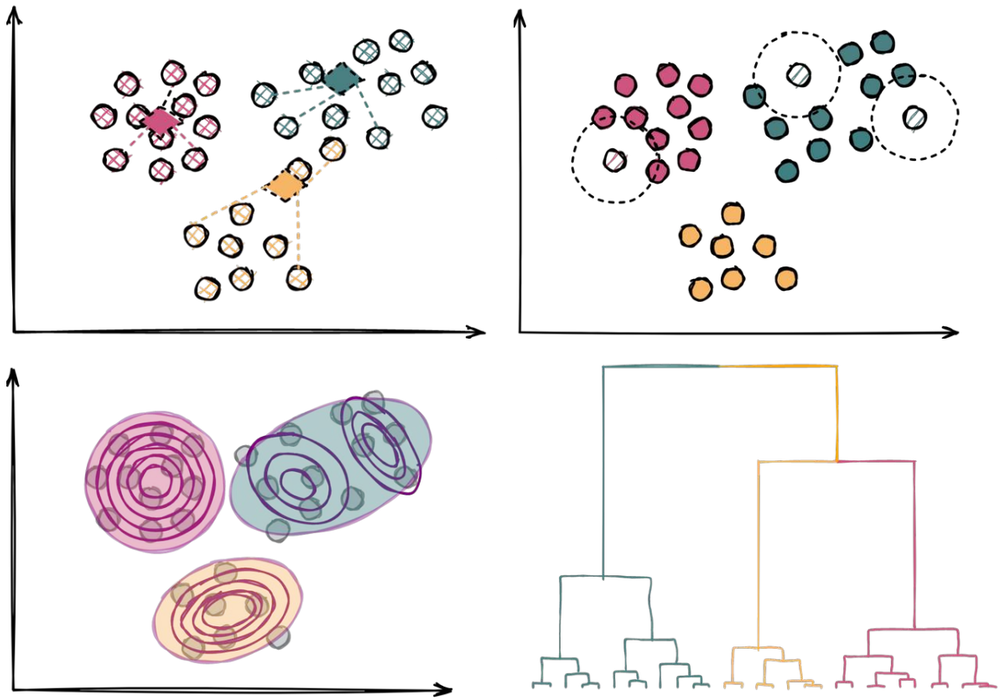


Clustering

Week 20

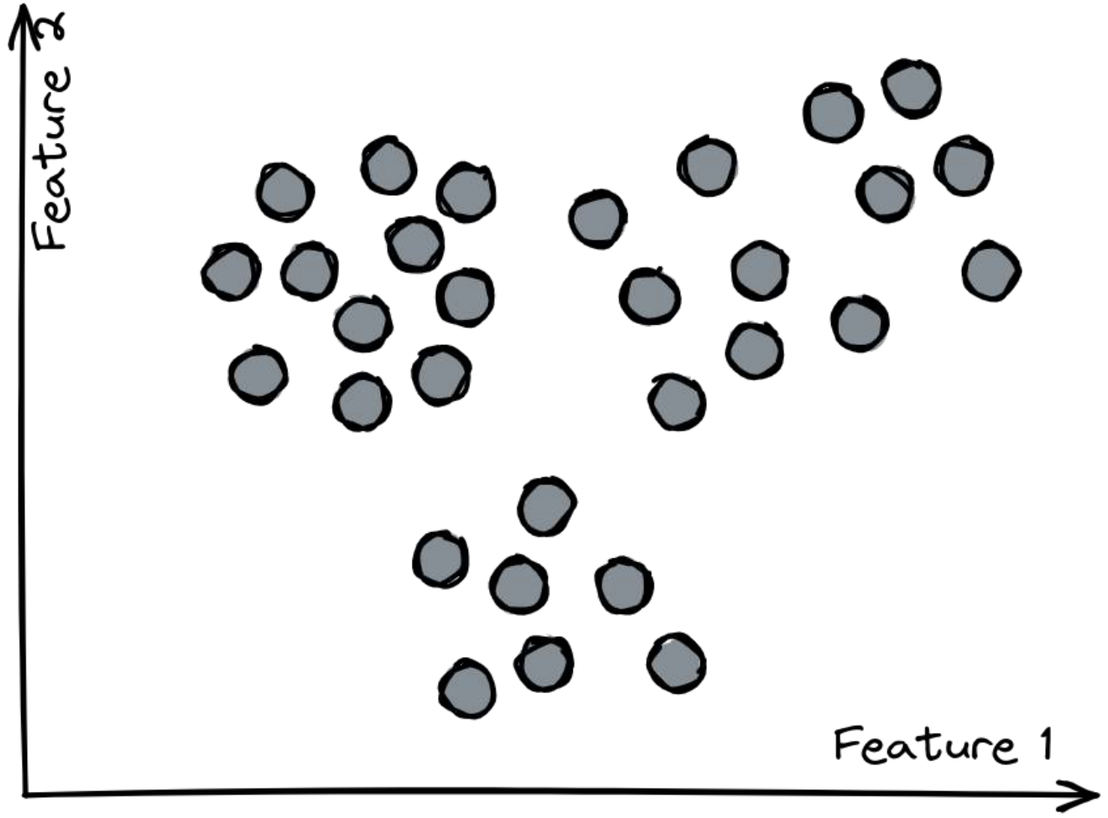
Middlesex University Dubai; CST4050 Fall21;
Instructor: Dr. Ivan Reznikov

Plan

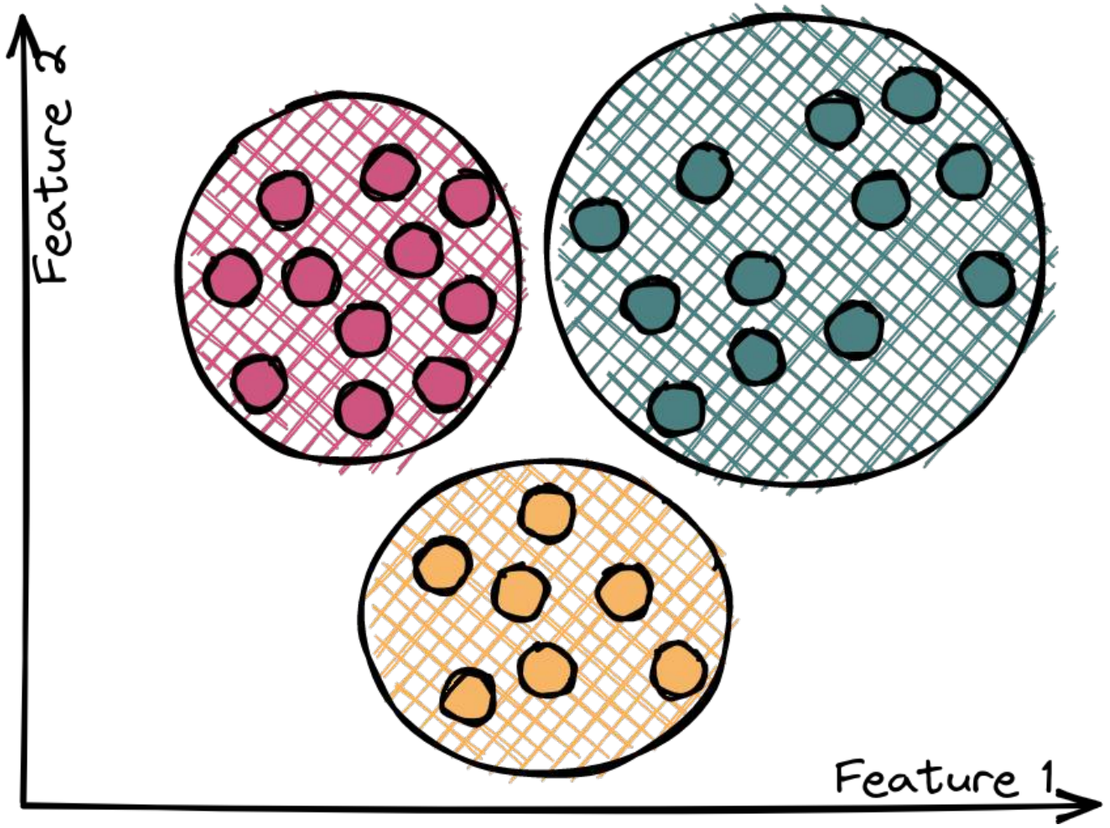
- Clustering and their types
- Centroid-based clustering and K-Means
- Density-based clustering and DBSCAN
- Model-based clustering and GMM
- Distance-based clustering and Hierarchical clustering
- Advantages and Disadvantages

Unsupervised Learning

Unlabeled Data

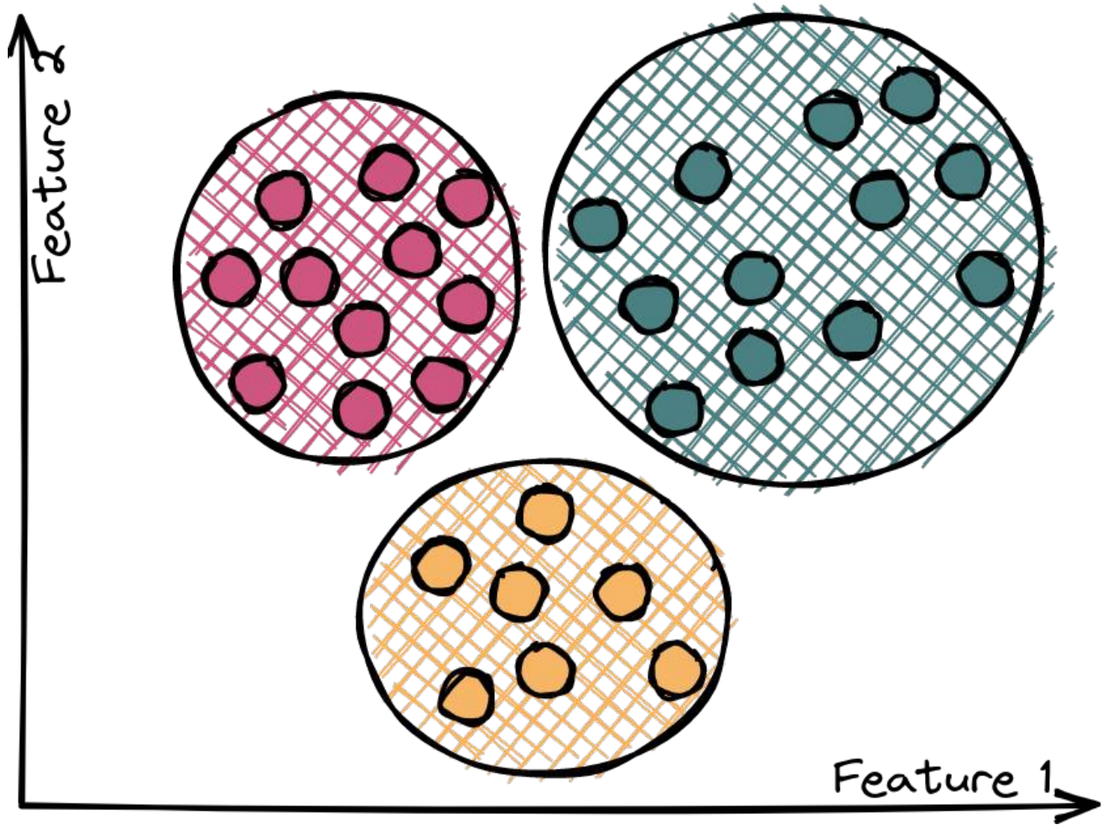


Labeled Clusters

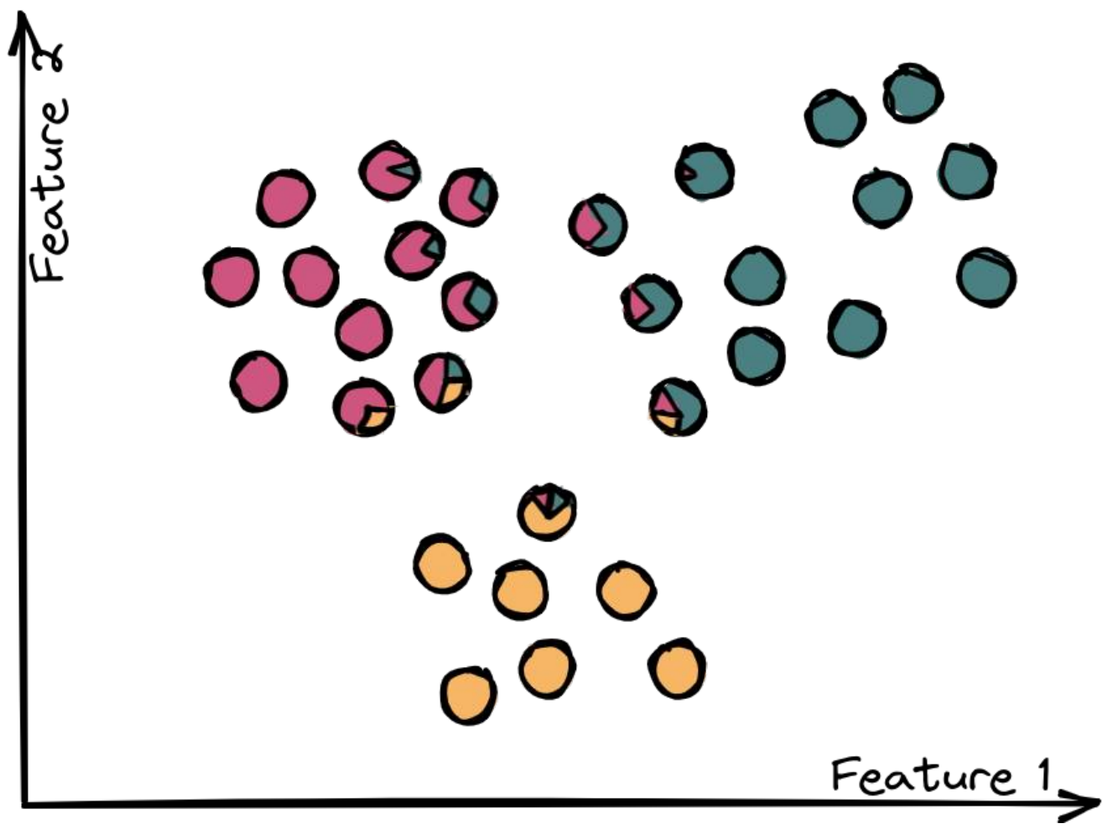


Types of clustering

Hard Clustering

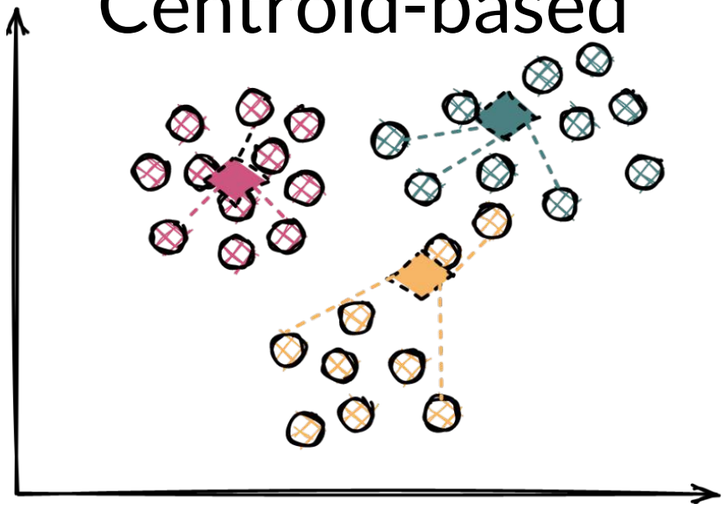


Soft Clustering

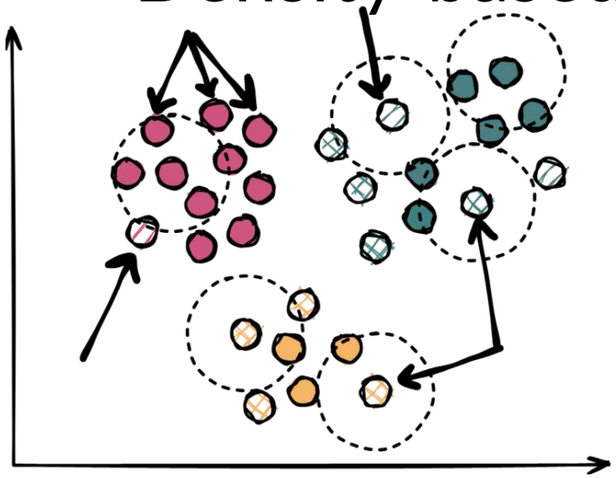


Types of hard clustering:

