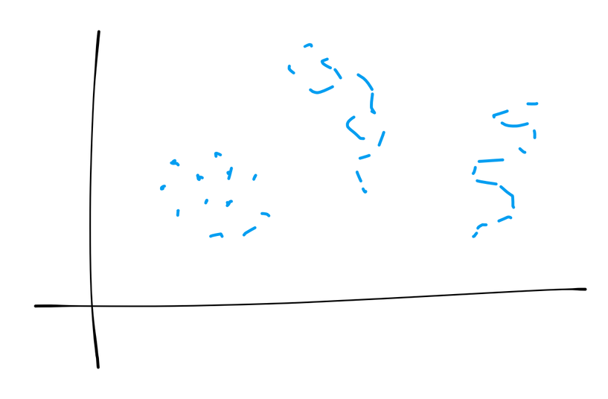
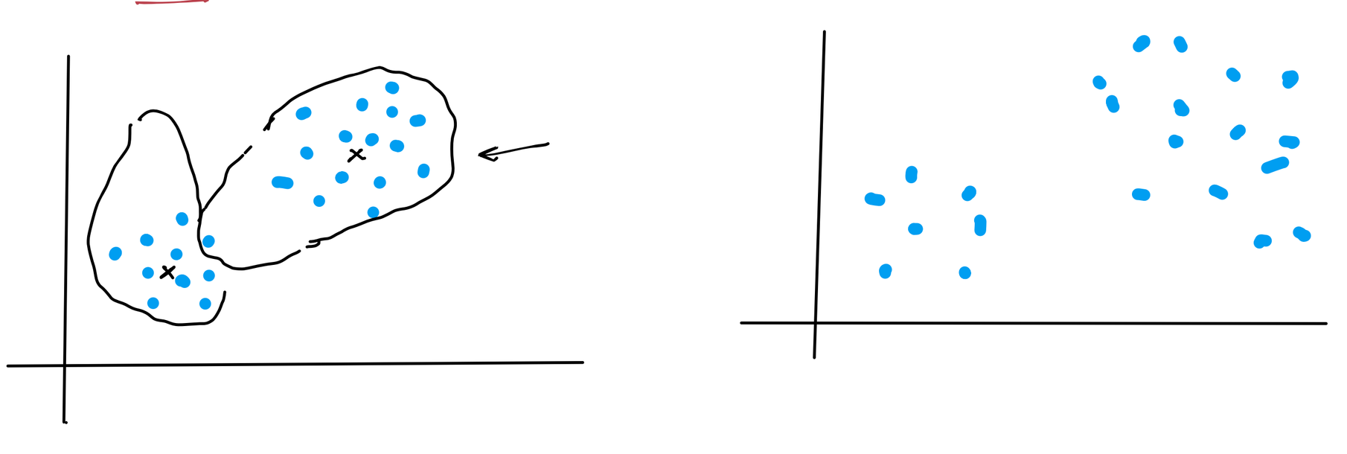
Clustering

