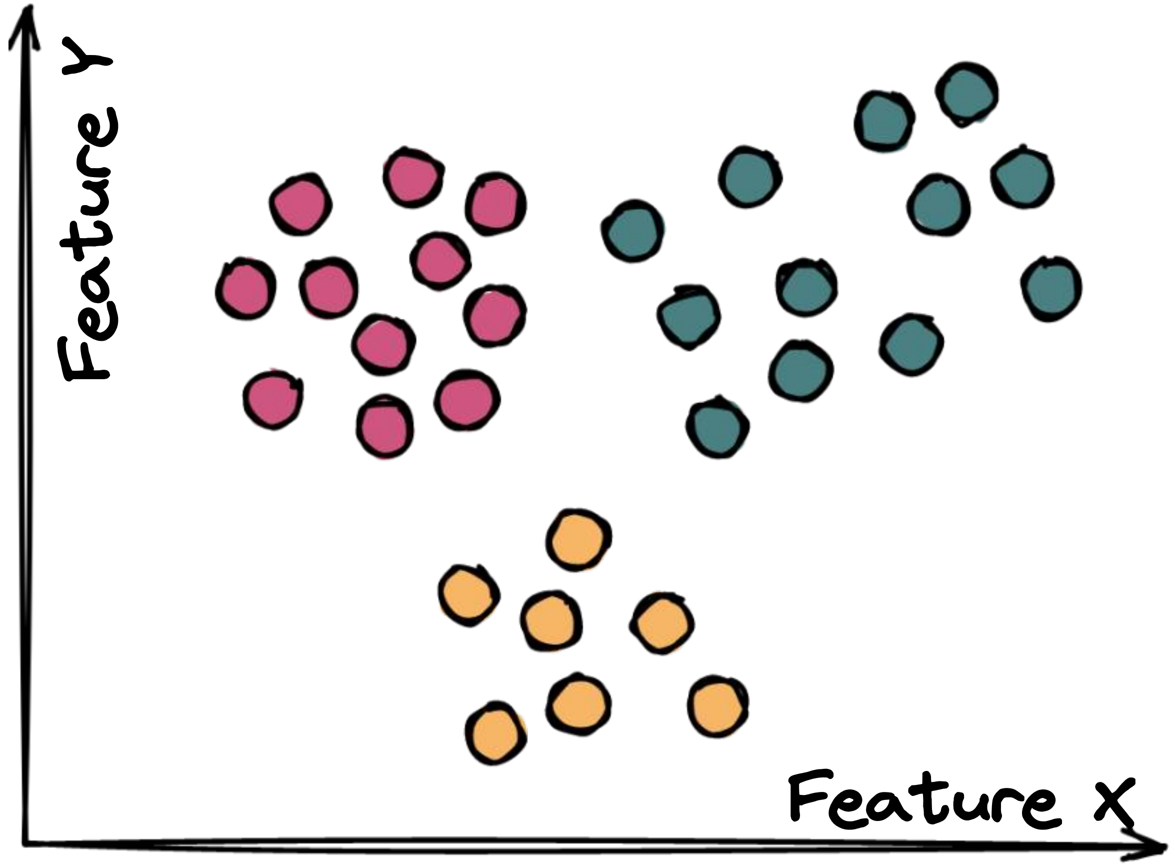
Distance-based clustering

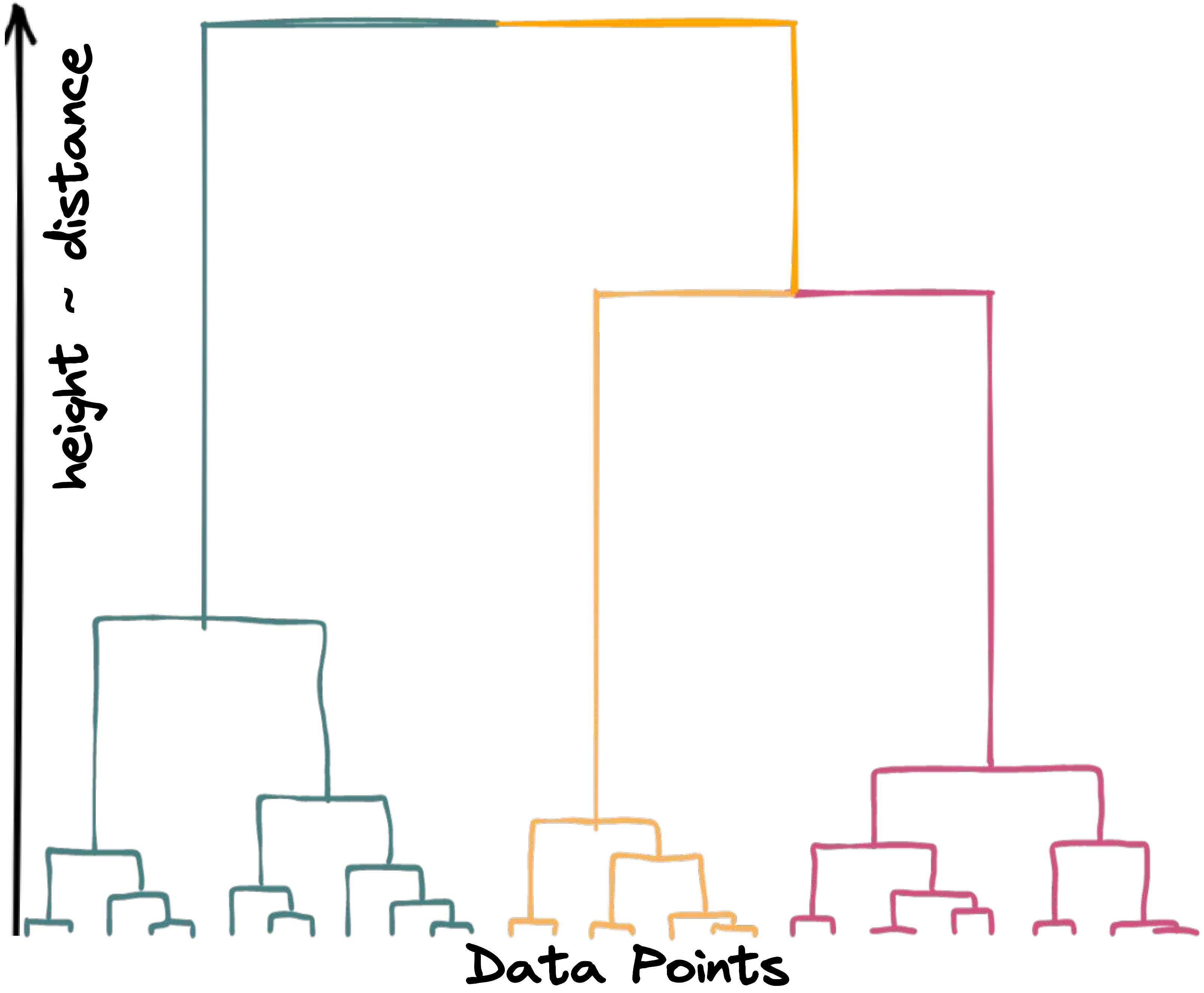


Distance-based clustering



Distance-based clustering





Advantages

	Centroid	Density	Model	Distance
performance and scaling	●			
can return probability of points belonging to cluster K	●			
overlapping clusters can be identified as several		●		
can work with weird-shaped clusters		●		
can find clusters surrounded by other clusters			●	
can provide object ordering				●
can return dendrogram				●

Disadvantages

	Centroid	Density	Model	Distance
required K number of cluster	🔴		🔴	🟠
sensitive to chosen inputs	🔴	🔴		
scaling problems with high dimensions	🔴			
strongly dependent on random	🔴			
varying sizes and densities problems	🔴	🔴		
exposed to noise and outliers	🔴		🟠	🔴
fails if sparse data		🔴		
requires a large amount of data			🔴	
needs to know the type of distribution			🔴	
can't regroup clusters if done wrong				🔴