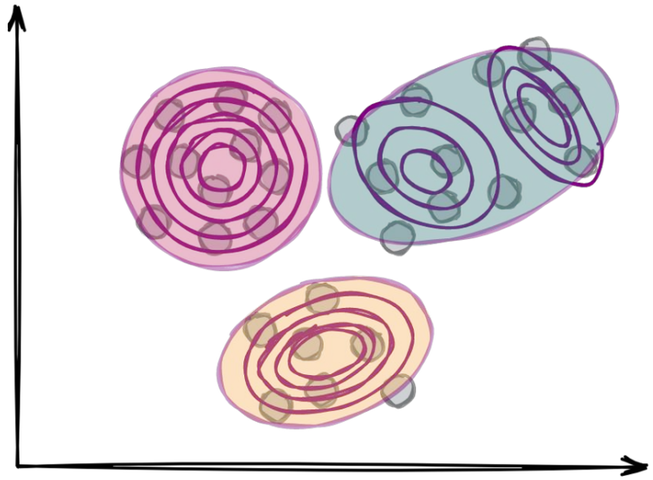
Centroid-based



Density-based

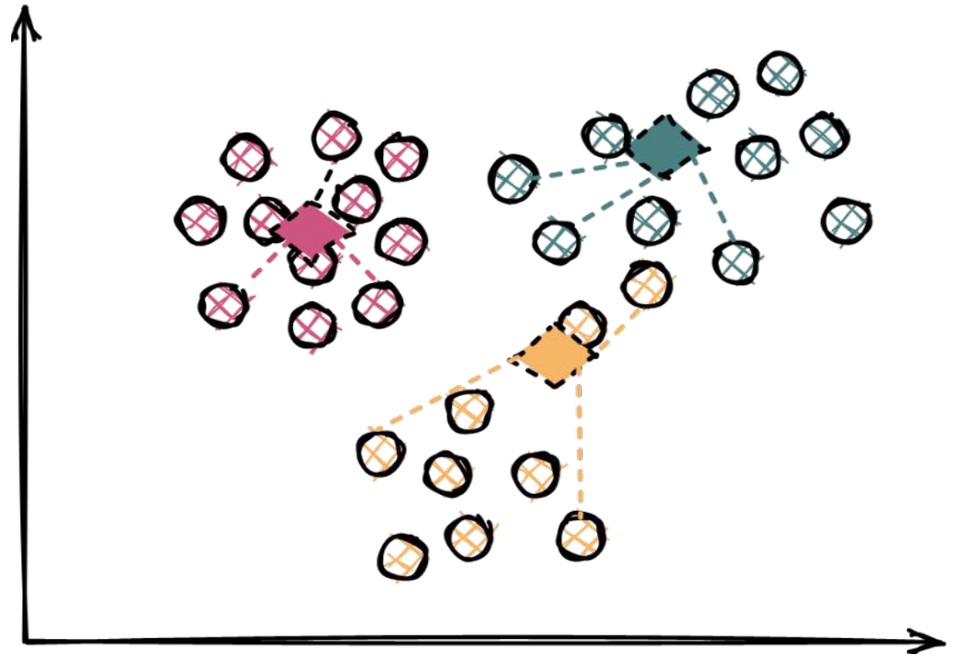


Model-based



Distance-based

Centroid-based clustering

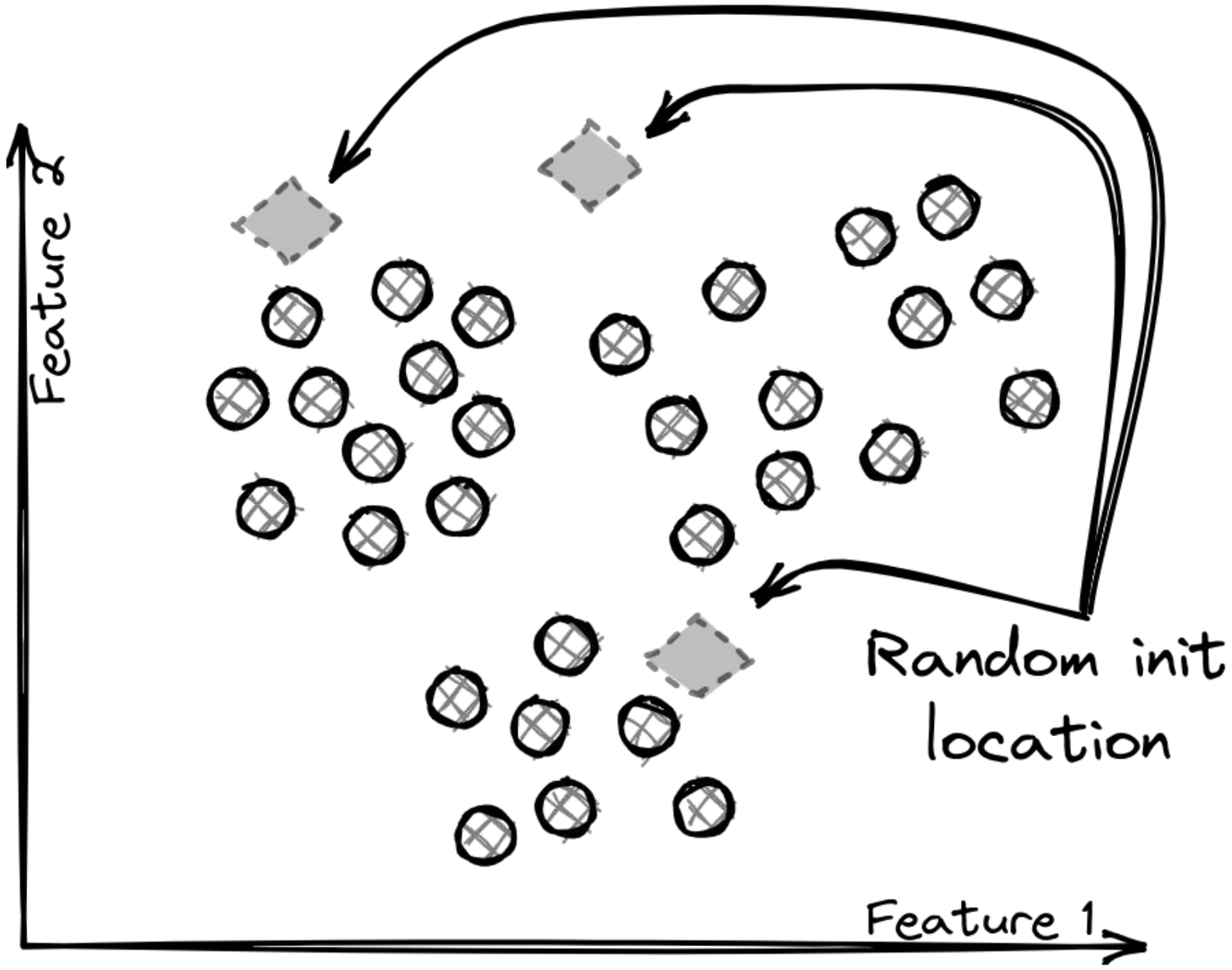


Main idea:
Minimize the squared distances of all points in the cluster to cluster centroids.

Most used method:
k-Means

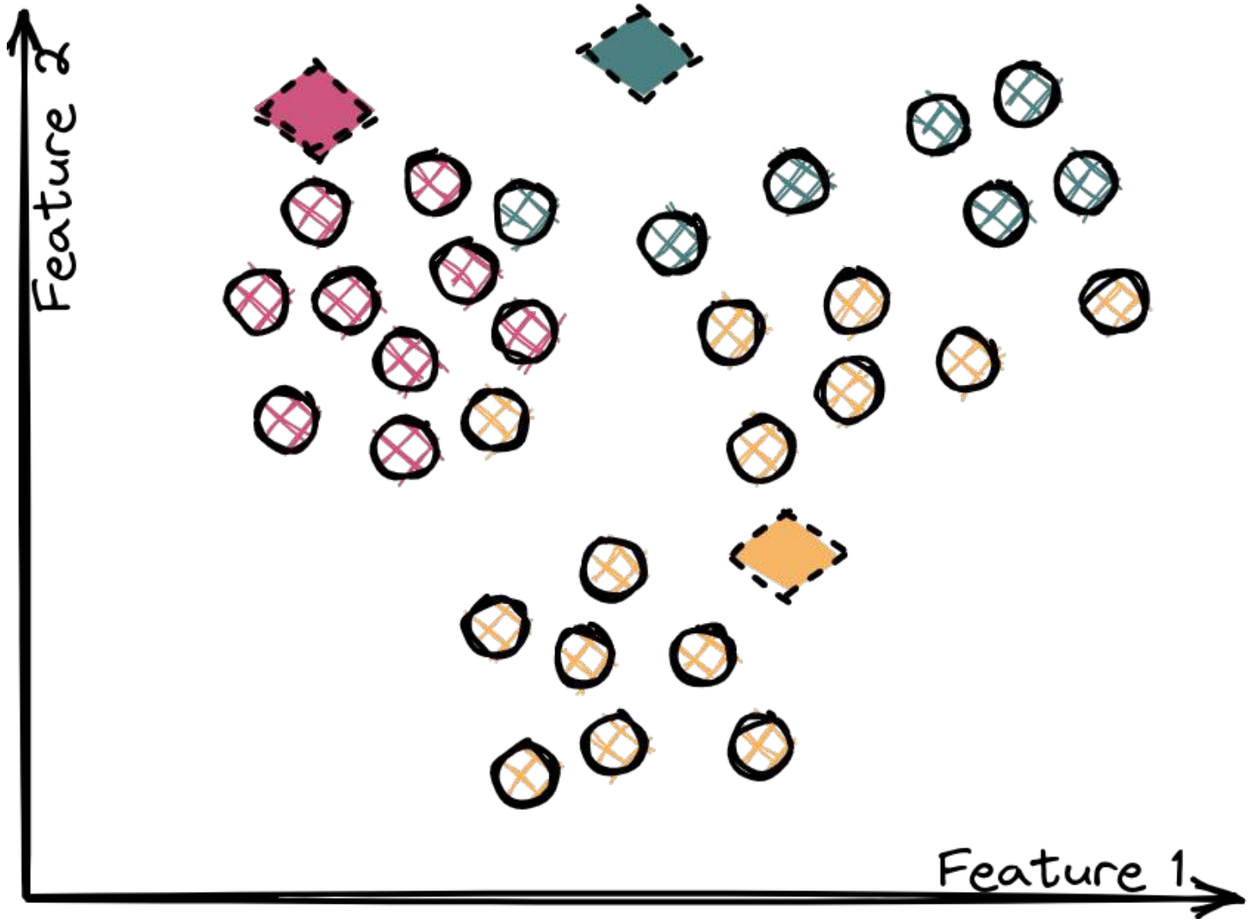
K-Means: Step 0a

Step 0a. Randomly set cluster centroids



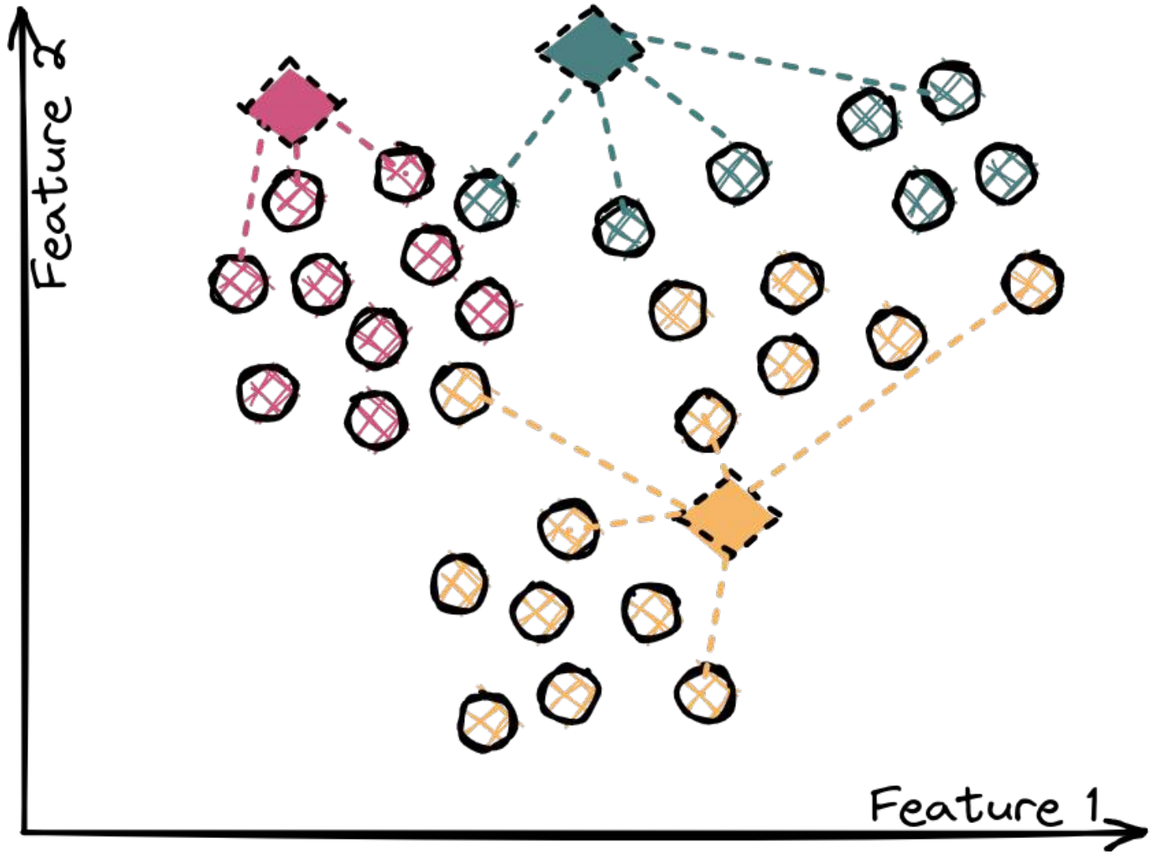
K-Means: Step 0b

Step 0b. Assign all data points to the closest centroid. In our example we'll use color coding.