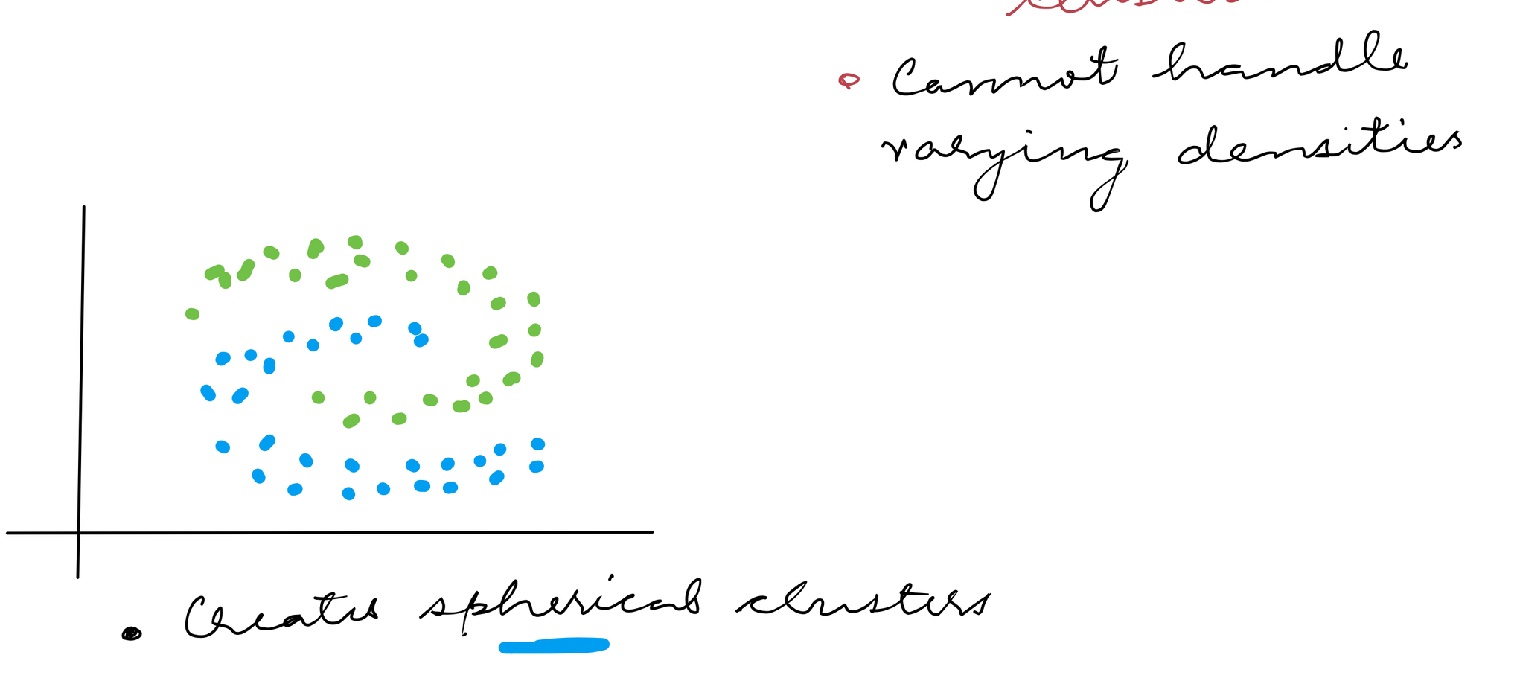


No labels
Supervised → X, Y
↳ X → Y

K-Means clustering

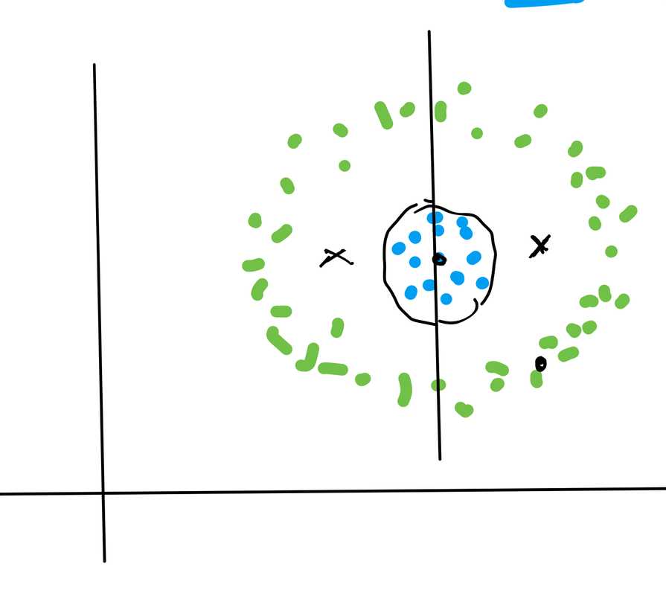


1. K clusters
2. Centroids ← Initialize
   ↳ Take the average of each dimension for all the points in the cluster.
4. Recompute Centroids
   ↳ Mean
3. Compute the distance of each point to each centroid. Assign it to the closest
5. Repeat till centroids stop updating or a preset number of iterations have passed.

