L8:

Hierarchical Agglomerative Clustering

Sep 15, 2025
Data Mining

Jeff M. Phillips

When data is well clusterable, most clustering algorithms / formulations will work well.

When data is not well clusterable, no clustering algorithm / formulation will find a good clustering.

Clustering

Input

Data $X \subset \mathcal{M} (= \mathbb{R}^d)$

Distance $D : \mathcal{M} \times \mathcal{M} \to \mathbb{R}_{\geq 0}$

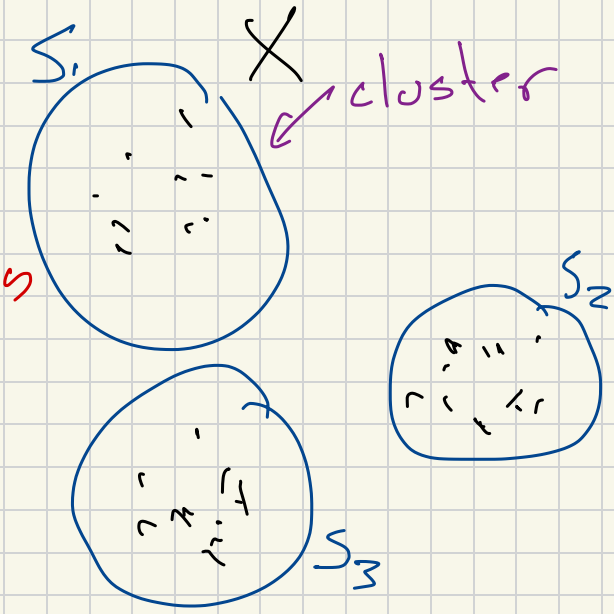
choice → parameters (e.g. # clusters k)

Output

clusters

$\mathcal{C}(X) = \{S_1, S_2, \dots, S_k\}$

- each $S_j \subset X$
- each pair $j \neq j'$ $S_j \cap S_{j'} = \emptyset$ hard clustering
- $\bigcup_{j=1}^{k} S_j = X$ (o.w. outliers)

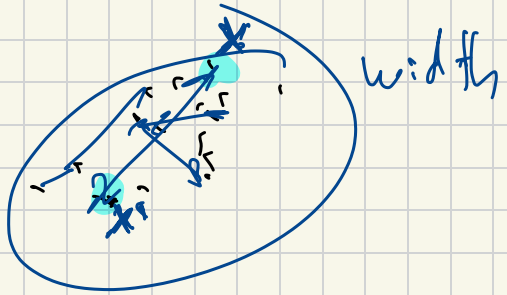


What makes a good clustering?

width For each S_j, $x, x' \in S_j$: $D(x, x')$ small

split For $S_j, S_{j'}$ (most) $x \in S_j$ $x' \in S_{j'}$:

$D(x, x')$ large

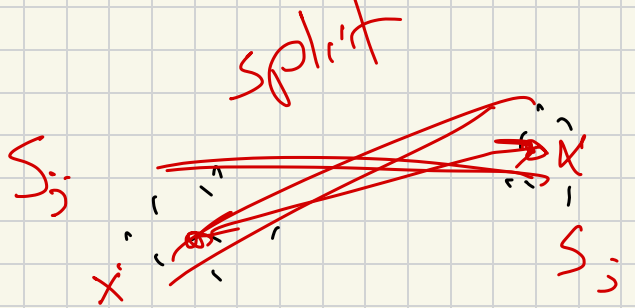


overall

width − split

or

$\frac{\text{width}}{\text{split}}$ small



Hierarchical Agglomerative Clustering (HAC)

If two clusters (points) are close, → merge them

HAC

0. Each $x_i \in X$ in own cluster S_i
1. while (two clusters are close enough)
   - Find closest pair S_i, S_j
   - Merge $S_i, S_j \rightarrow S' = S_i \cup S_j$

$D(S, S')$

$S, S' \subset X$

u = represent(S) u' = repres(S')

① u = mean(S) u' = mean(S')

