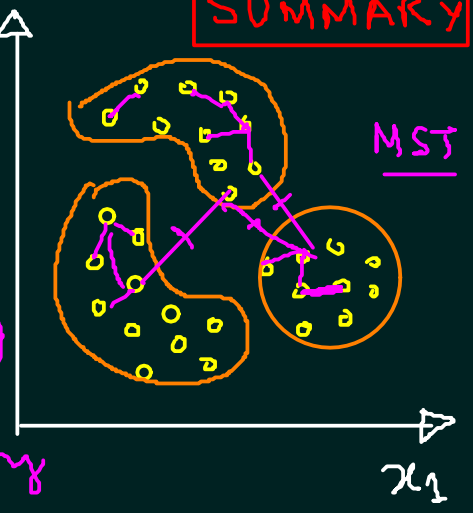
▷ Clustering (Unsupervised Learning) :

SUMMARY

↳ Hierarchical Clustering

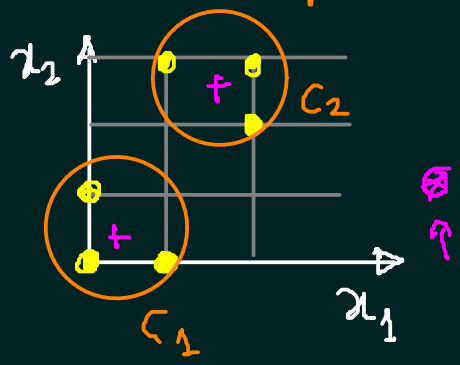
- Agglomerative Clustering (bottom up)
  - Single Linkage (min)
  - Complete Linkage (max)
  - Avg. Linkage (avg)
- Divisive Clustering (top-Down)

▷ dist metrics
- ↳ Symmetry
- ↳ Reflexivity
- ↳ Triangle Inequality



↳ Partitional Clustering

↳ K-Means Clustering



B_1: (0,0), (1,0)

B_2: (0,1), (1,3), (2,2), (2,3)

