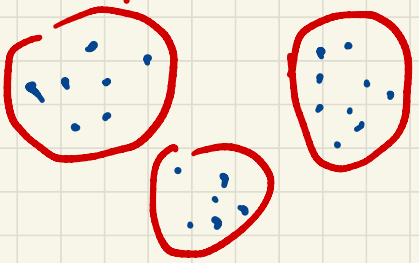
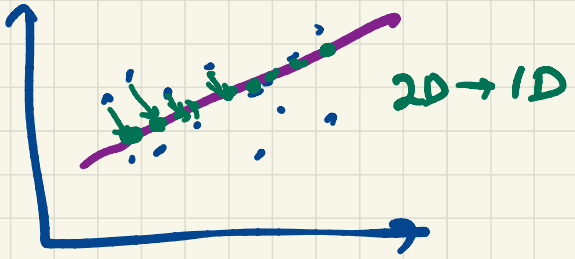
Unsupervised Learning → Data has no labels

Goal: Discover structure & patterns in data.

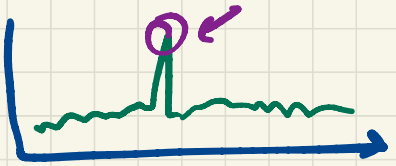
① Clustering



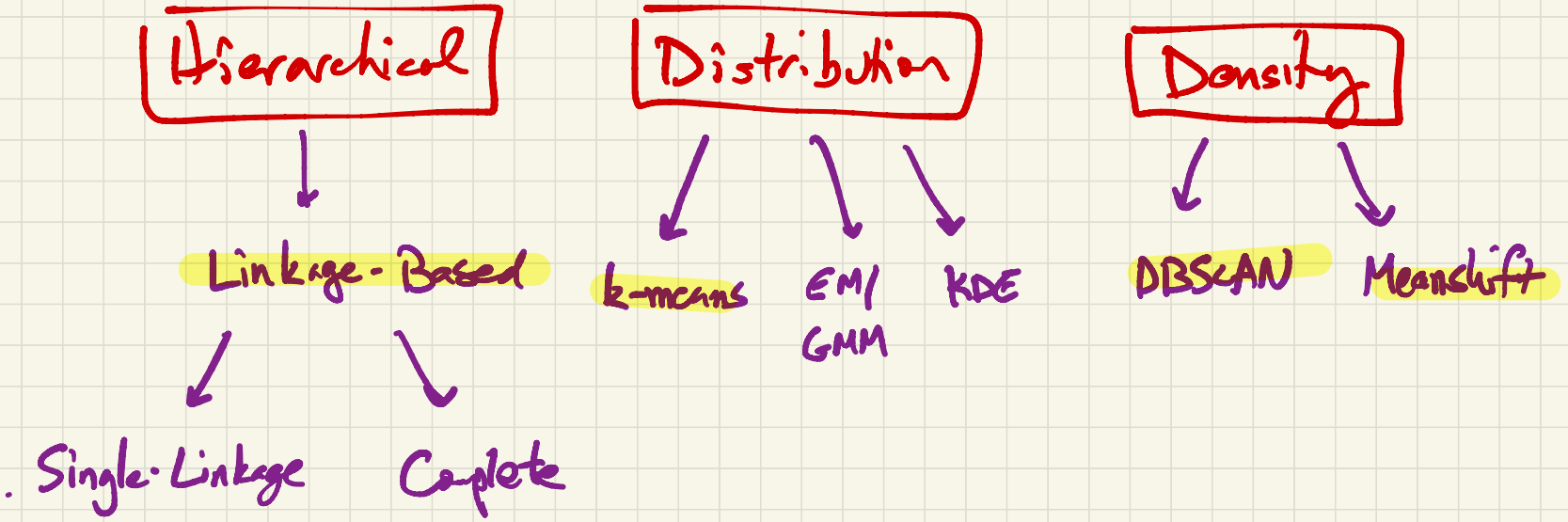
② Dimensionality Reduction



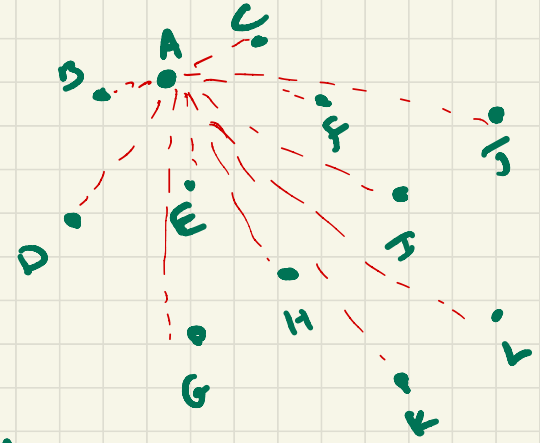
③ Outlier Detection



Clustering


Hierarchical
Distribution
Density
Linkage-Based
k-means
EM/
GMM
KDE
DBSCAN
Meanshift
Single-Linkage
Complete