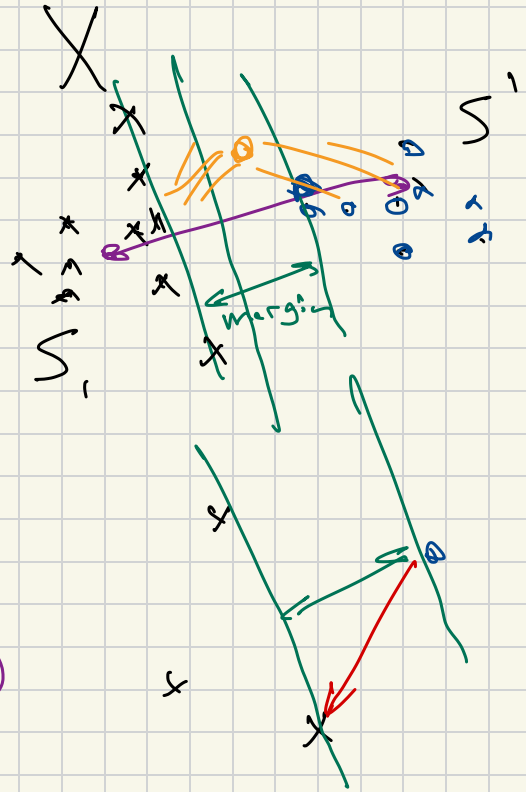
$$D(S, S') = D(u, u')$$

② $$\frac{1}{|S|} \sum_{x \in S} \left(\frac{1}{|S'|} \sum_{x' \in S'} D(x, x') \right)$$

$$= \frac{1}{|S|} \frac{1}{|S'|} \sum_{x \in S} \sum_{x' \in S'} D(x, x')$$

③ $$D(S, S') = \min_{x \in S,\, x' \in S'} D(x, x')$$

④ $$D(S, S') = \max_{x \in S,\, x' \in S'} D(x, x')$$



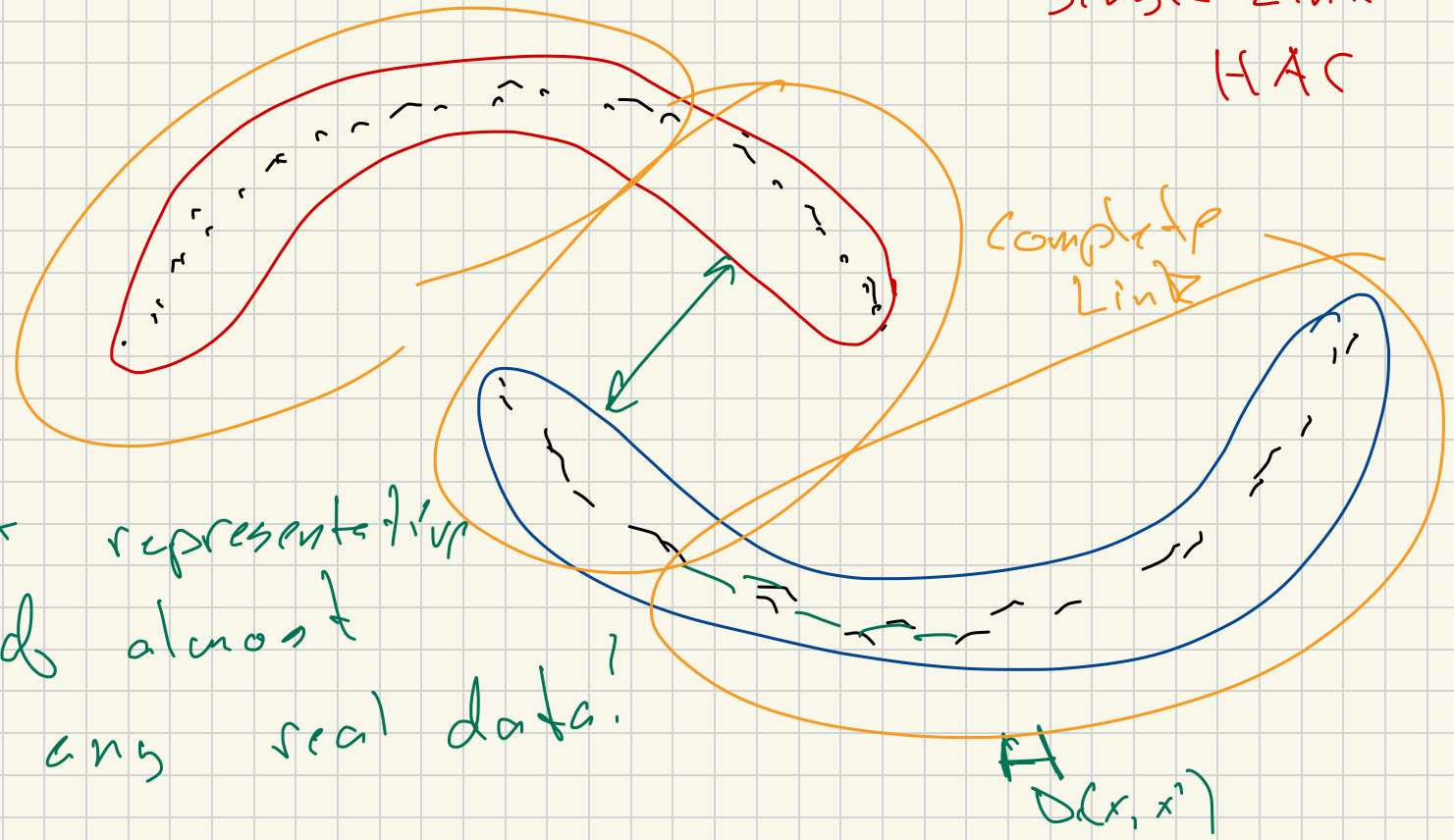
[single Link]