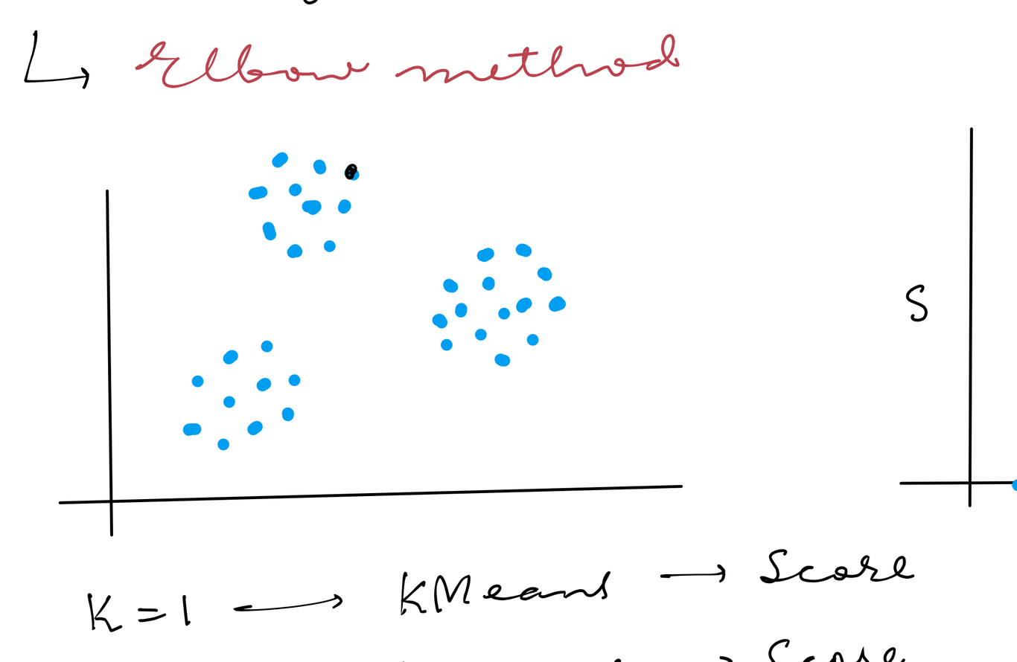


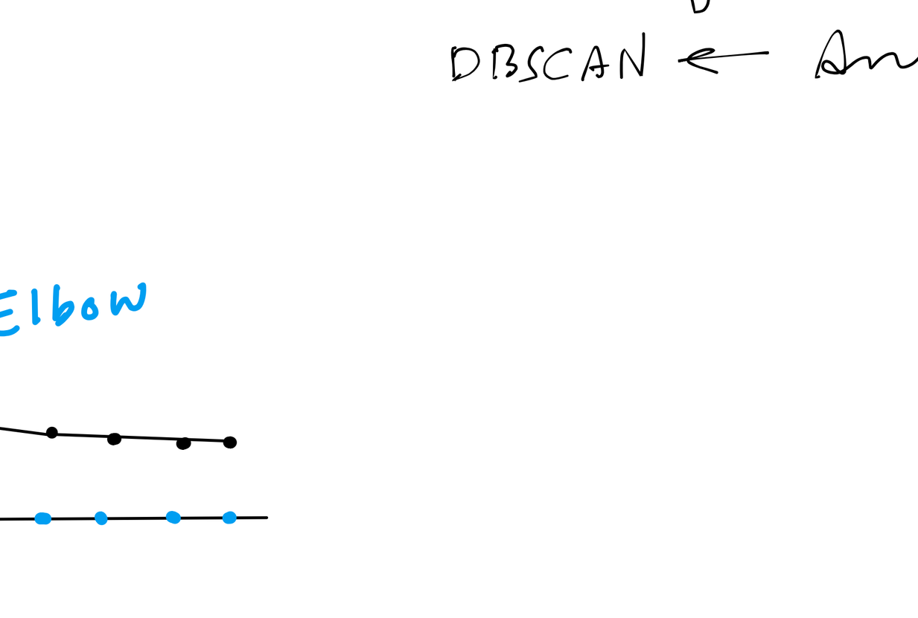
Incorrectly clustered

- Cannot handle varying densities



- Creates spherical clusters
- Creates equal partitions

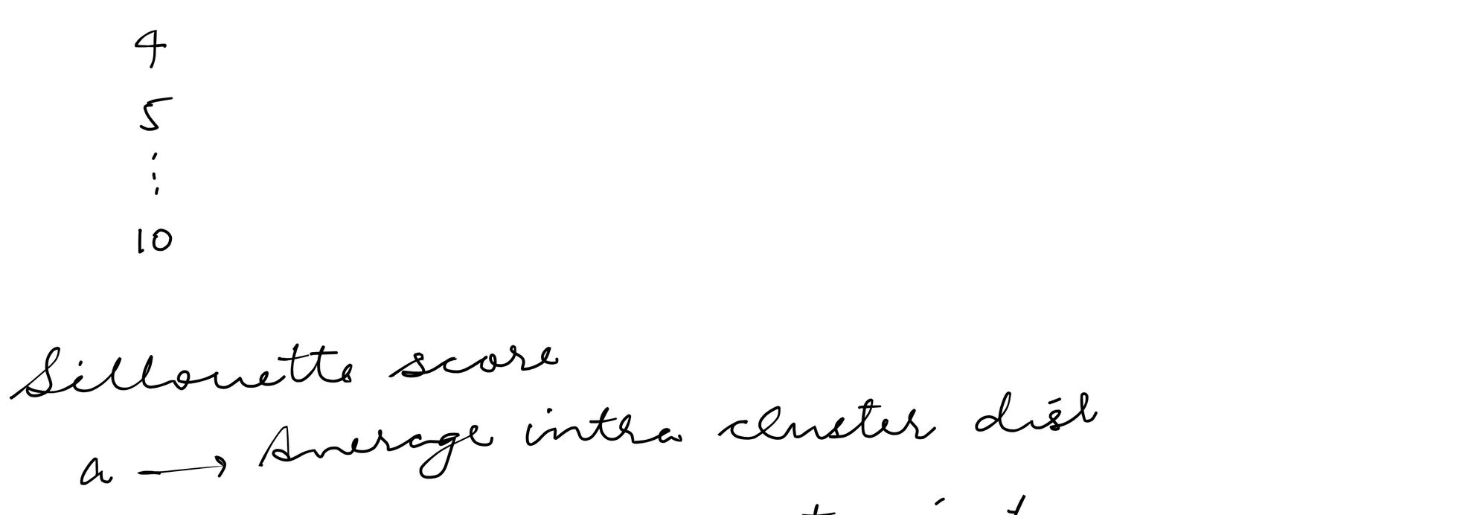




Non-core
↳ Non-DR

KMeans } Cluster vs Anomaly
HC }
DBSCAN ← Anomaly

How do you select K?
↳ Elbow method



K=1 → KMeans → Score
K=2 → KMeans → Score
K=3 → → →
4