Linkage-Based Clustering (Hierarchical)

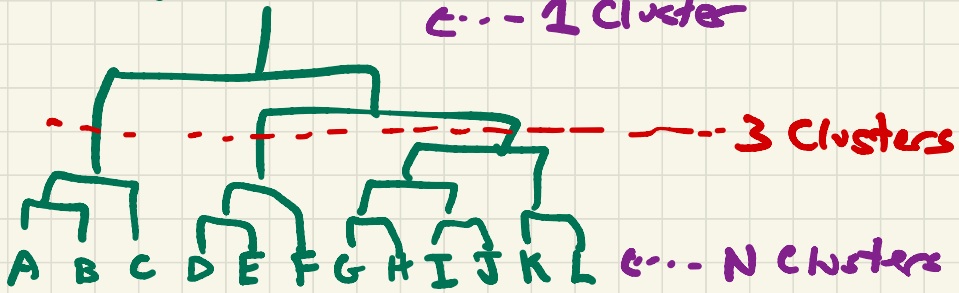


M_{ij} → M_{ij} = Distance b/w i & j

Linkage: Distance between two data points

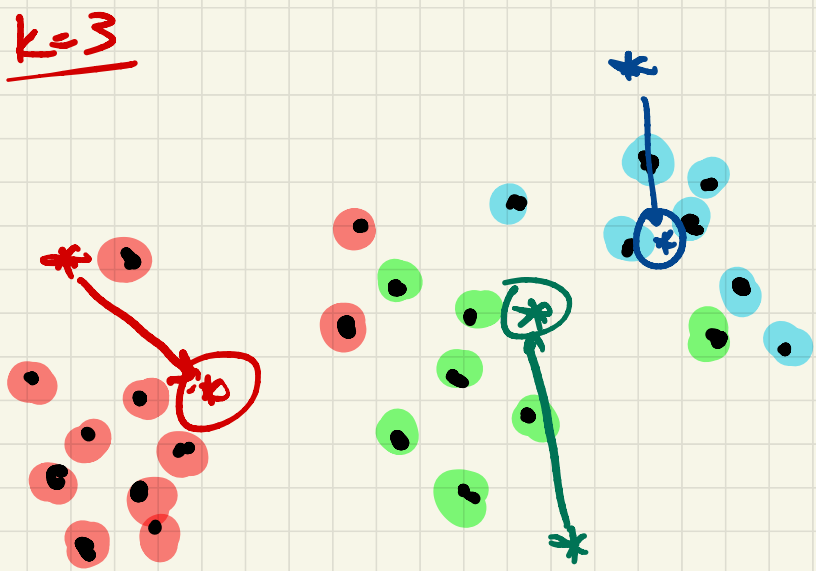
Single: Given clusters, join two clusters w/ closest pair of points

Dendrogram

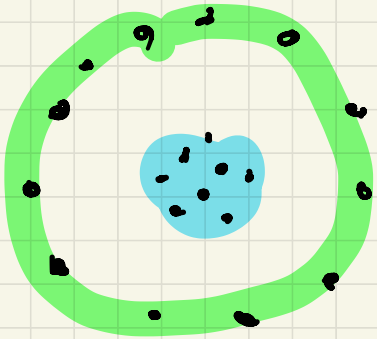


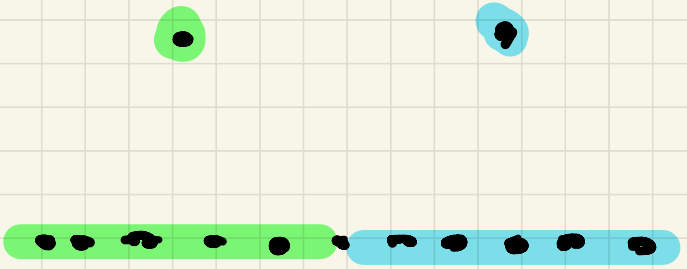
k-means Clustering

1. choose k
2. choose cluster centers
3. assign points to cluster based on closest cluster center
4. recompute cluster centers
5. repeat steps 3 & 4 until convergence