$m_1' = (0.5, 0)$

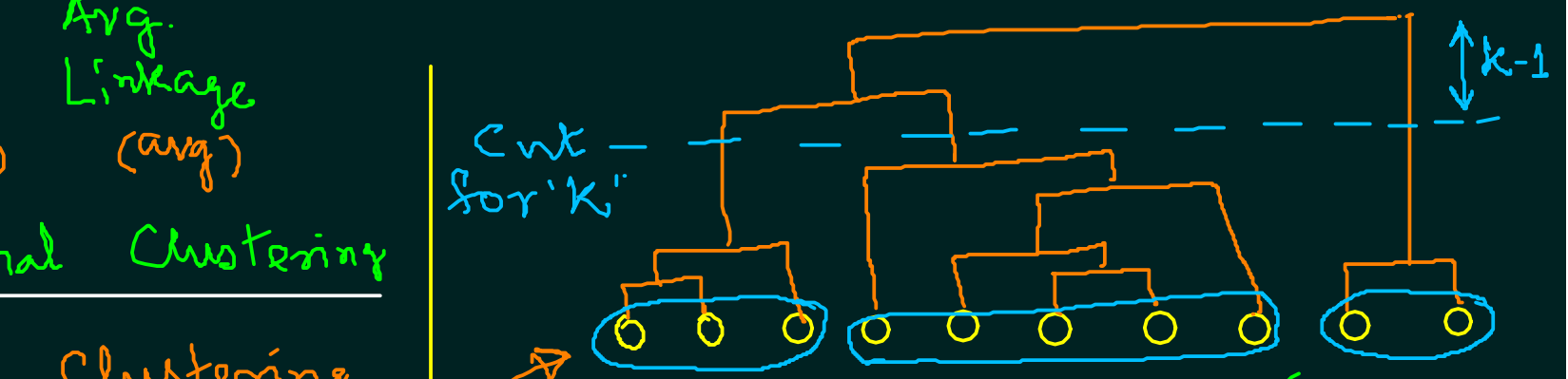
$m_2' = (1.25, 2.25)$

$[\ldots]_{B_1}$ $[\ldots]_{B_2}$

$m_1'' = (1/3, 1/3)$

$m_2'' = (5/3, 8/3)$

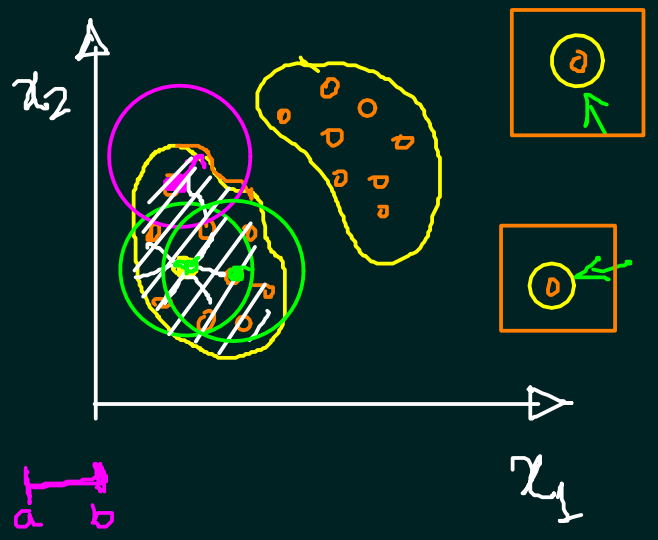
↳ Iteration Stops



Dendrogram (tree)

↳ Density Based Clustering ▷▷ TODAY

Evaluation of Cluster
↳ Scatter Coeff.



— dense points / core points

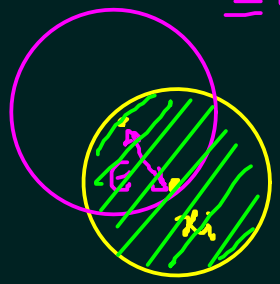
$x_i \geq [mp]$ [mp]

— Border points

— Noise points

core

$|x_j - x_i| < \epsilon$ $= R$



▷ Connectedness :

① Core points (x_i)

② $|x_j - x_i| < \epsilon \iff x_i \rightarrow x_j$: directly connected

③ $x_i \rightarrow x_{k_1} \rightarrow x_{k_2} \rightarrow \cdots\cdots \rightarrow x_{k_m} \rightarrow x_j$: path connected

$< \epsilon$

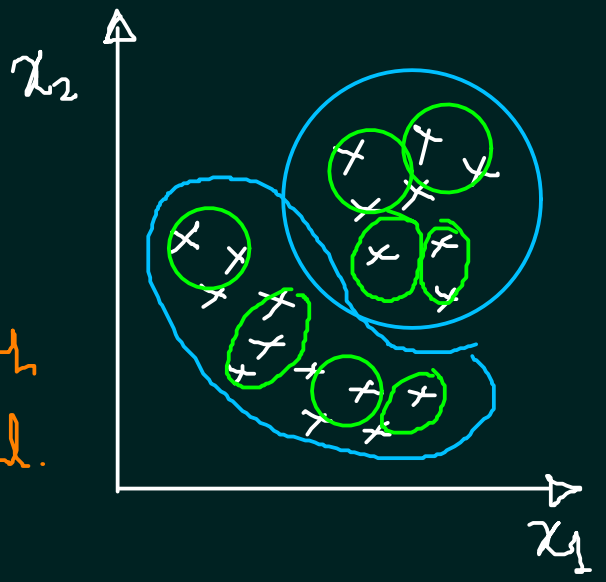
Algo: → A { All points that are path connected from Core points form cluster