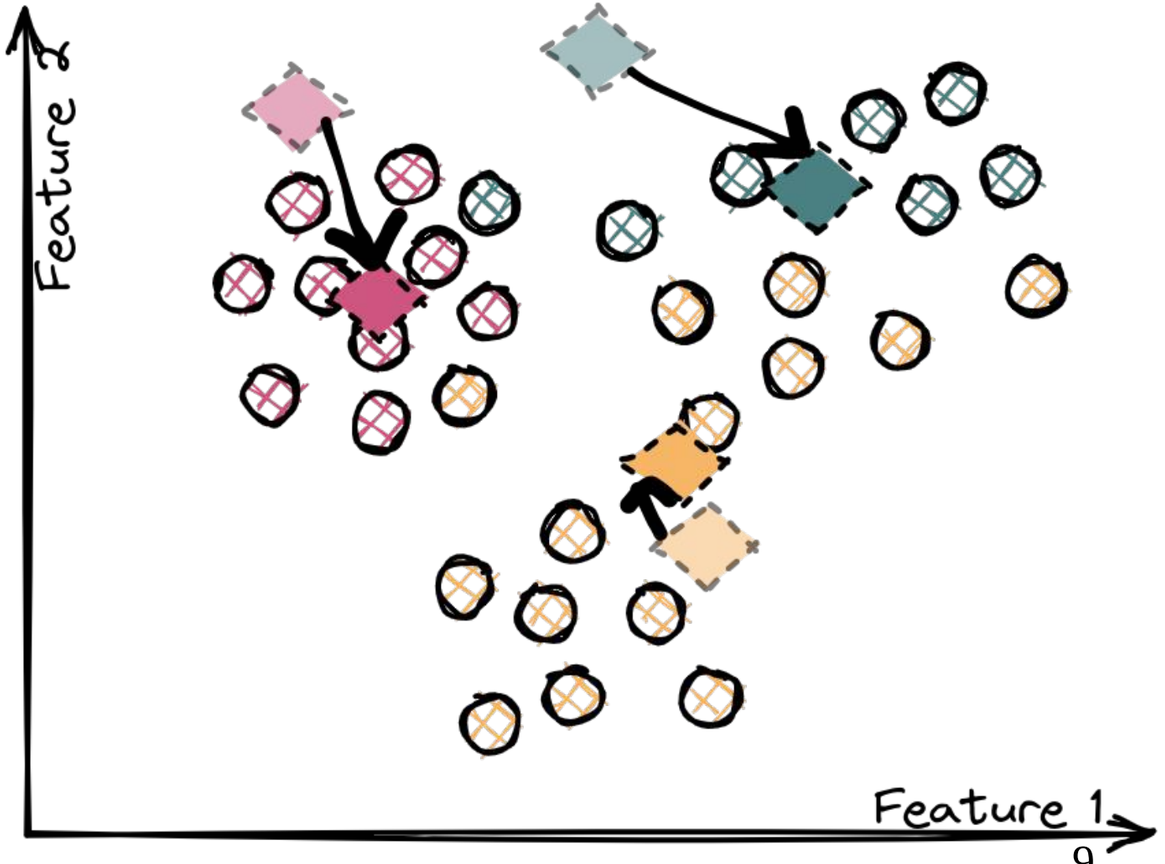


K-Means: Step1

Step1a. Calculate distances to points

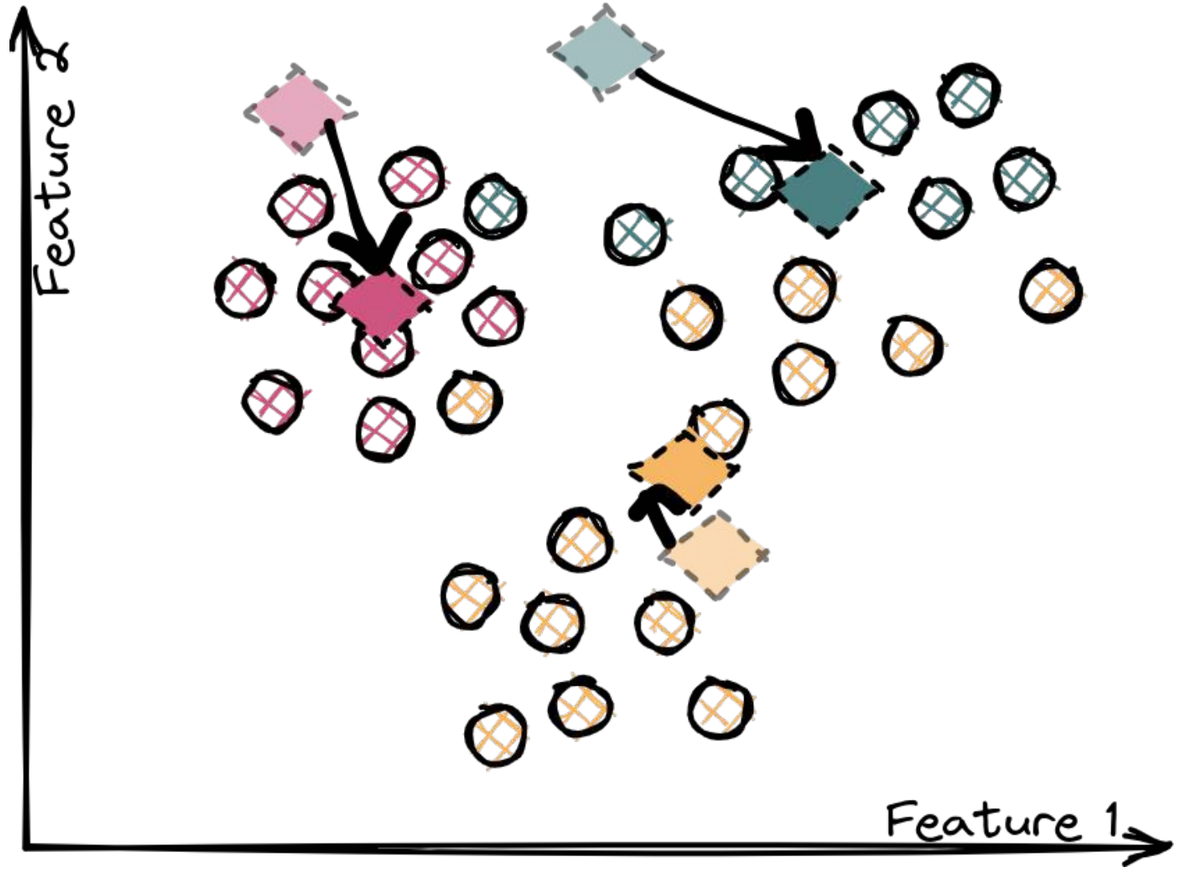


Step1b. Relocate centroids to minimize point distances

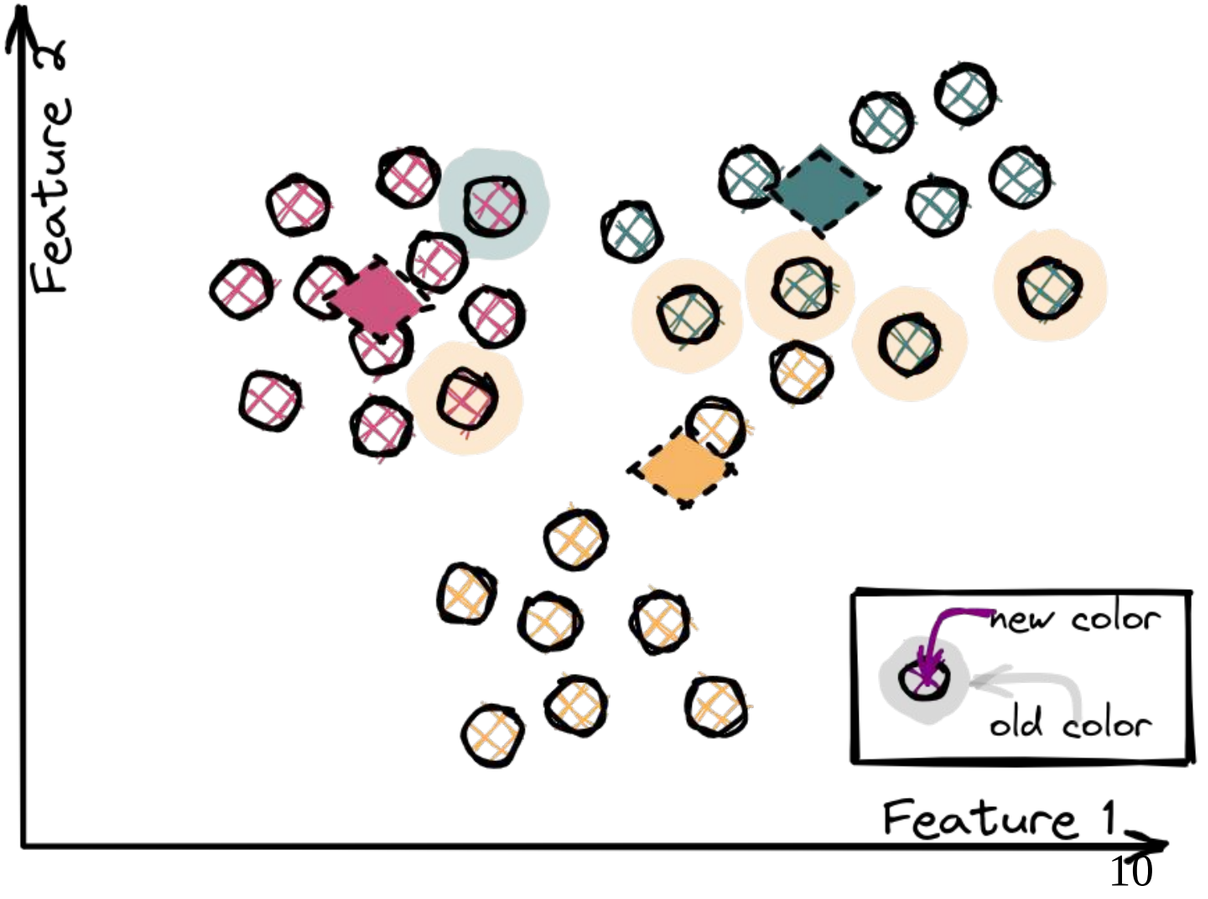


K-Means: Step1

Step1b. Relocate centroids to minimize point distances

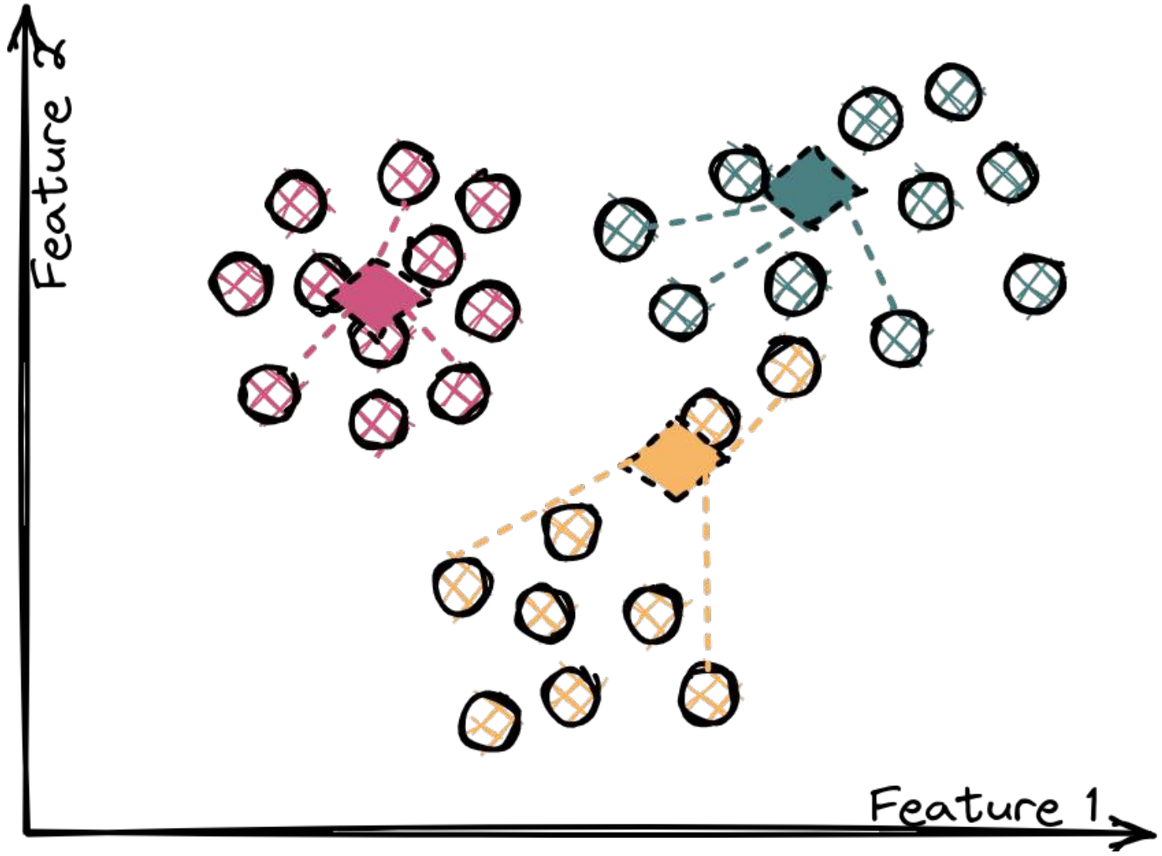


Step1c. Reassign nearest points

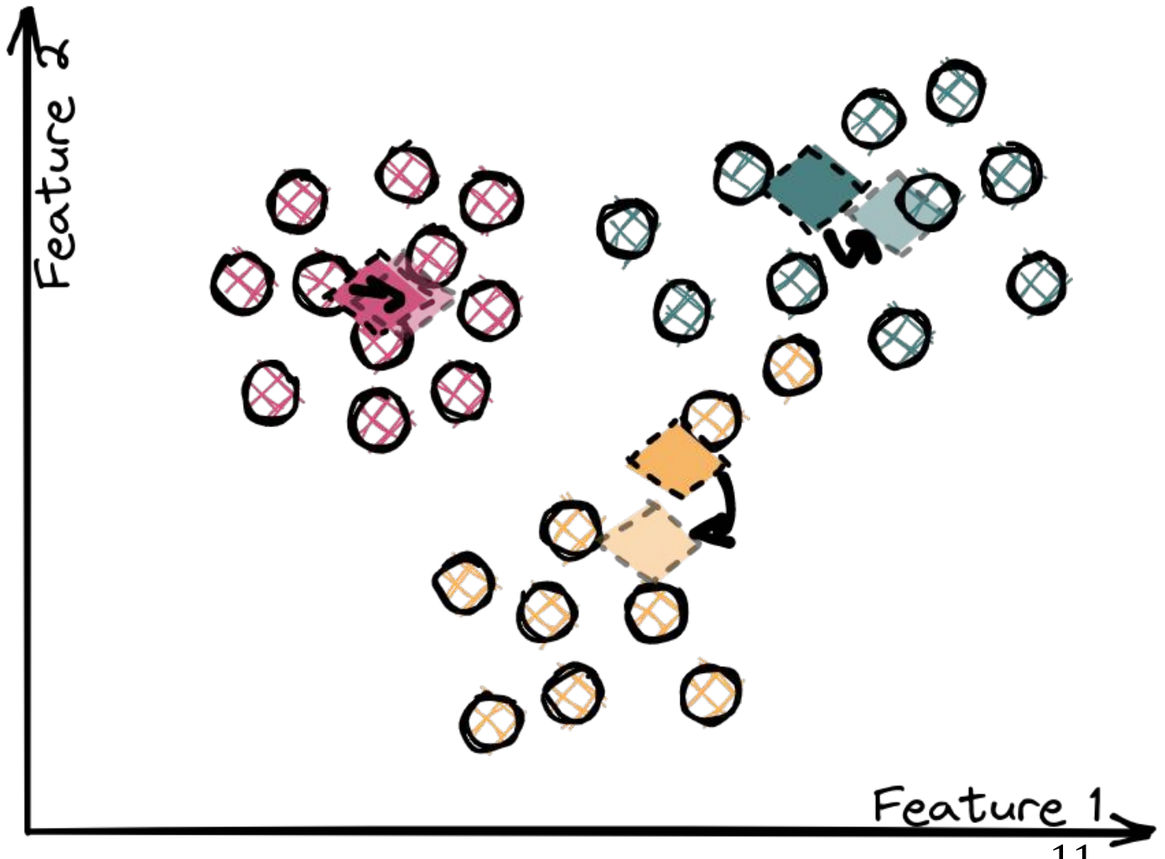


K-Means: Step2

Step2a. Calculate distances to points

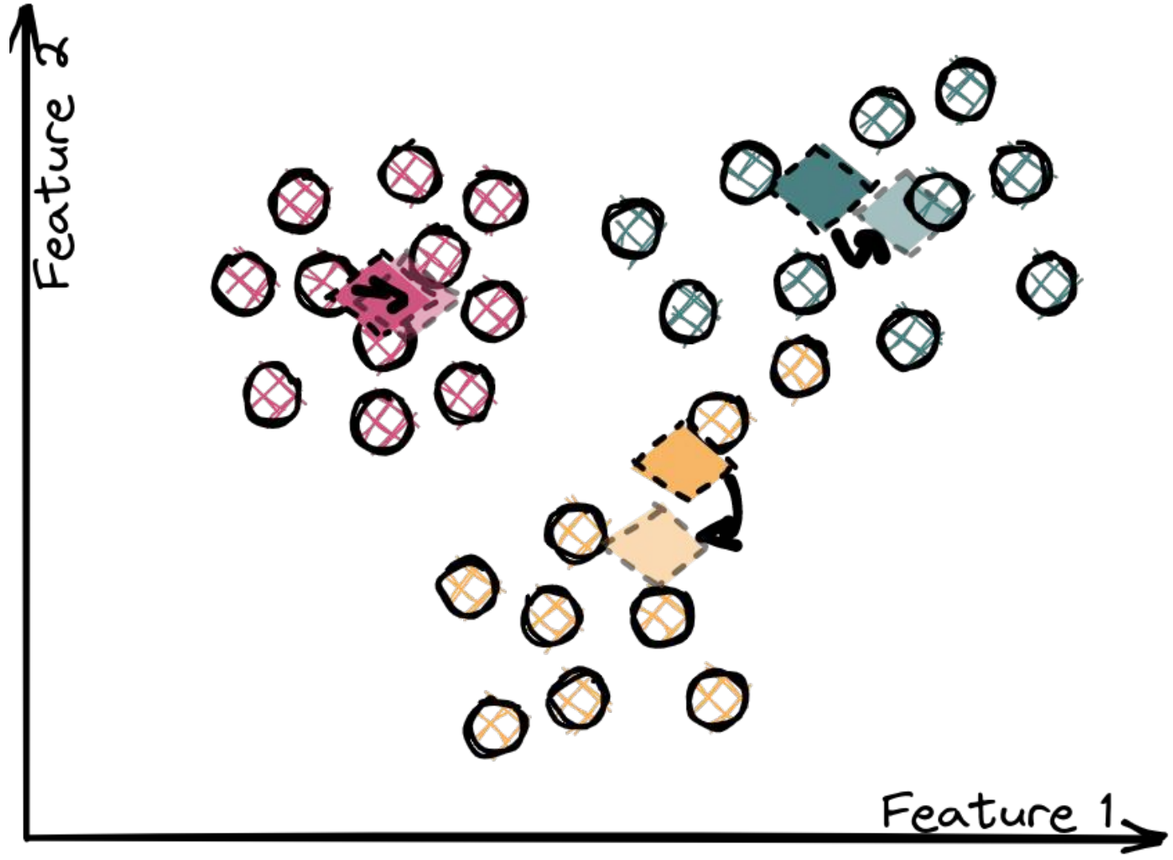


Step2b. Relocate centroids to minimize point distances

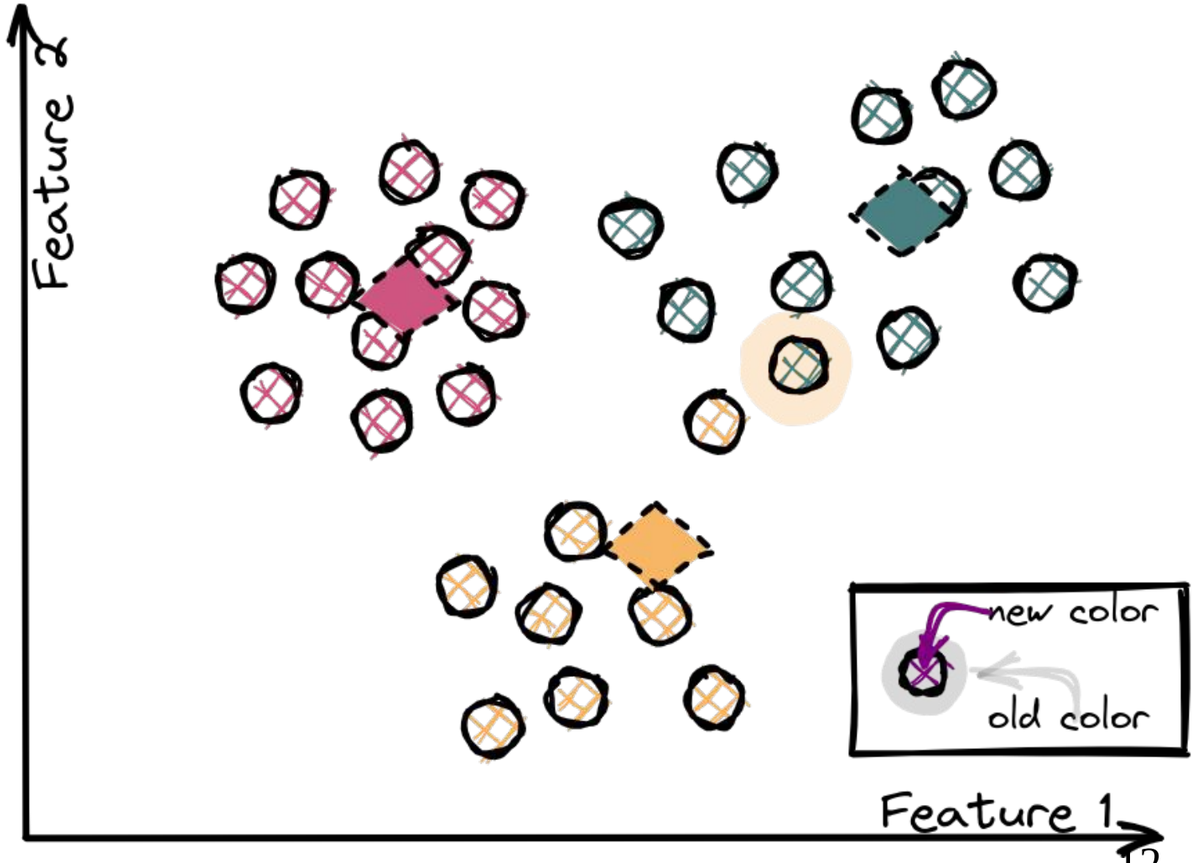


K-Means: Step2

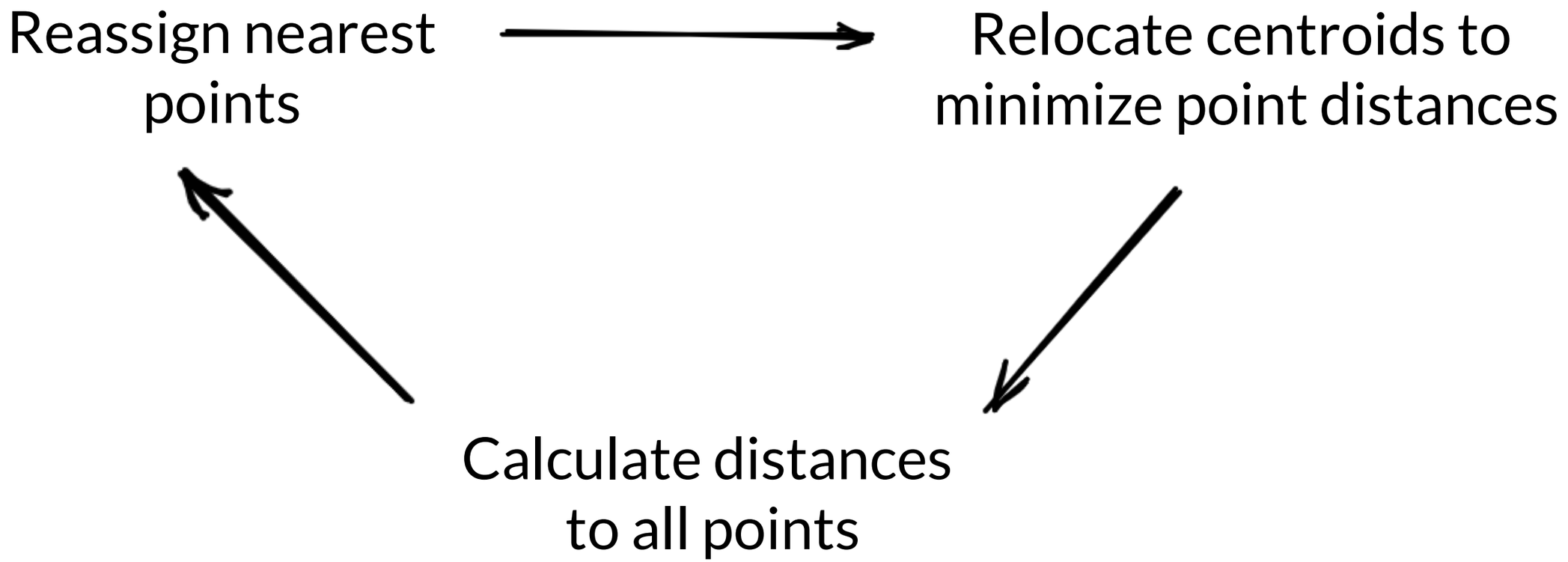