Failure Modes



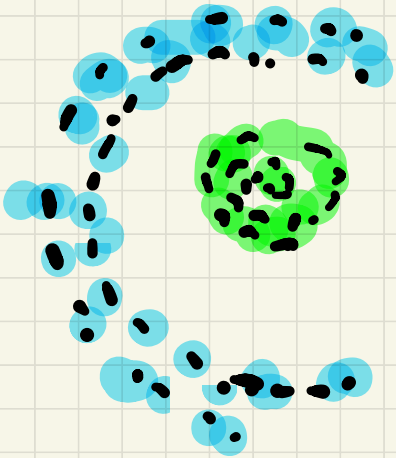


Problem: (1) Clusters don't have "centers"
(2) Clusters not hierarchical

Intuition: Similar points are "near" one another.

Solution: Density-based clustering.

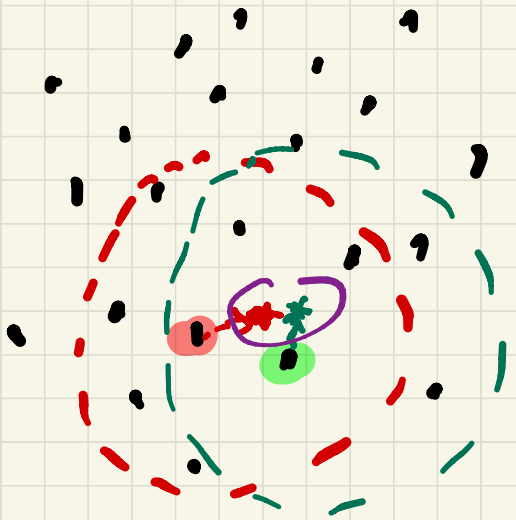
Density-Based Clustering: DBSCAN



1. Select point (random)
2. Declare all points in "neighborhood" to be in cluster.
3. Repeat step 2 for all points in cluster.

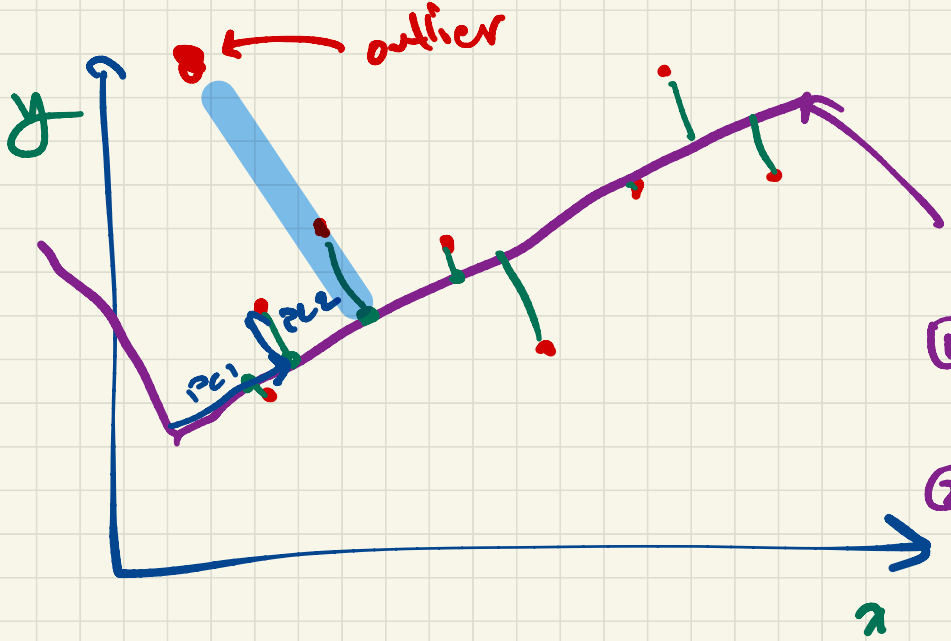
4. When no more points in neighborhood of cluster, go to step 1.

Density-Based Clustering: Mean-Shift



1. Every point is a cluster center
2. Each point has "radius" (bandwidth)
3. Mean of data points within each radius is a cluster center.
4. Move radius around new center
5. Repeat until centers do not move.

Dimensionality Reduction → Principal Components Analysis (PCA)



Vector is in the direction that
① captures highest variance.
② minimizes distance of projection.