[complete Link]

Two Moons Data Set

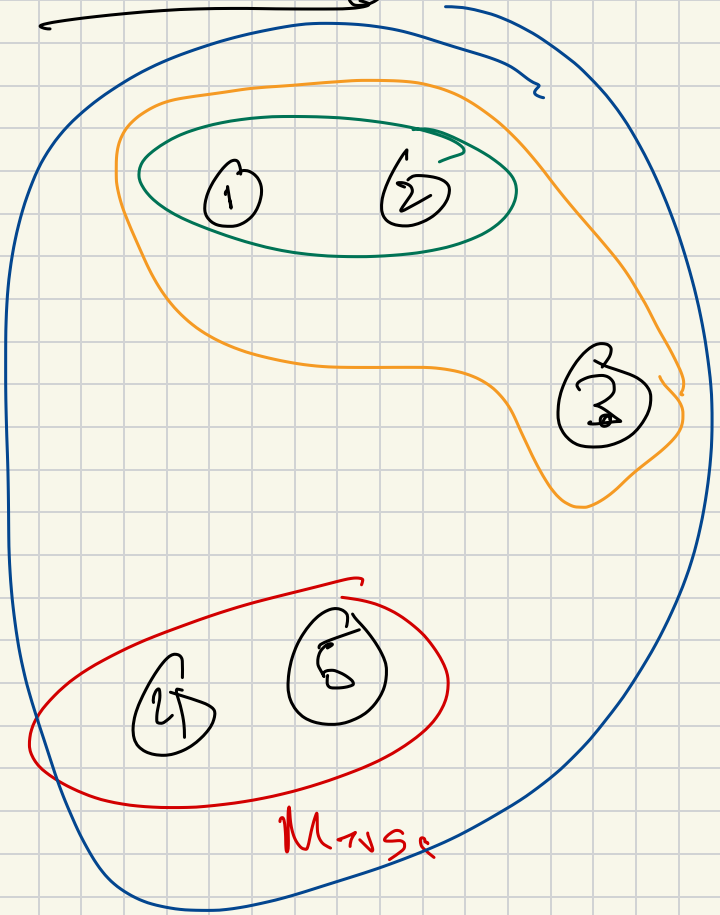
Single Link HAC

X

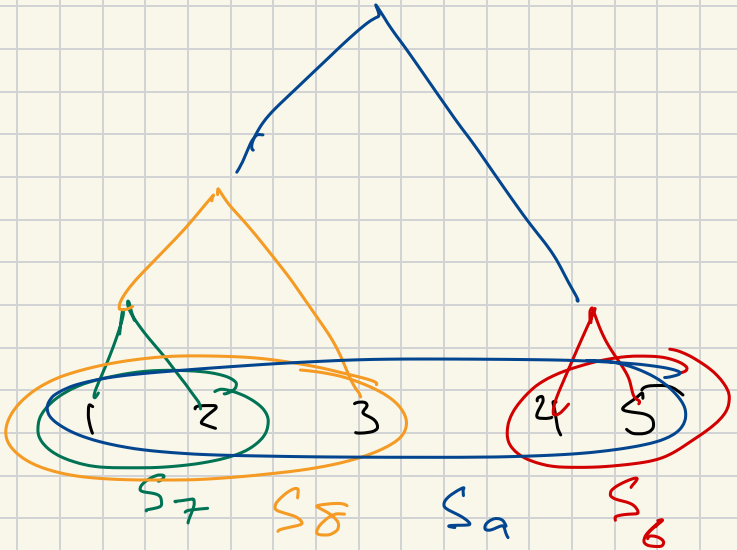


Not representative of almost any real data!

Hierarchy



Single Link Hierarchy



Phylogenic Trees

How efficient is HAC?

n rounds of while loop.

- $D(s_i, s_j)$ $O(n^2)$ if $(|s_i|, |s_j|) = (a, b)$
  - maybe $O(n^2)$ pairs
  - $O(n^2)$

→ overall $O(n^3)$

representative-based approach

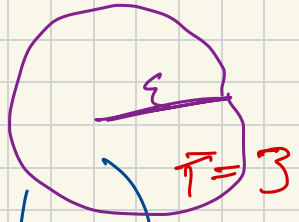
↳ $O(n^2 \log n)$

Density-based clustering (DBScan)

Two-Parameter

ε = radius (what is close)

τ = MinPts (density parameter)



1. Find core pts (w/ τ pts in ε-radius ball)
2. Edges
   - $\text{dist}(x, x') \le \varepsilon$
   - x or x' is core
3. clusters are connected components