Step2b. Relocate centroids to minimize point distances



Step2c. Reassign nearest points

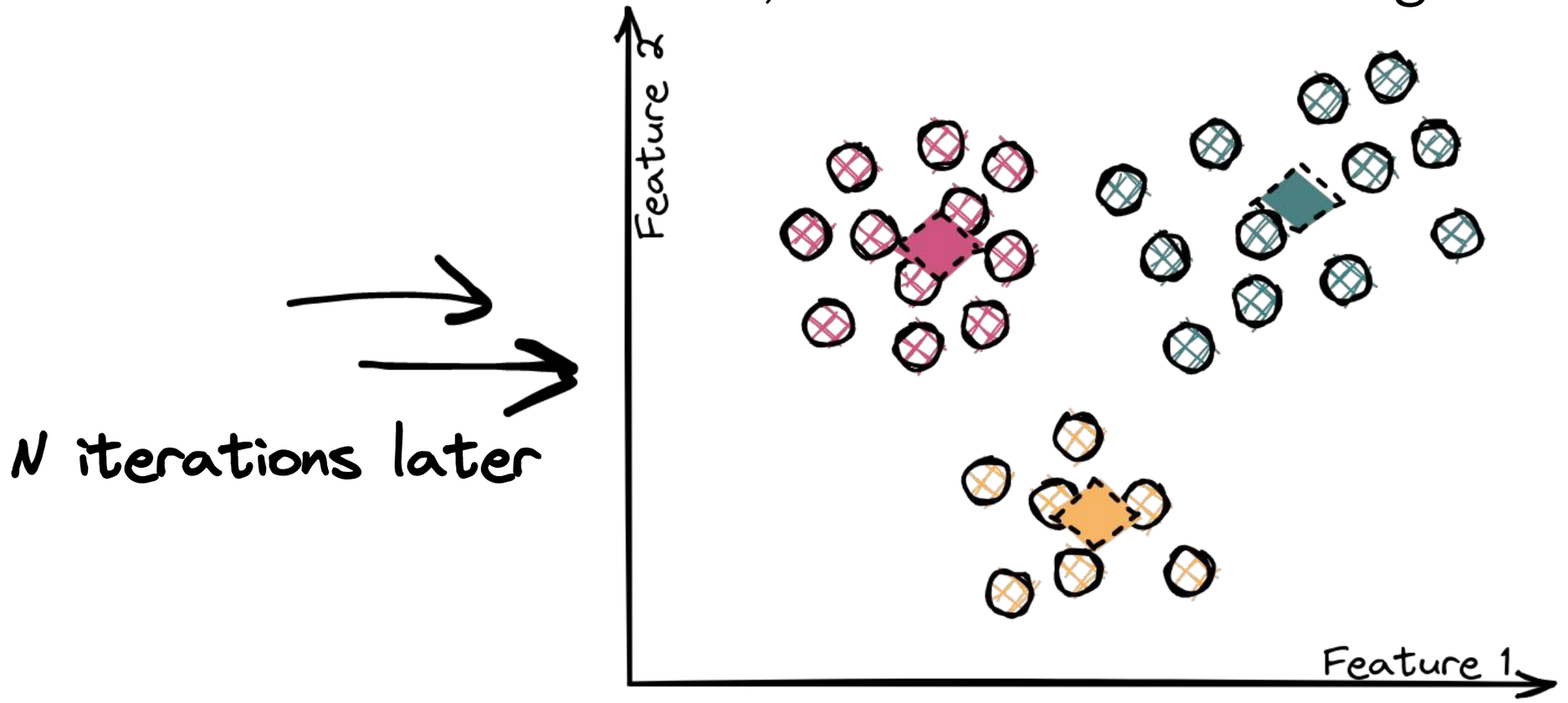


K-Means: iteration logic

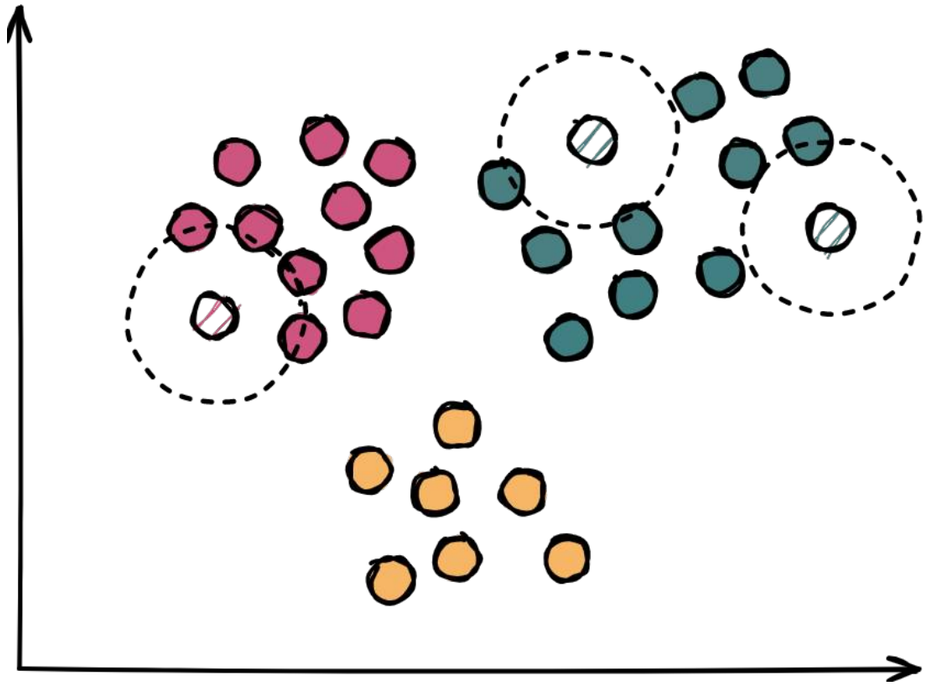


K-Means: StepN

After a while the shifting of centroids will stop. Now we assume we found the true location of centroid, and finished clustering



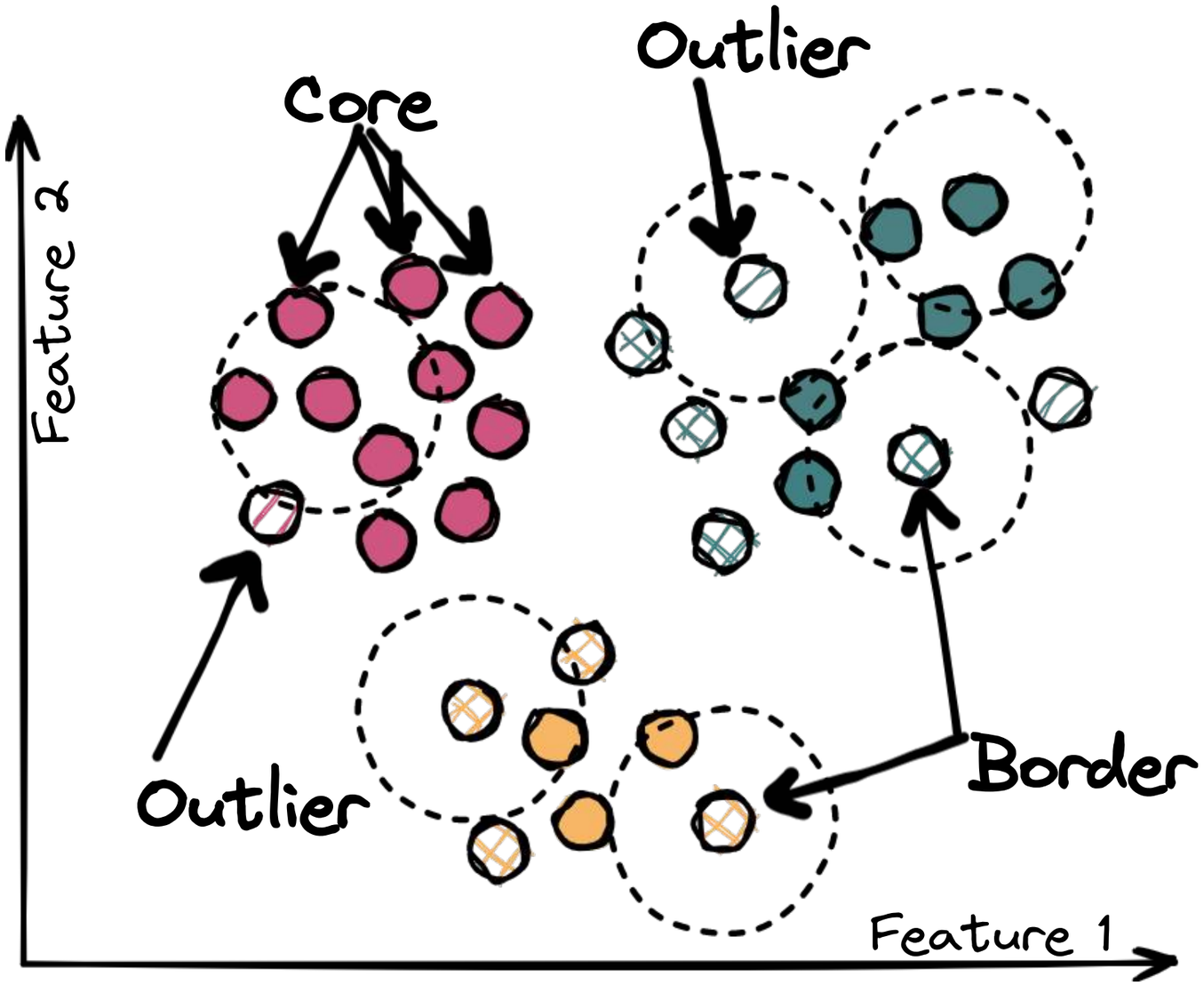
Density-based clustering



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

Most used method:
DBSCAN

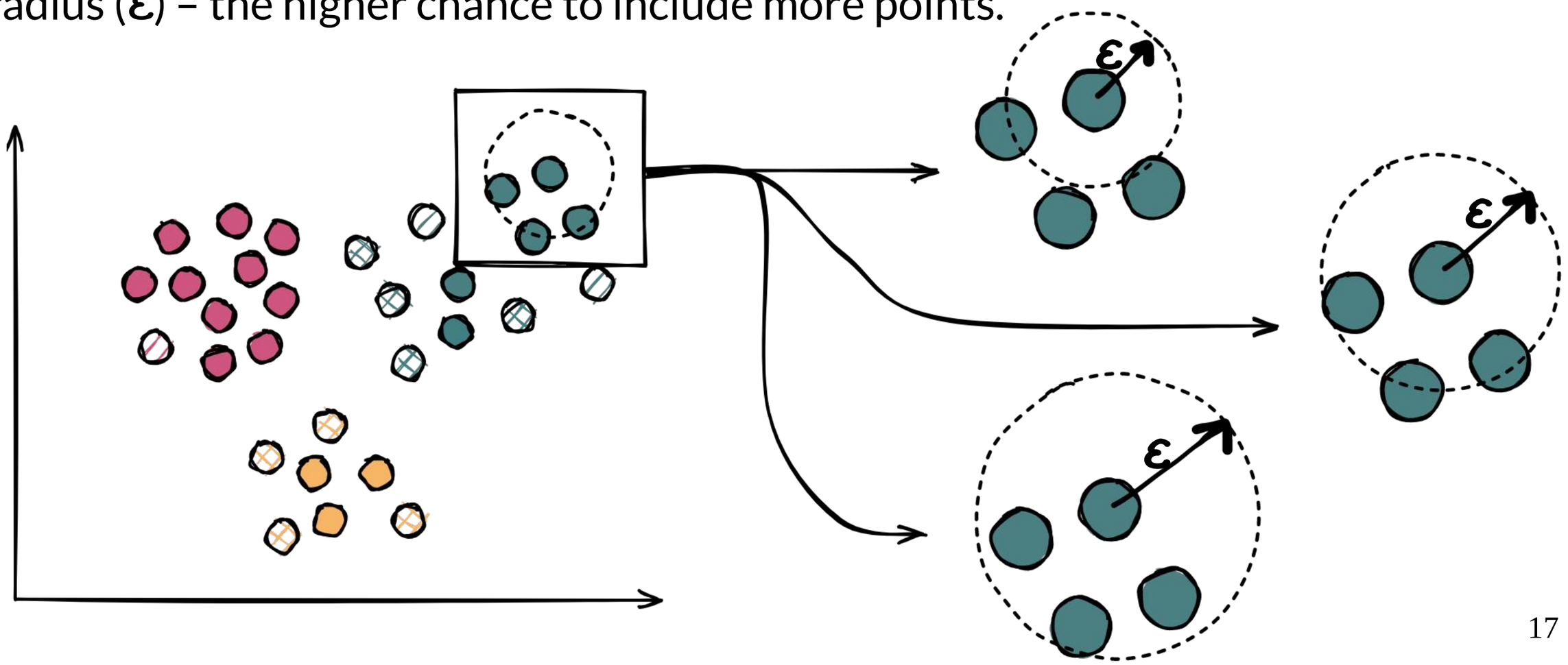
Density-based clustering



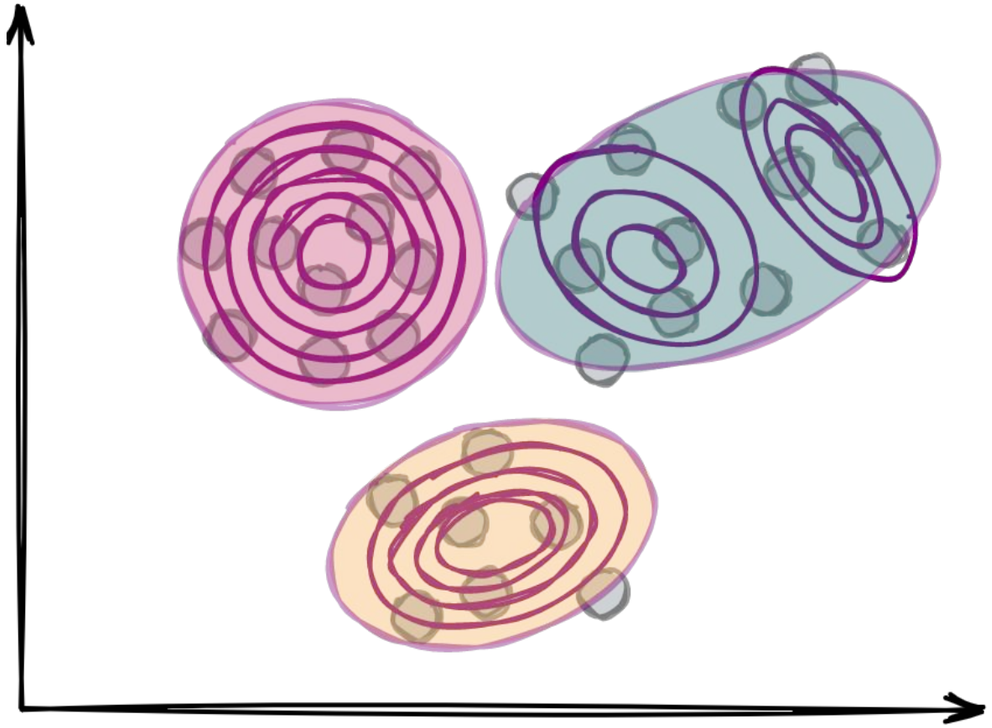
One of the parameters varied is the number of neighbors to be considered as part of cluster. N=1 may be a good default value.