5
⋮
10

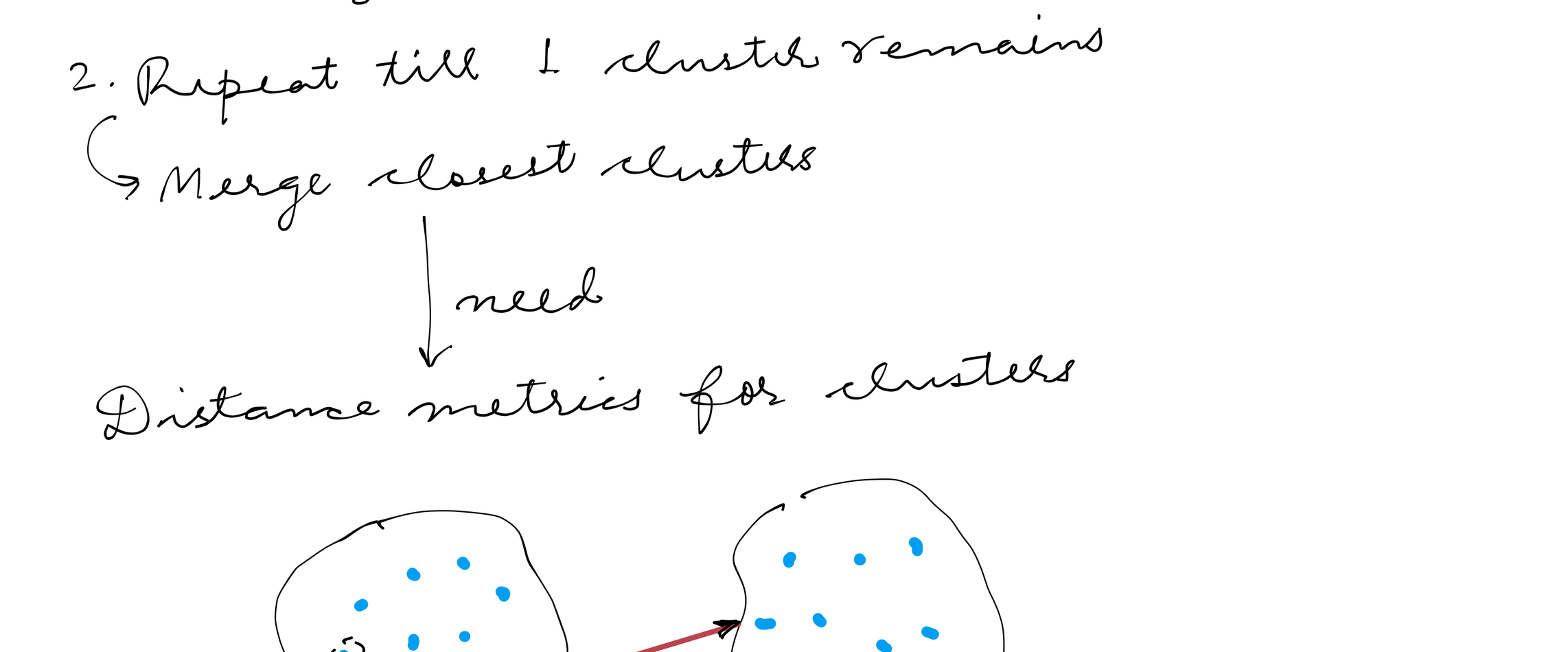
Sillouette score
a → Average intra cluster dist
b → Distance to closest point from another cluster

$$\frac{b - a}{\max(a, b)} \rightarrow 1$$

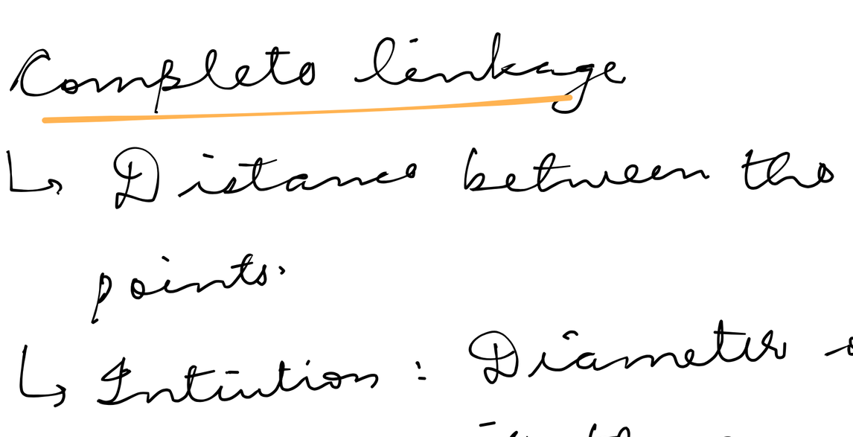
$\frac{\text{Avg. intra clust.}}{\text{Avg. inter clust.}}$ ← ↓ ↑ } ↓

— o —

Agglomerative Hierarchical clustering



1. Initialize each point as a seperate cluster.
2. Repeat till 1 cluster remains
   ↳ Merge closest clusters
   ↓ need
   Distance metrics for clusters



1. Single linkage
   ↳ Distance between the closest 2 points in the clusters.
2. Complete linkage
   ↳ Distance between the 2 farthest points.
   ↳ Intuition: Diameter of the cluster if the smaller 2 are merged.
3. Average linkage.
   ↳ Average the distance between all pairs of points in the two clusters.

