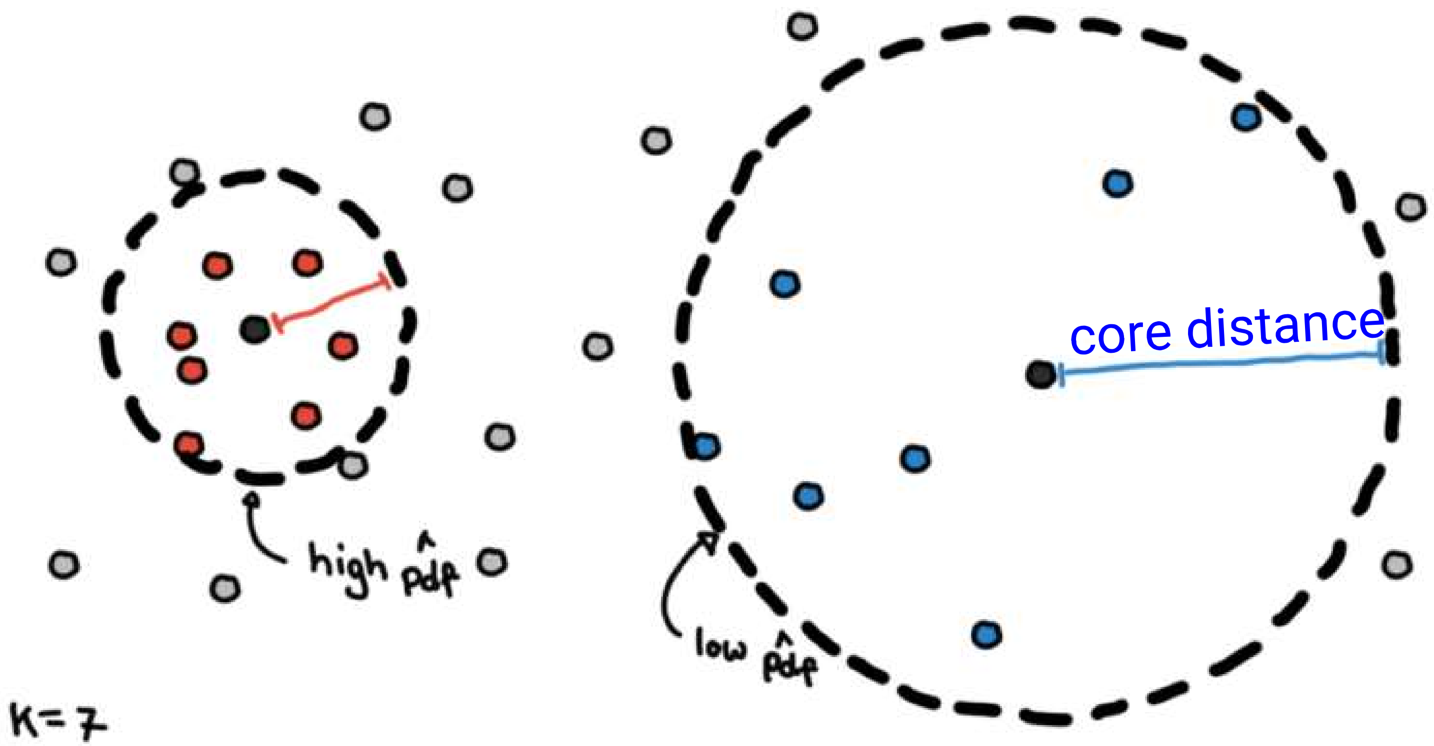
HDBSCAN

Hierarchical DBSCAN

Estimating densities



Core distance with min_samples =7

No need to specify epsilon.

類似於DBSCAN, 但只需要用一個參數 min_samples 來衡量密度
即找到 min_samples 個人的半徑 , 此稱為 Core distance.
故密度的衡量為1/ Core distance

How does HDBSCAN do this?

At a high level, we can simplify the process of density-based clustering into these steps:

1. Transform the space according to the density/sparsity.
2. Build a minimum spanning tree of the data using the mutual reachability distances as edge weights.
3. Build Hierarchical Tree (Condensed Tree):Sort edges of the MST by increasing weight and create a hierarchy of connected components.
4. Condense the Tree : Filter the tree to keep stable clusters by measuring how long clusters persist (i.e., how stable they are across distance scales).
5. Extract the stable clusters from the condensed tree.