Density-based clustering

Another parameter to vary is the radius of the "neighbor" circle. The bigger the radius (ε) – the higher chance to include more points.



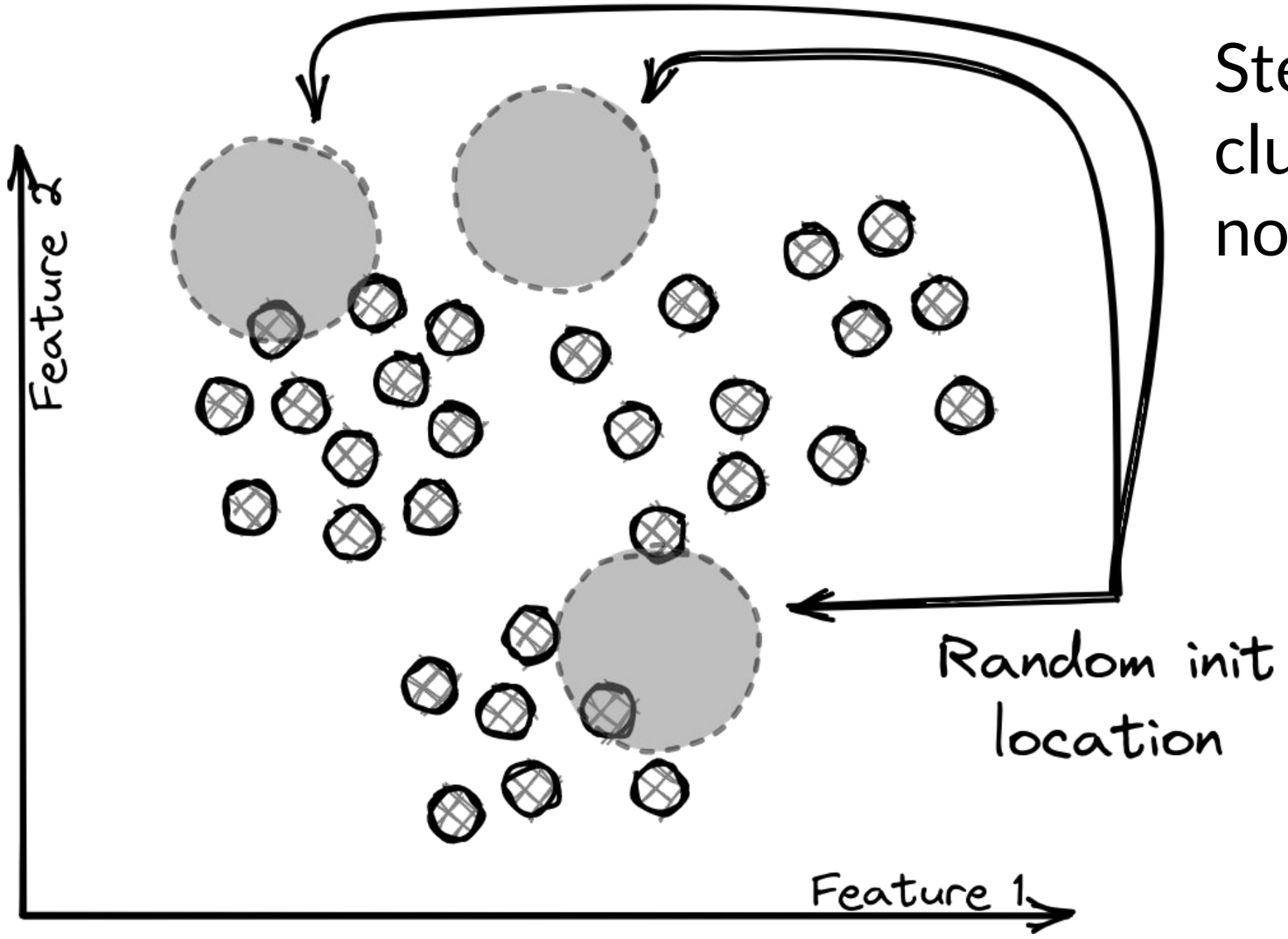
Model-based clustering



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

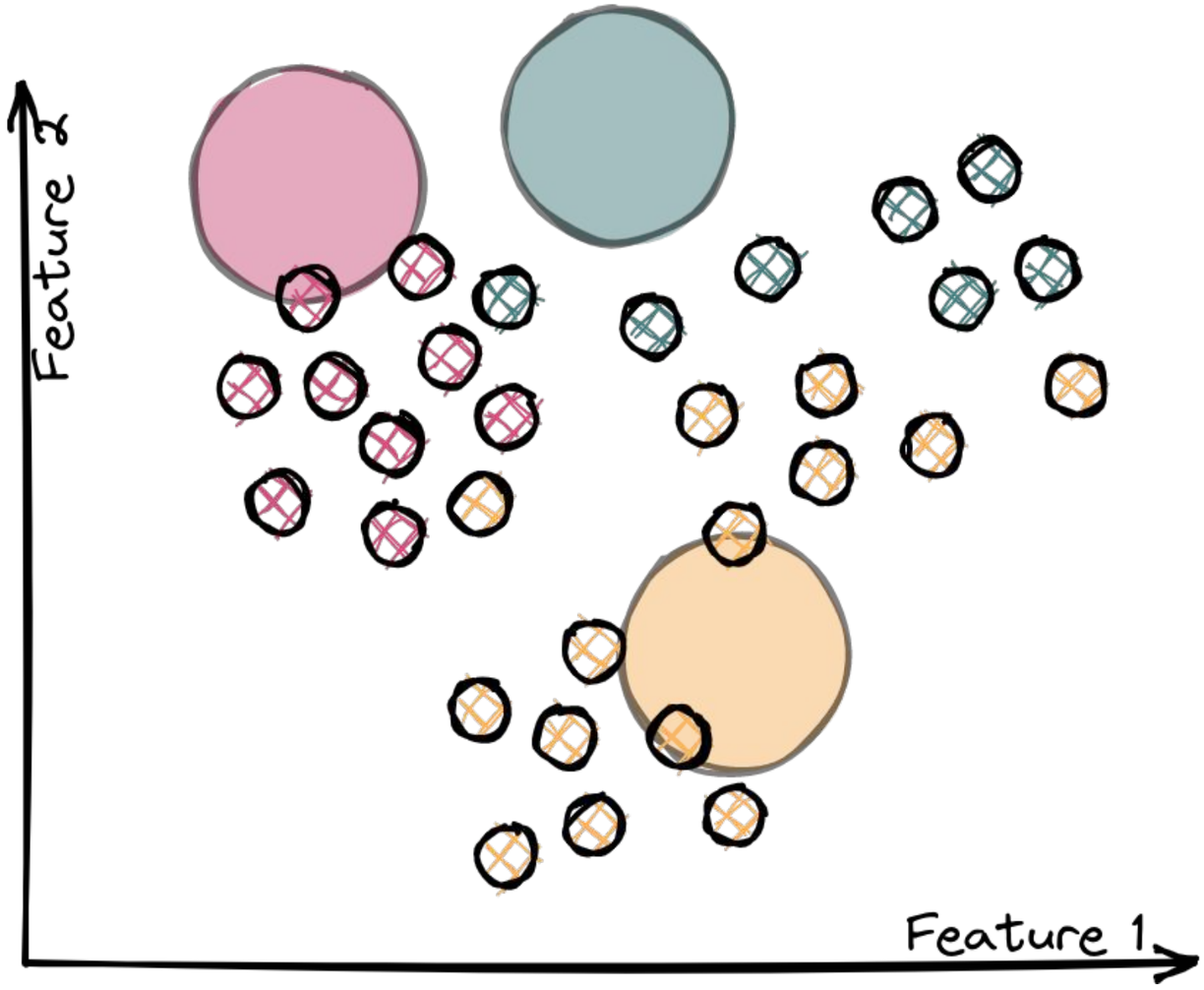
Most used method:
GMM – Gaussian Mixture Models

GMM: Step 0a



Step 0a. Randomly set cluster centers and normal distributions

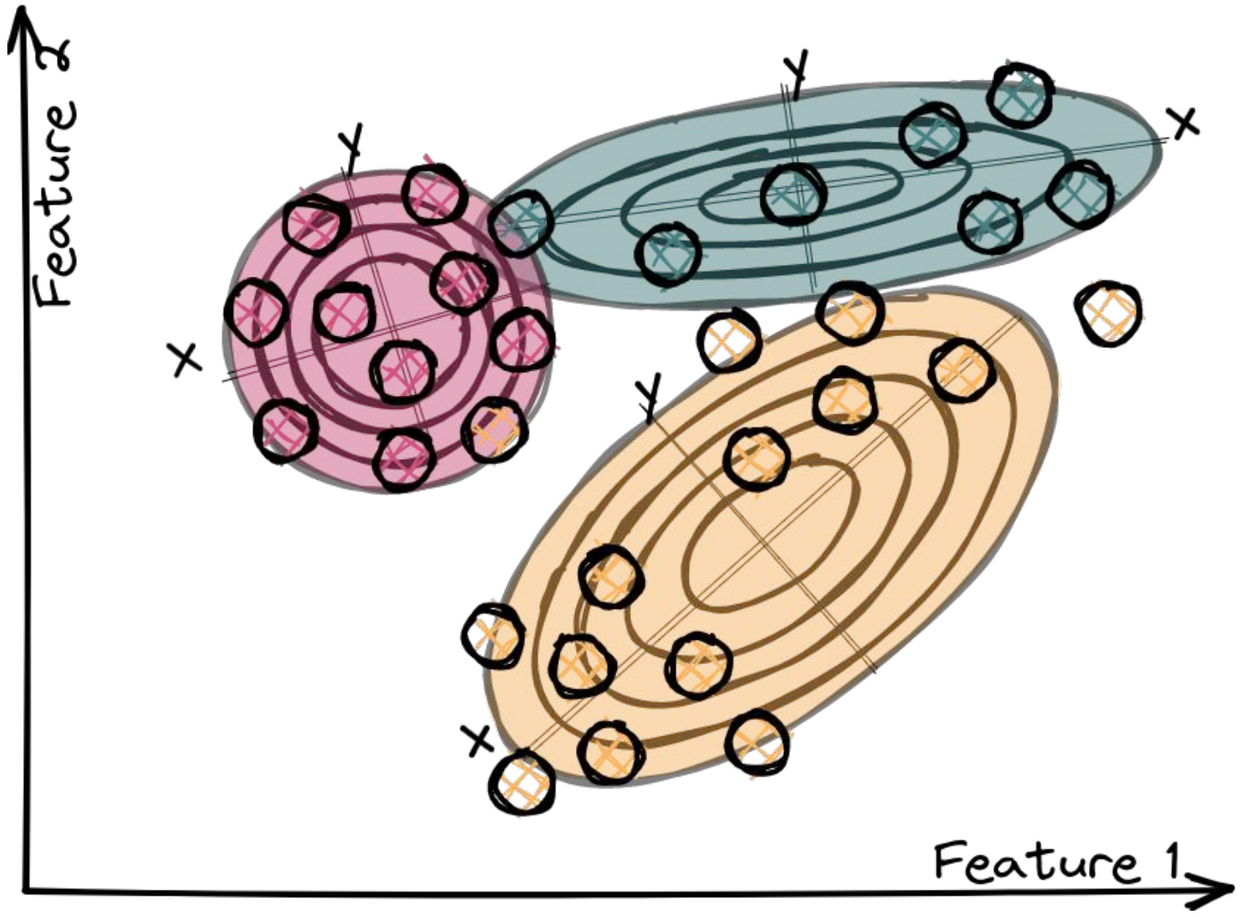
GMM: Step 0b

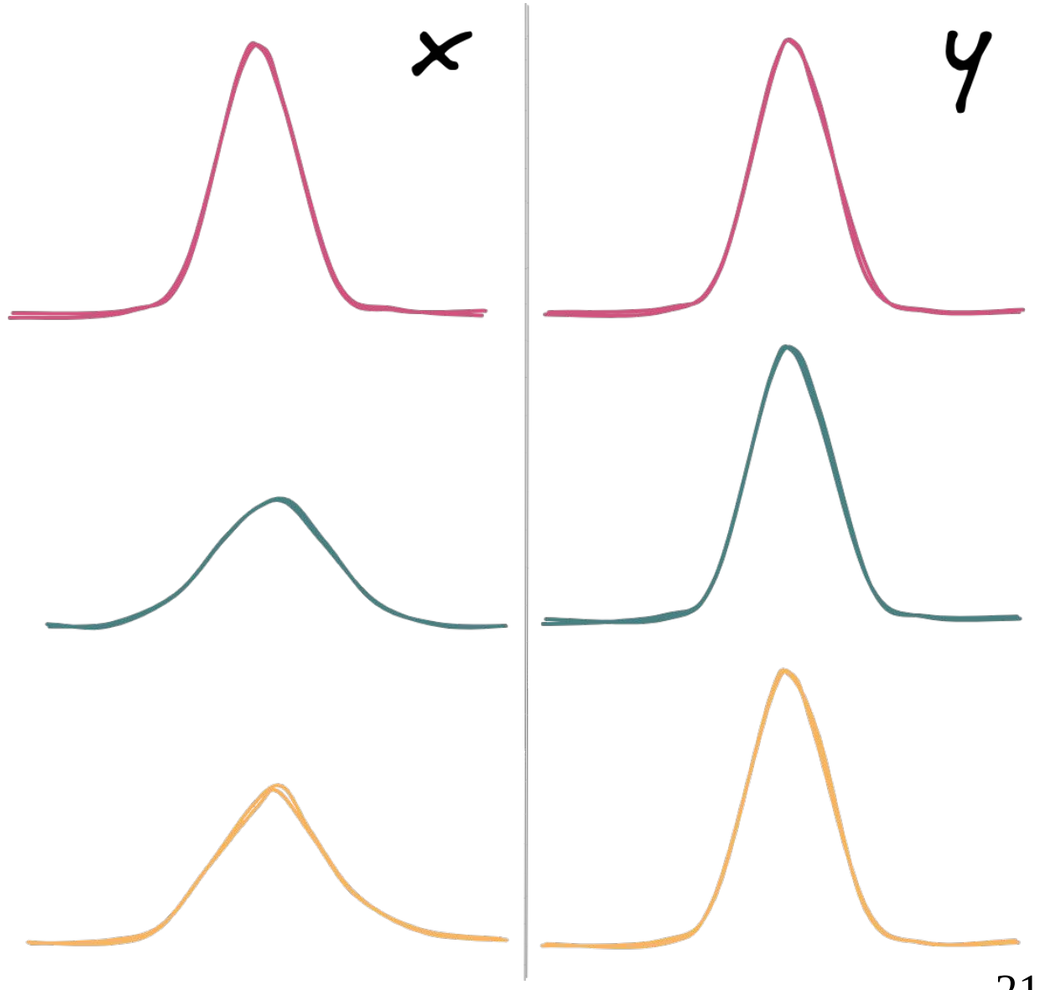


Step 0b. Assign all data points to the closest distribution center. In our example we'll use color coding.