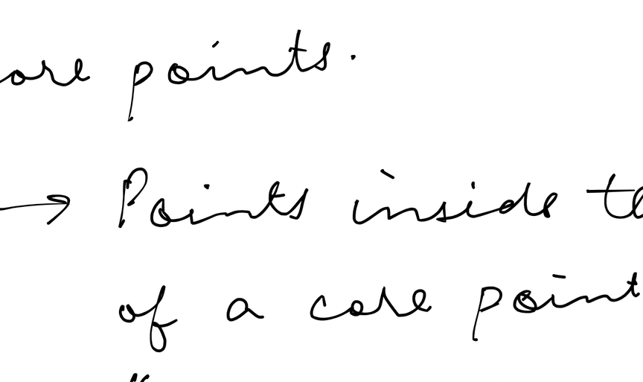
Limitations
↳ Where to cut the dendogram?
↳ Which linkage to use?
↳ Expensive

— o —

DBSCAN

↳ You don't have to specify the #clusters

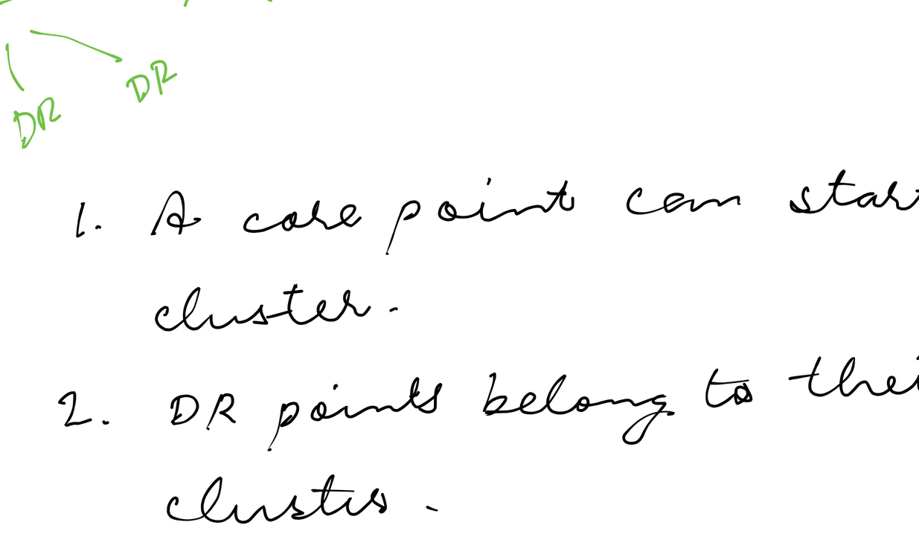
Terminology
ϵ : radius
minSup : #neighbours to label a point as core point.



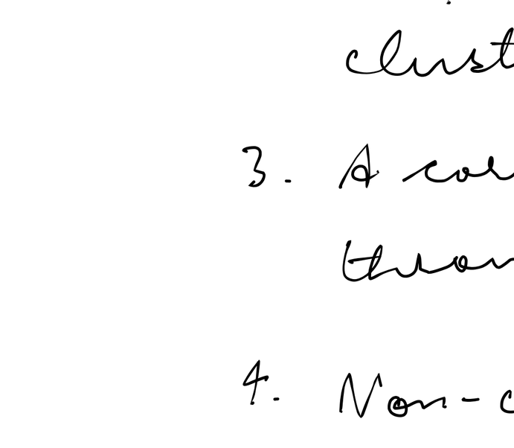
If a point has more than minSup points in its ϵ-neighborhood
↳ It is a core point.

① Core points.
↳ Points inside the ϵ-neighborhood of a core point are said to be
② "Density Reachable" (DR)

③ Density Connected (DC)
↳ If a "density reachable" point is itself core point, it is said to be "density connected."



What if a point is DR from 2 diff clusters?



1. A core point can start its own cluster.
2. DR points belong to their core point's clusters.
3. A core point expands its cluster through DC points.
4. Non-core, non DR points are anomalies.