B Outlier detection

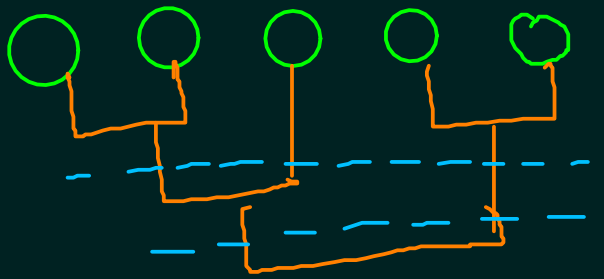
DBSCAN

Hybrid Clustering Algo:

→ Partitional
→ Hierarchical



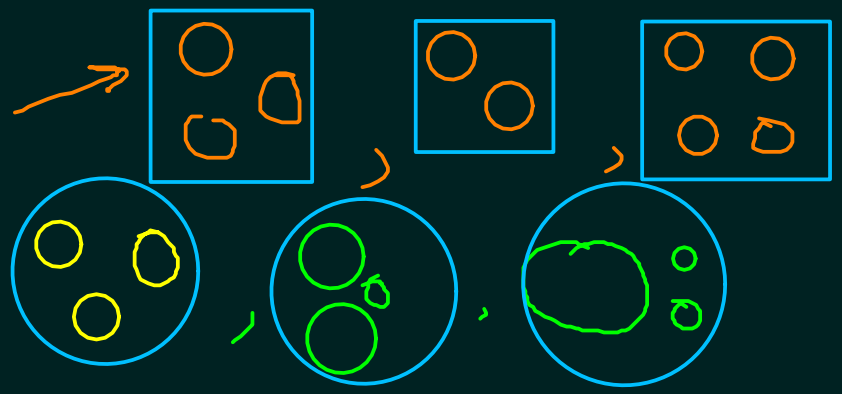
- K-Means apply with high "K".
- $k' << k$ and aggl. clustering with k' as final.
- Single Linkage K-Means (CLARA)



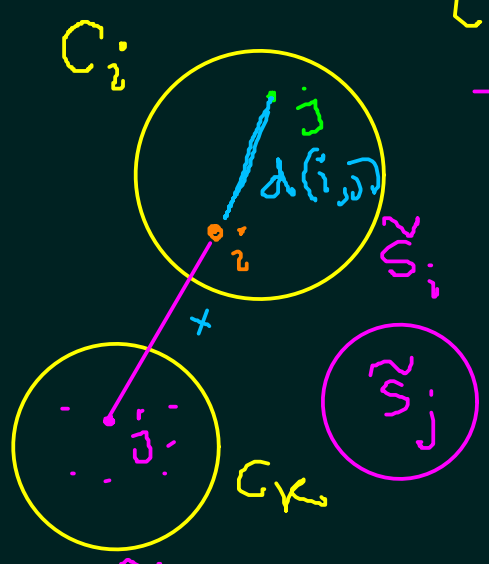
Experimental Clustering:

- $k = ?$ → 

- Hyper Parameters chice →



- How do we evaluate Clustering Soln?
  - Scatter, Silhouette Coeff.



$C_1, C_2 \dots C_m$

→ Cohesion (dissimilarity)

$\forall i \in C_i$

$a(i) = \frac{1}{|C_i|-1} \sum_{\substack{j \in C_i \\ j \neq i}} d(i,j)$ ∝ $\frac{1}{s}$ ← w

sim →

$\sum s(i,j) / (|C_i| - 1)$

→ separation:

$\forall i \in C_i \rightarrow b(i) = \min_K \left(\sum_{j \in C_K} d(i,j) / |C_K| \right)$

▷ Silhouette Value of i : $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$

$|C_i| = 1$

$s(i) = 0$

$b(i) >> a(i) \rightarrow s(i) = 1 - \frac{a(i)}{b(i)}$

$b(i) << a(i) \rightarrow s(i) = \frac{b(i)}{a(i)} - 1$

$-1 \leq s(i) \leq 1$

0

0

$\forall i \in C_i \rightarrow s(i) \Rightarrow \tilde{s}(i) =$ mean of all $s(i)$

▷ Silhouette Coeff → $\left[\max_K \tilde{s}(k) \right]$

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