GMM: Step1a

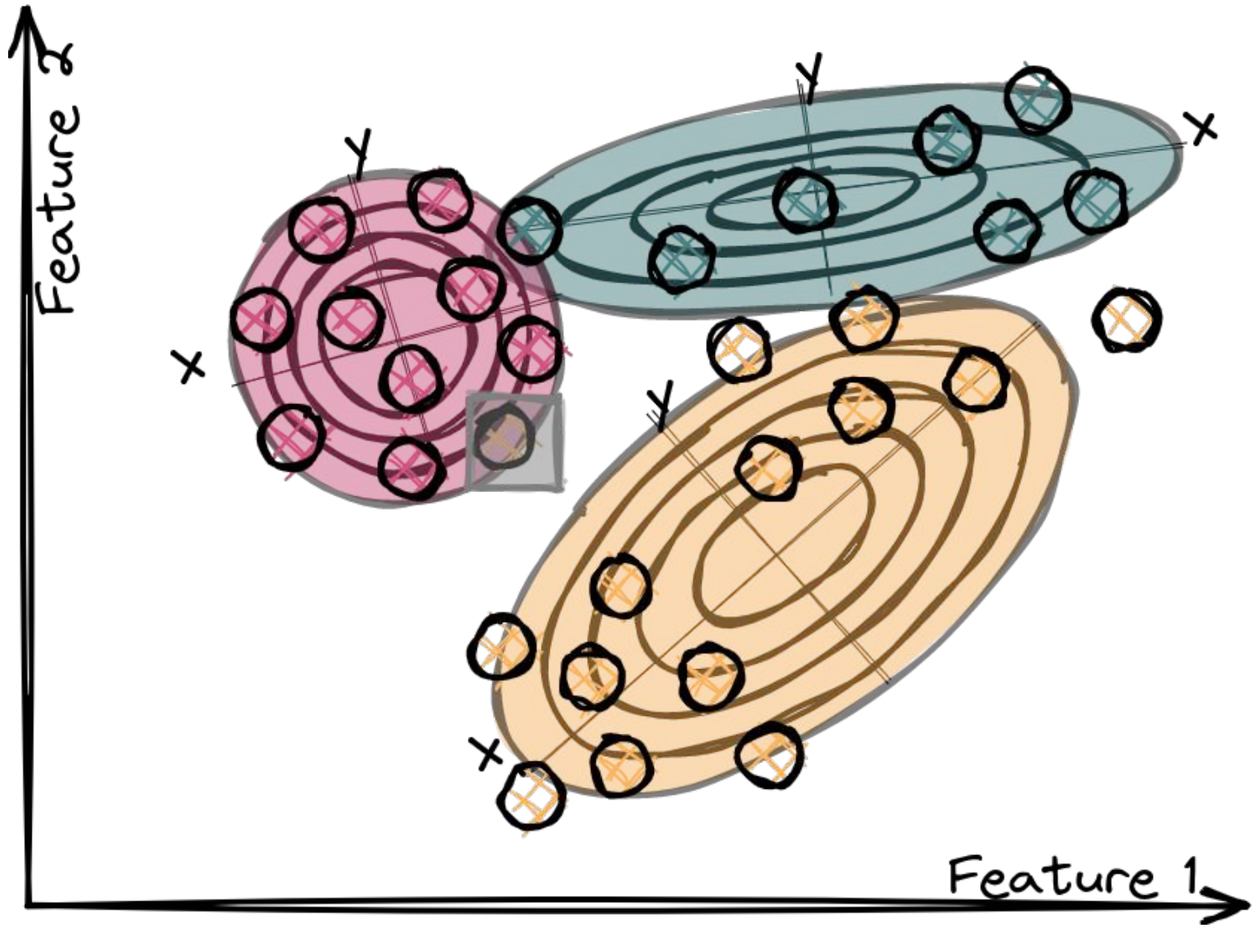
Step1a. Update distributions

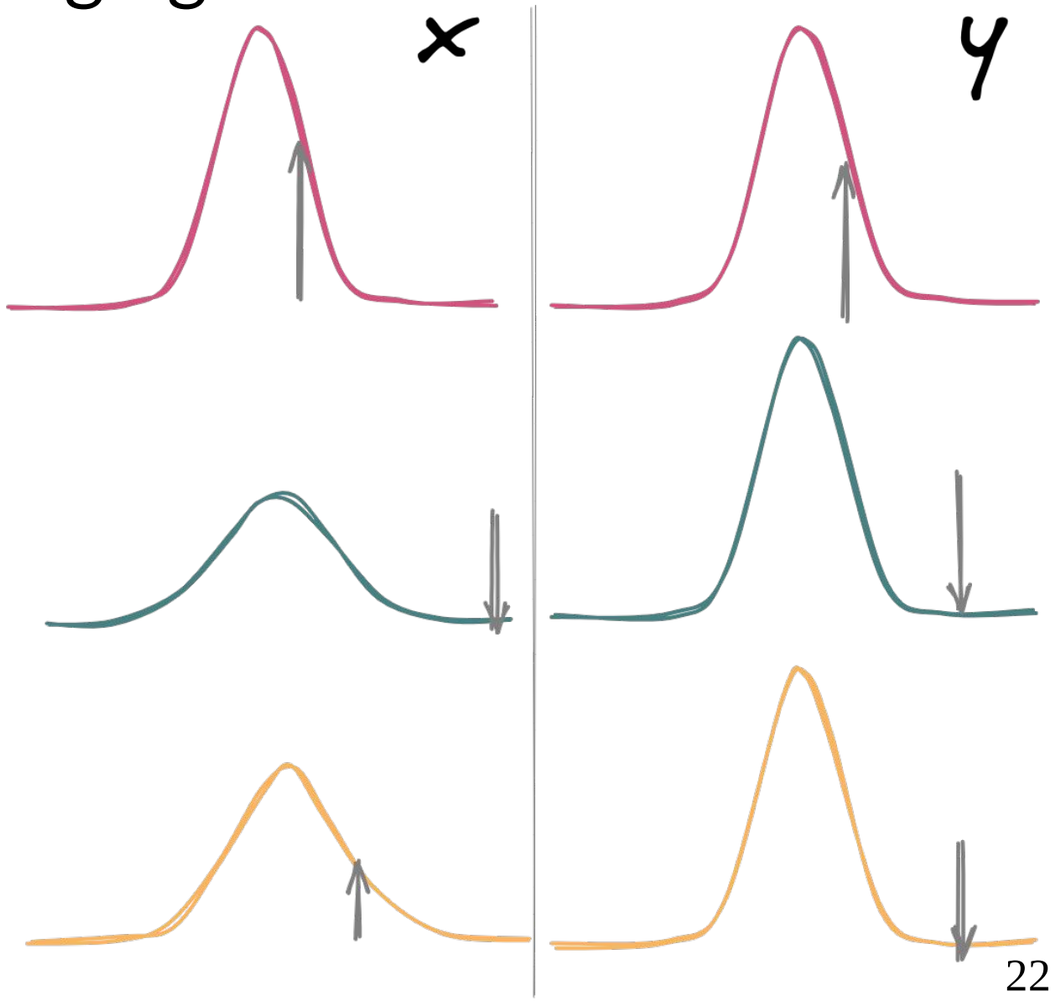




GMM: Step1b (example1)

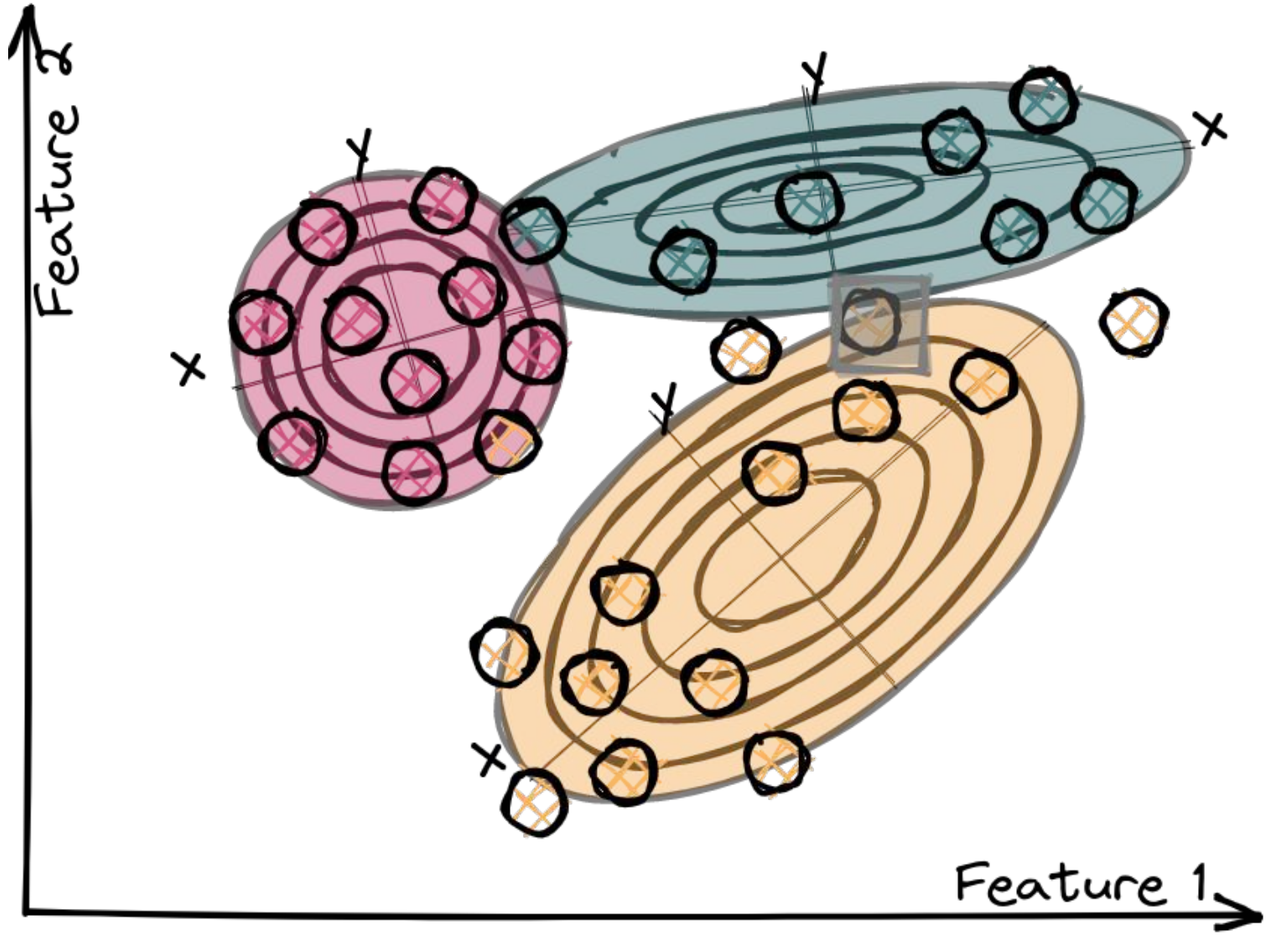
Step1b. Update data point cluster belonging

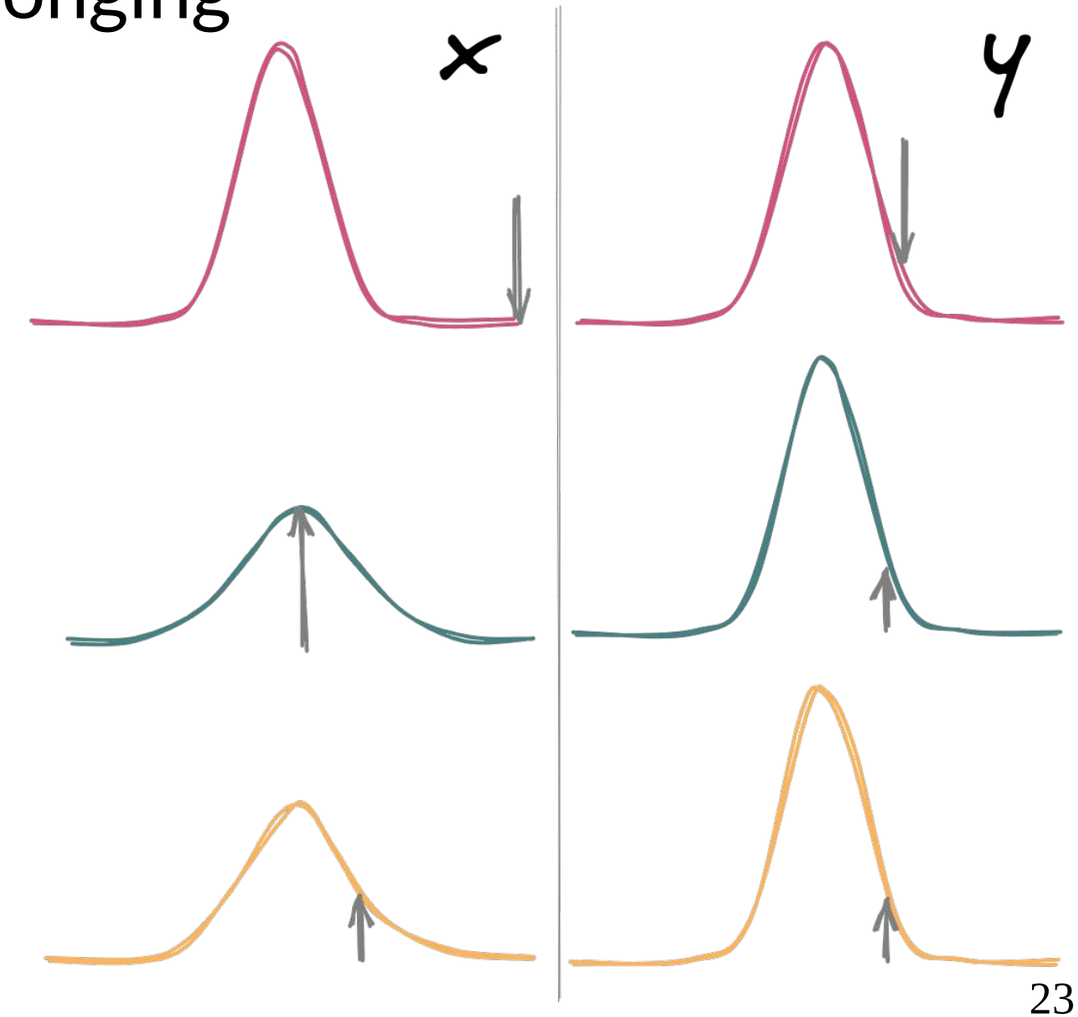




GMM: Step1b (example2)

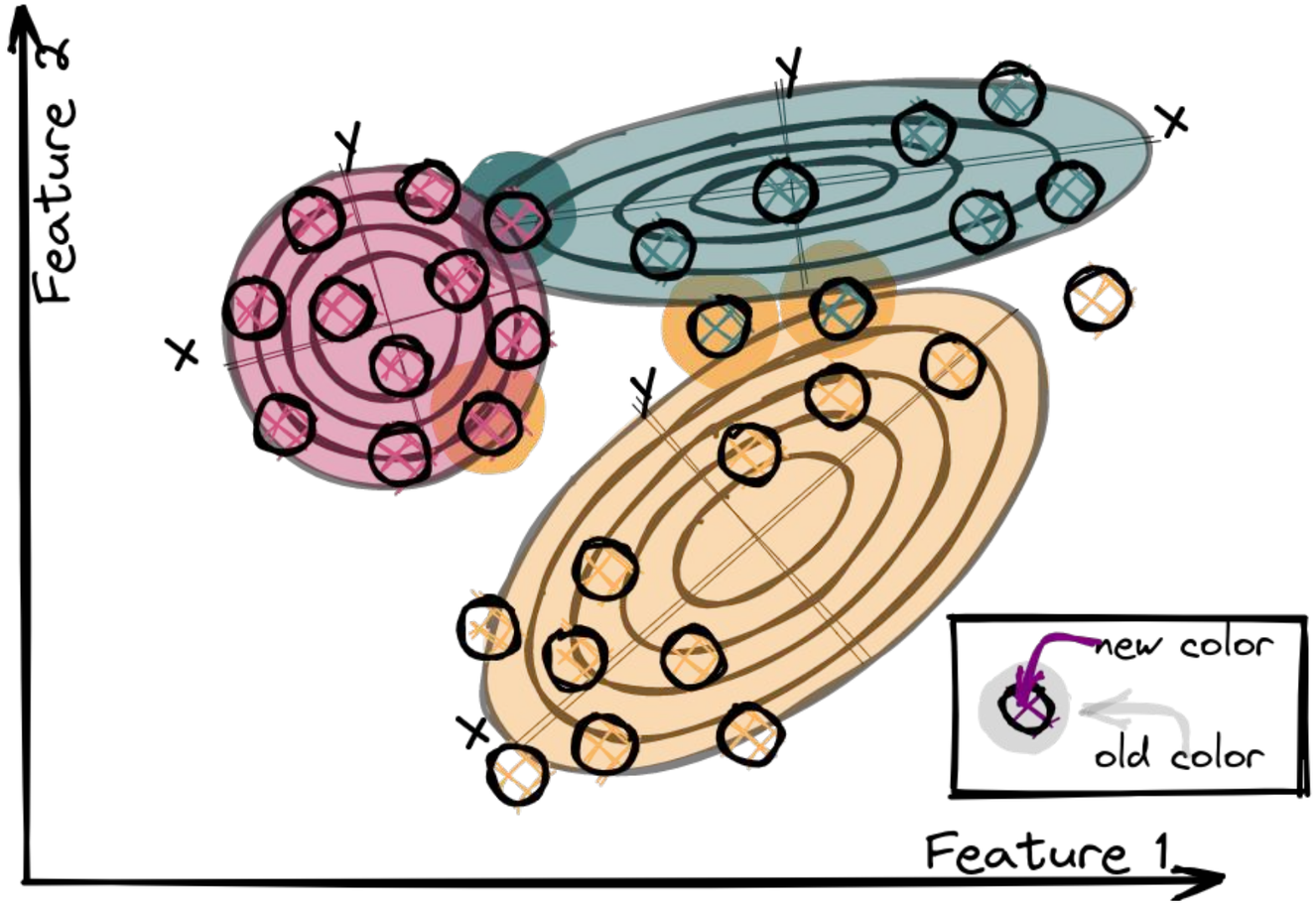
Step1b. Update data point cluster belonging



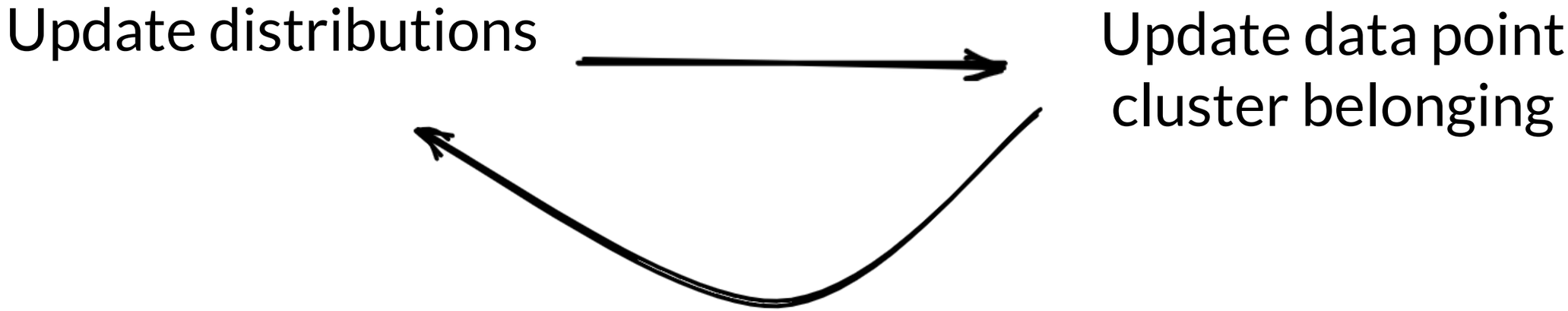


GMM: Step1b (overall)

Step1b. Update data point cluster belonging

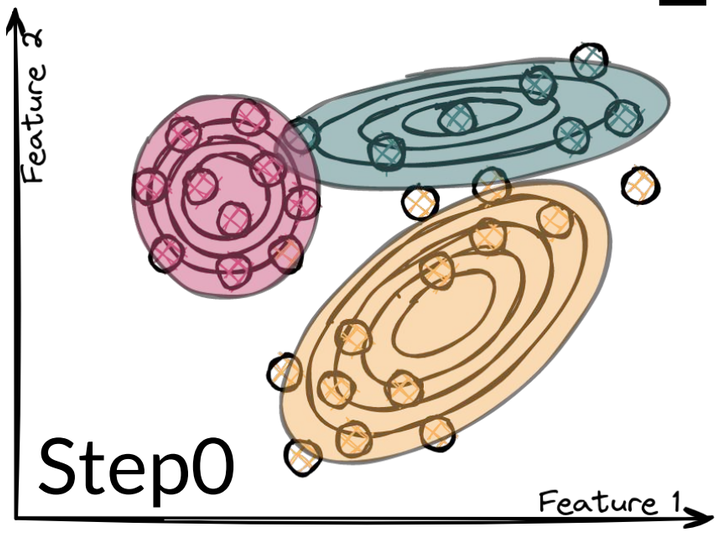


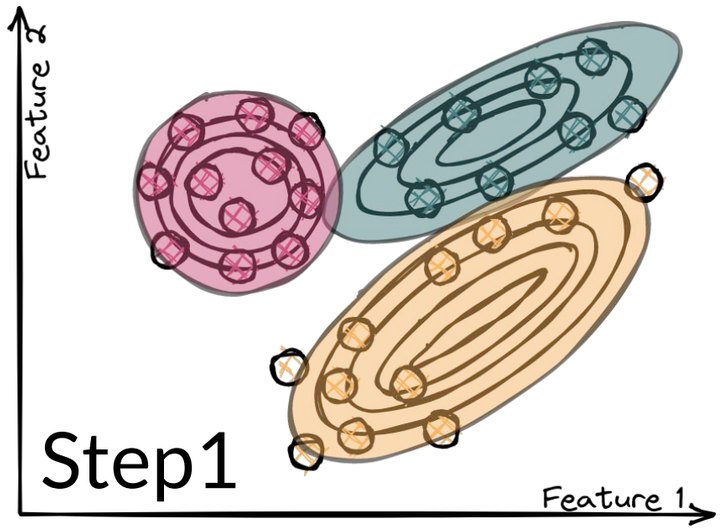
GMM: iteration logic

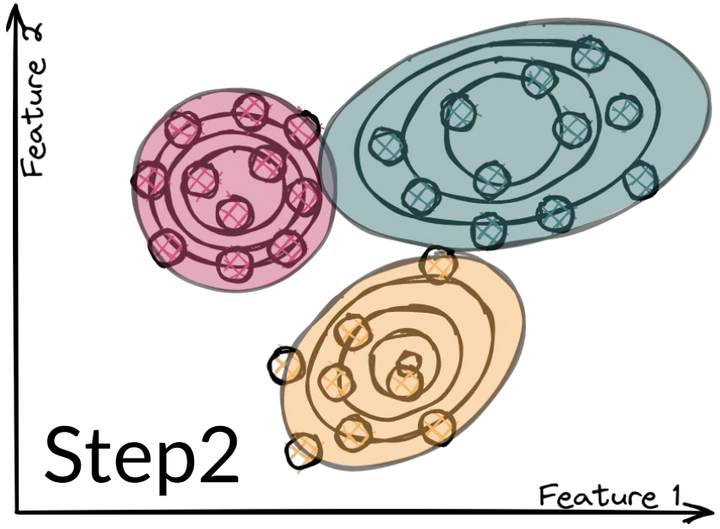


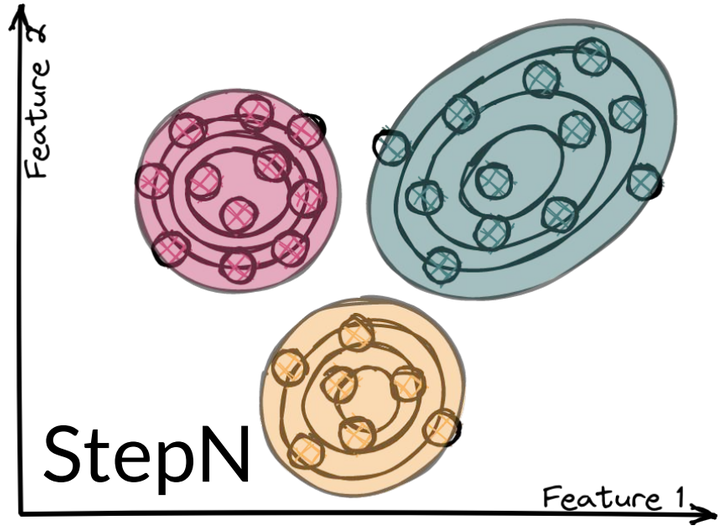
Expectation-Maximization is a form of optimization function. We start with random Gaussian parameters and check if a sample belongs to a cluster X (expectation). Afterwards, the Gaussian parameters are updated to fit the points assigned to the said cluster. The maximization step aims at increasing the likelihood of the sample belonging to the cluster distribution.

GMM: Step N









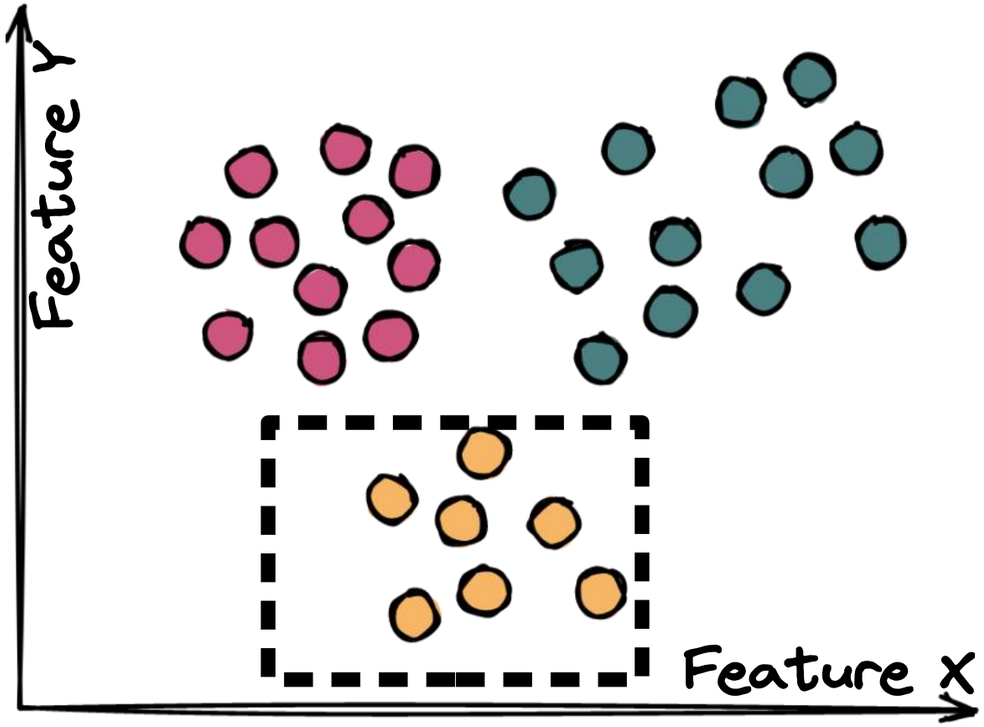
After a while the distributions will stop changing.
Now we assume we finished clustering

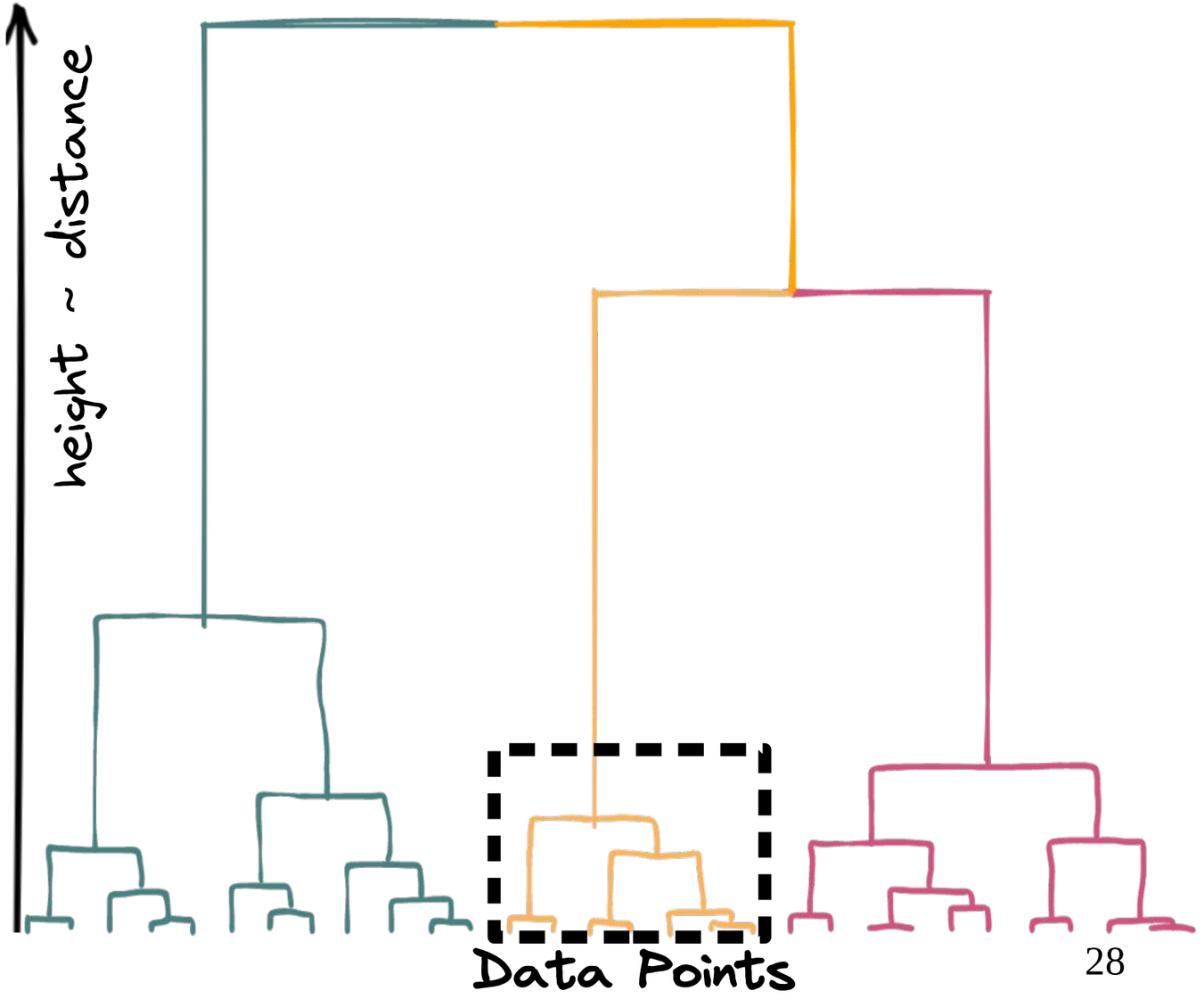
Distance-based clustering

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

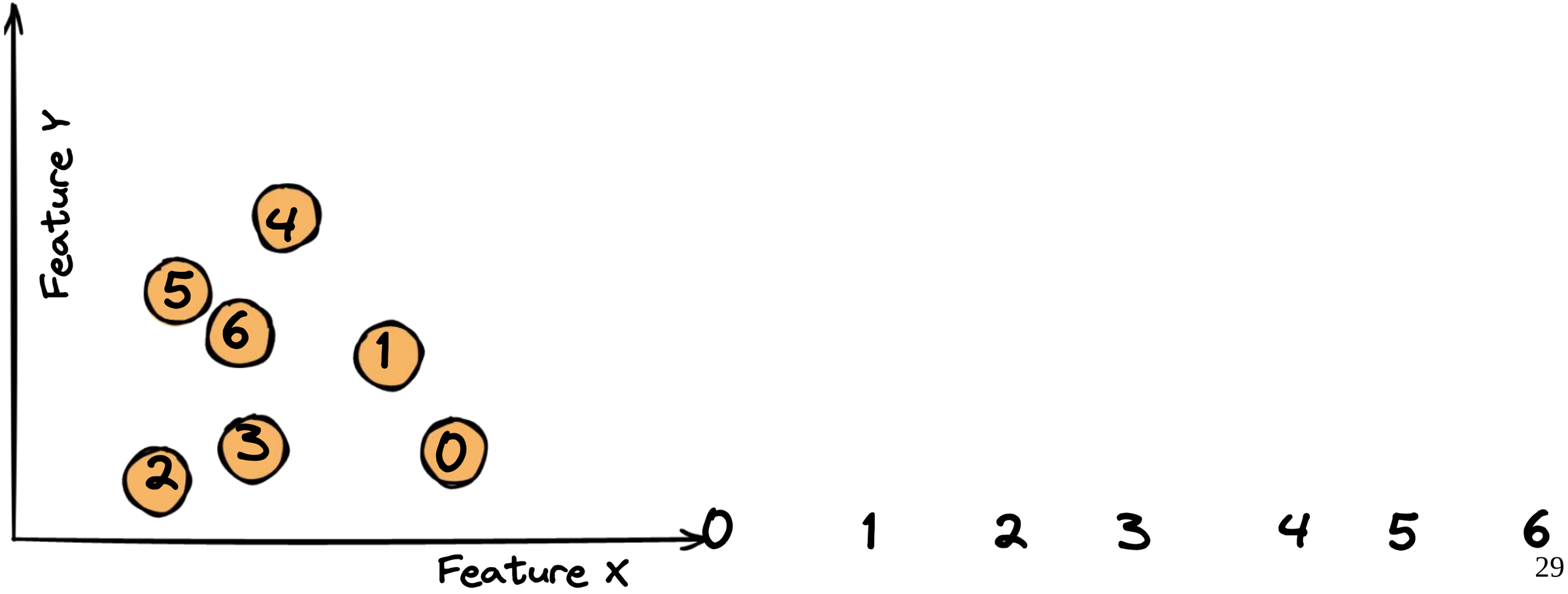
Most used method:
AHC – Agglomerative hierarchical clustering

Distance-based clustering

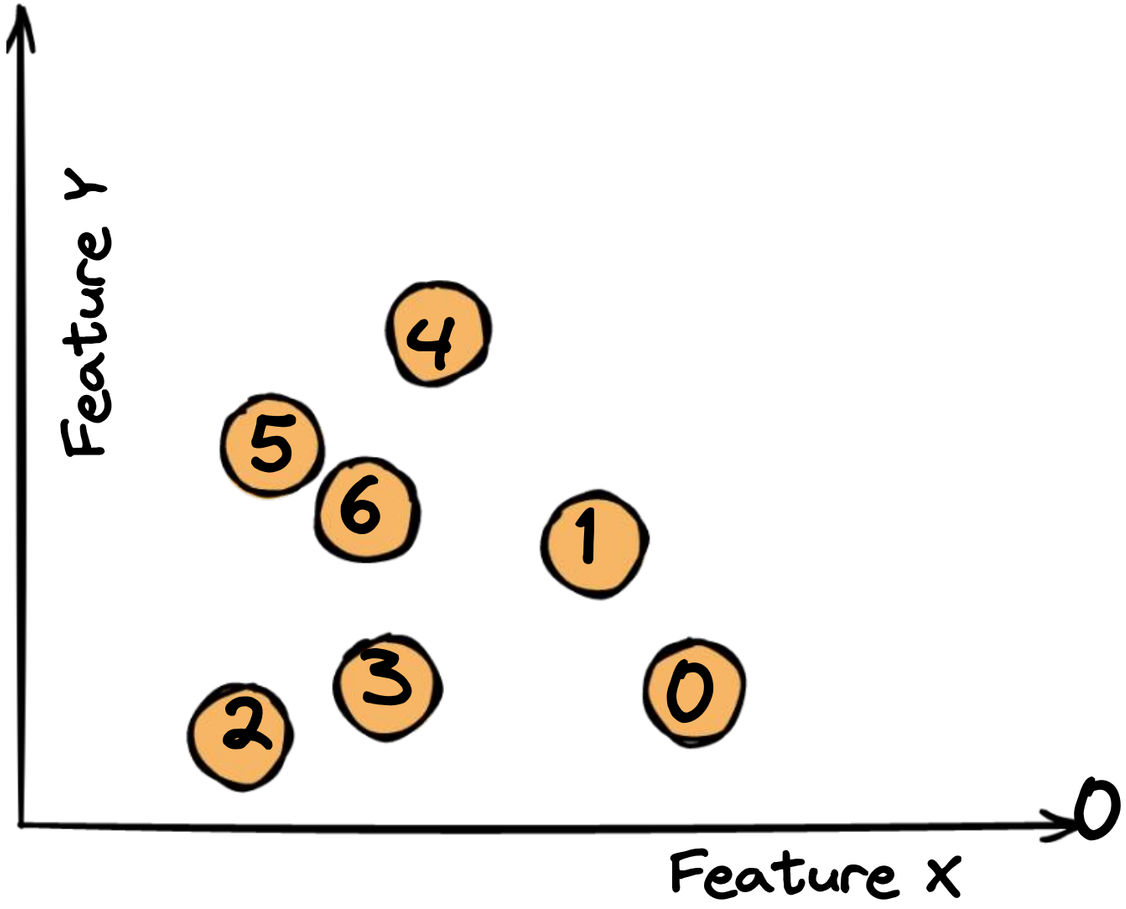


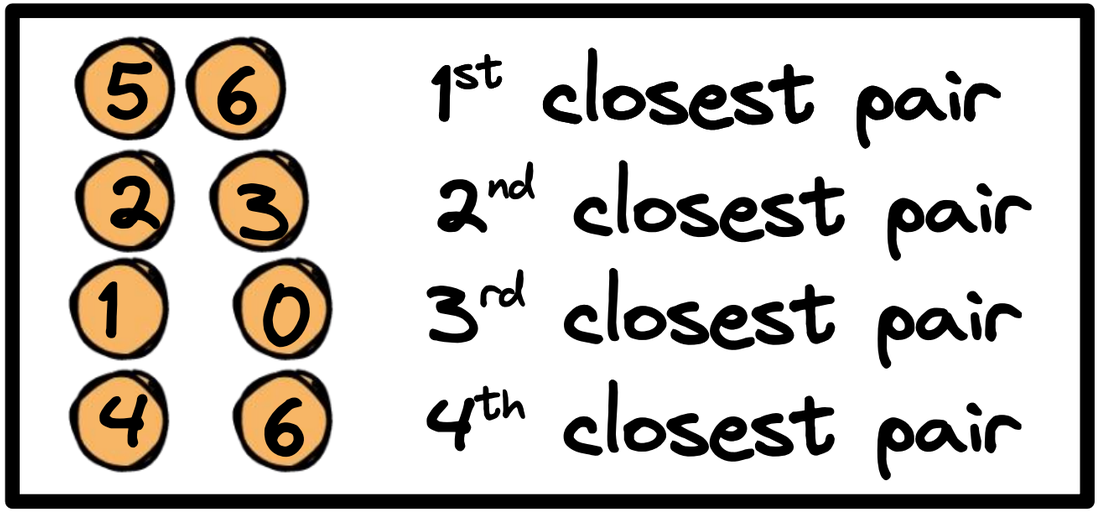


Distance-based clustering

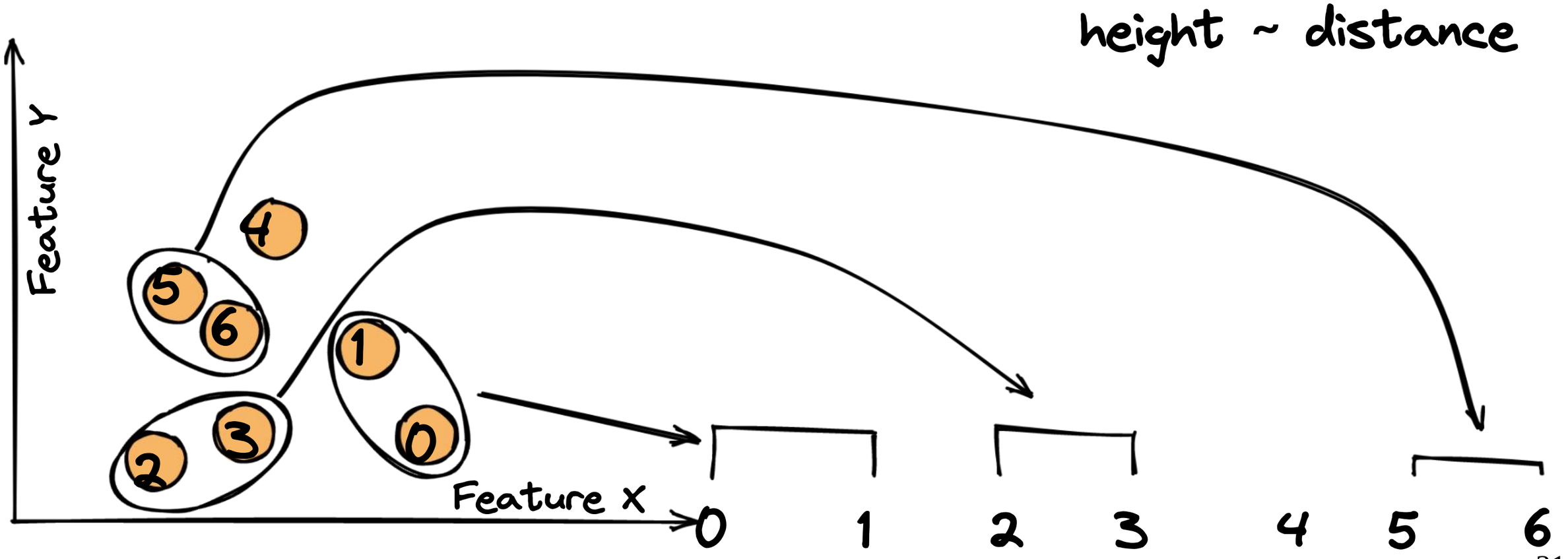


Distance-based clustering

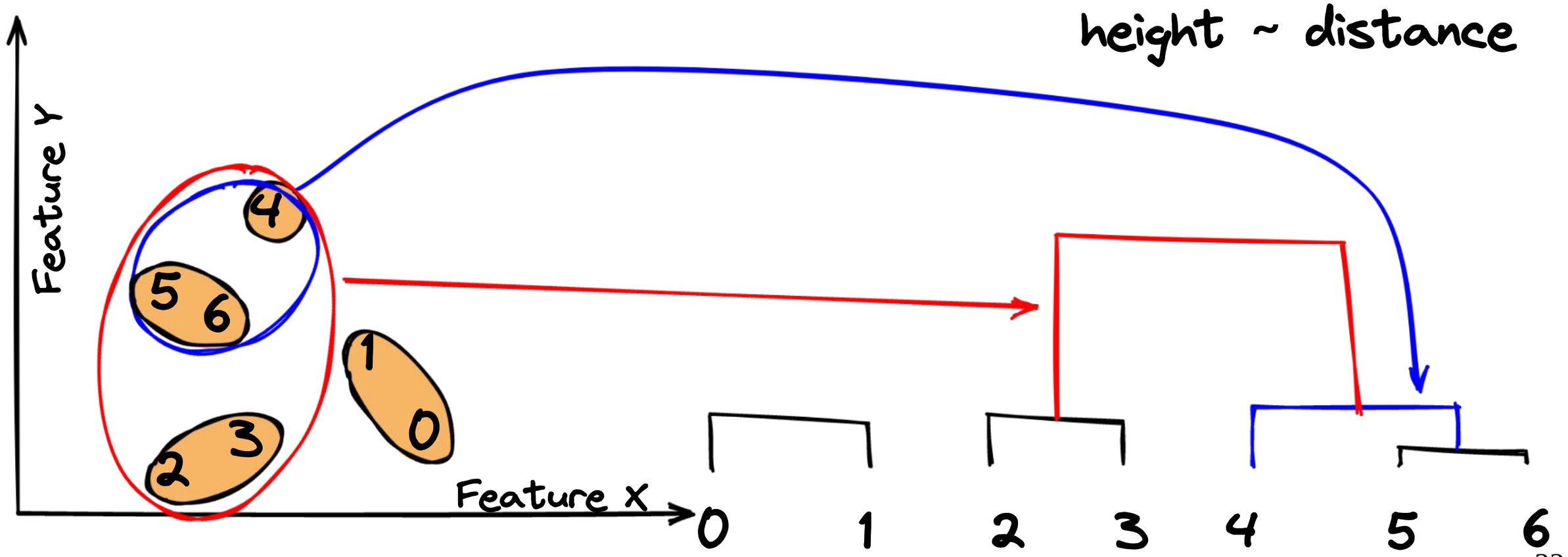




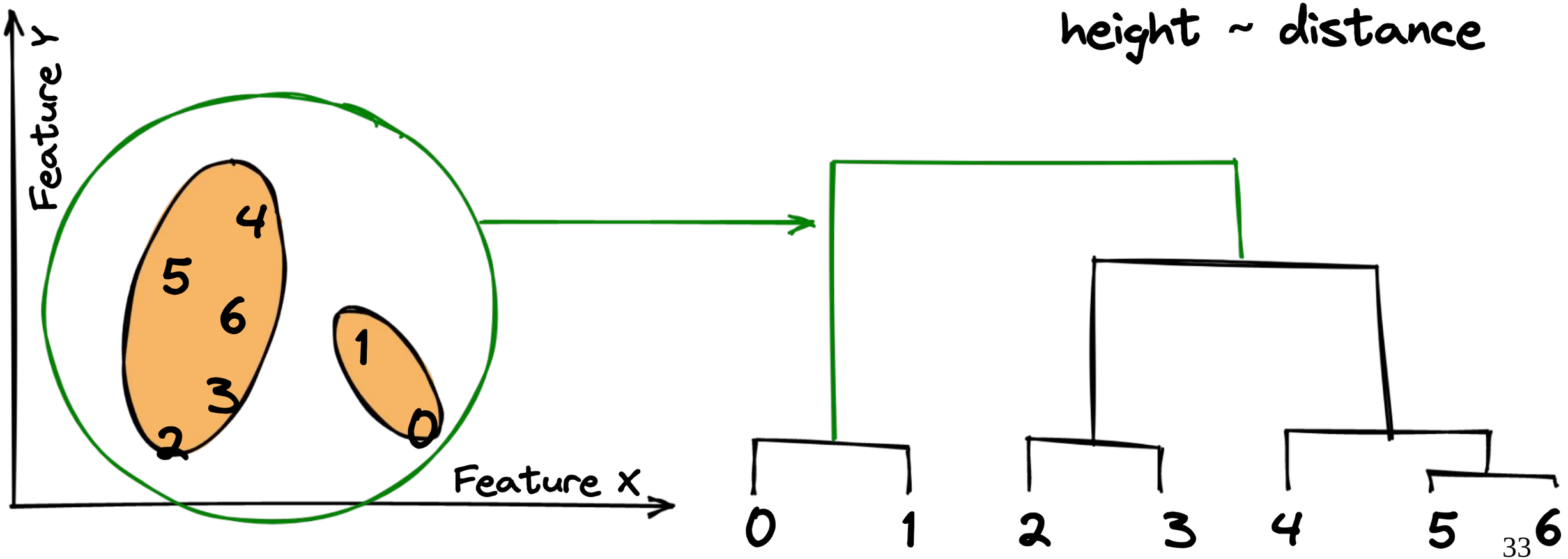
Distance-based clustering



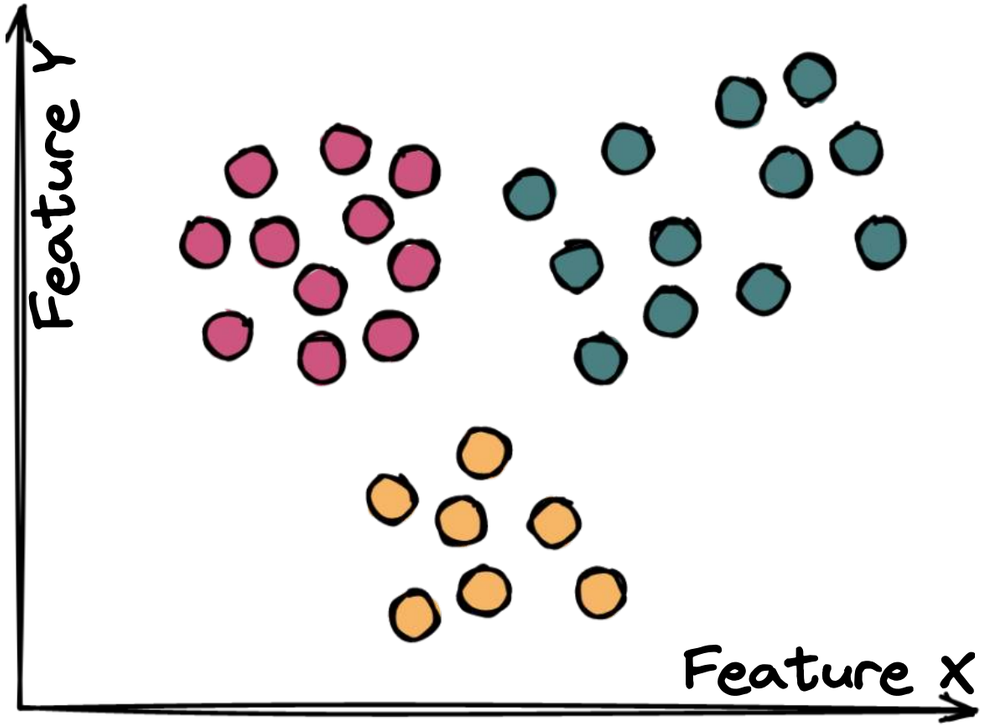
Distance-based clustering

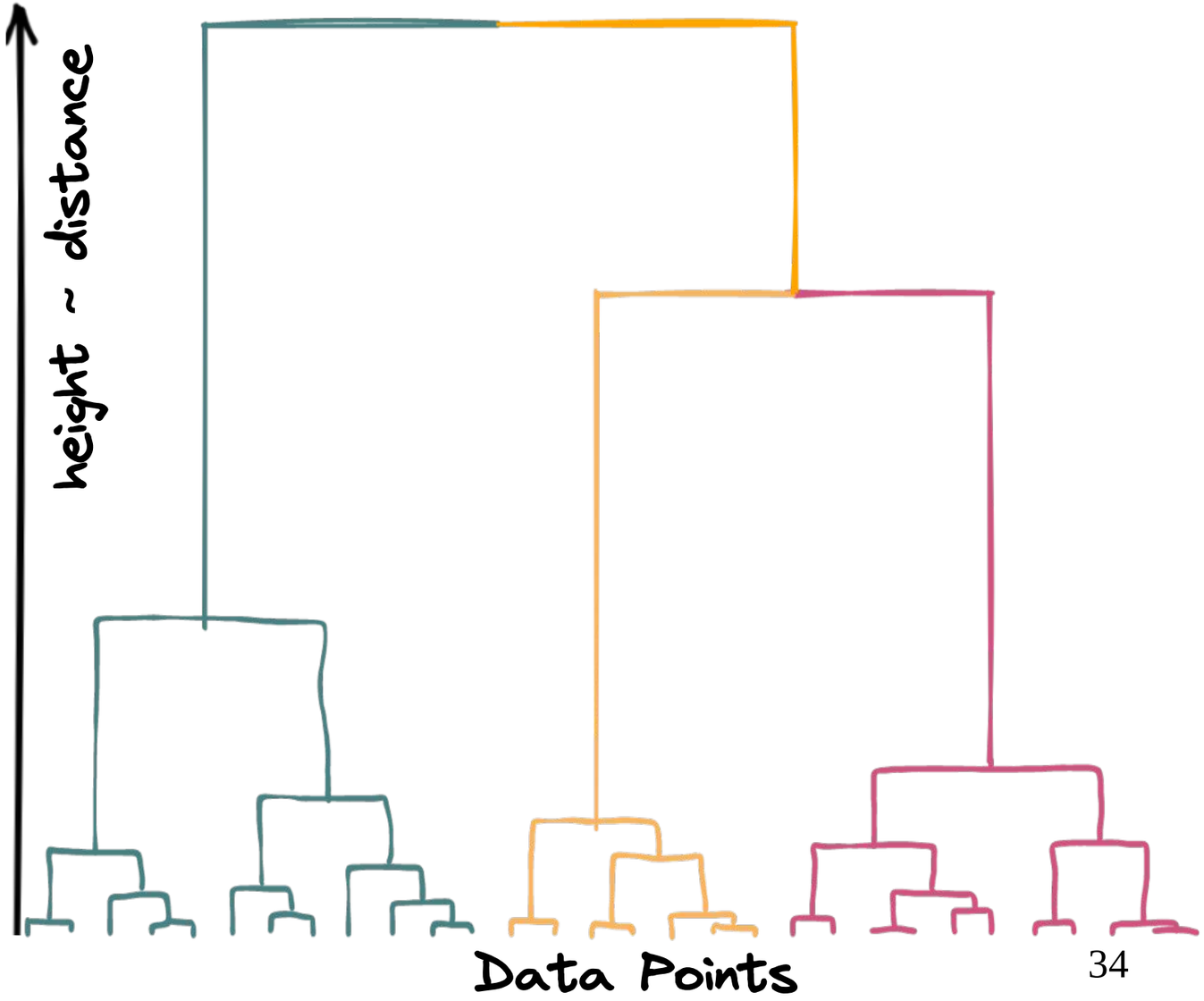


Distance-based clustering

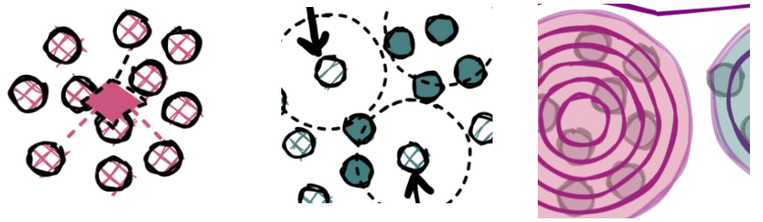


Distance-based clustering



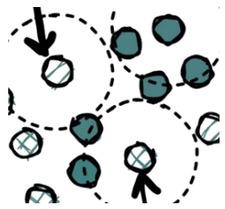
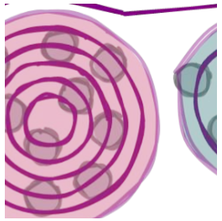
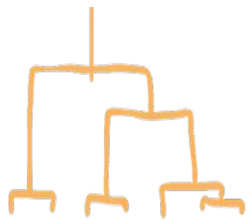


Advantages



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