1. Transform the space from data set

For each point, compute its core distance $Core_k(x)$

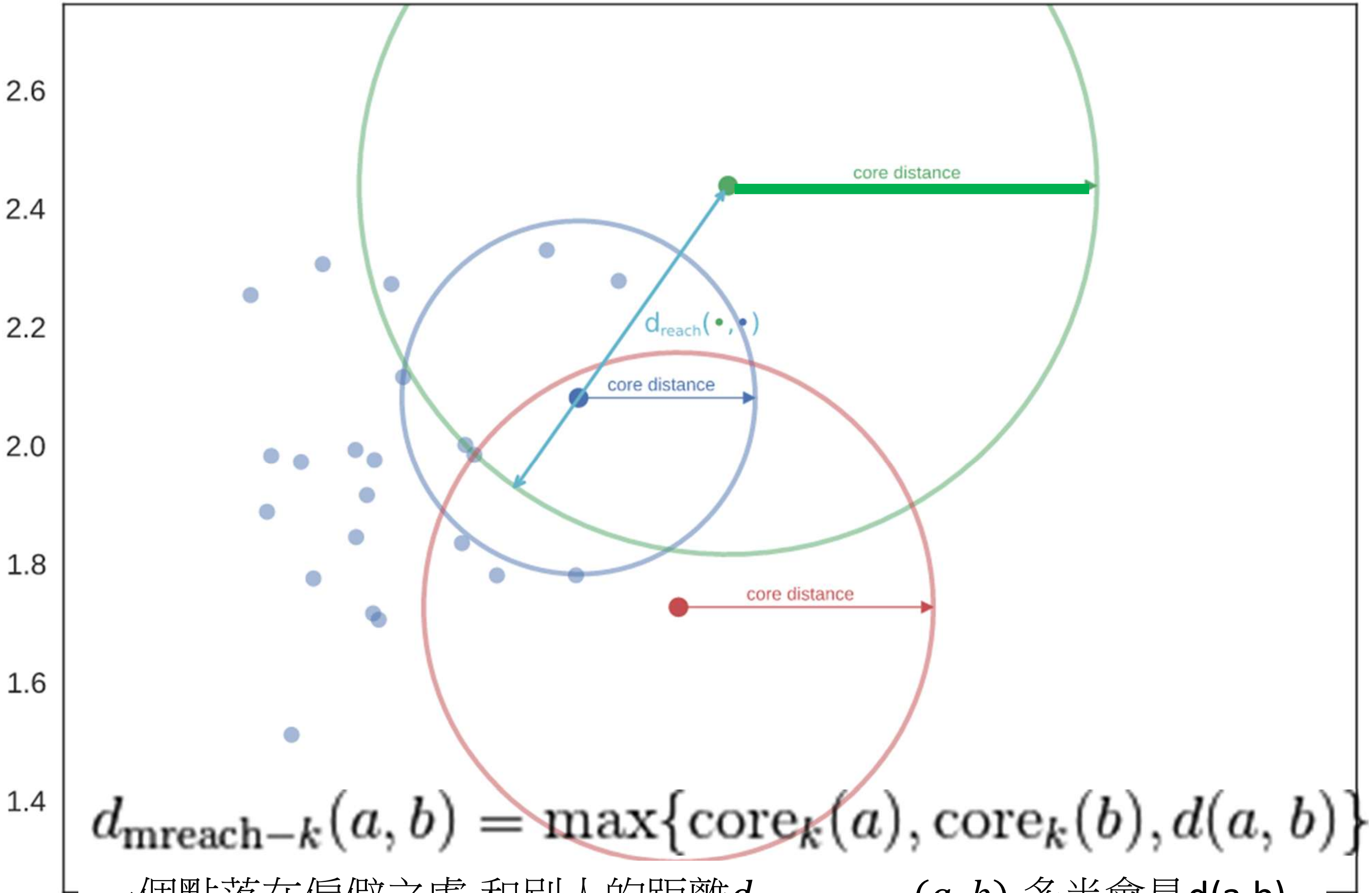
- The distance to its k-th nearest neighbor, where k = min_samples.
- Points in denser regions would have smaller core distances while points in sparser regions would have larger core distances.

Compute Mutual Reachability Distance

- For each pair of points (a, b), define the mutual reachability distance as:

$$d_{\mathrm{mreach}-k}(a,b) = \max\{\mathrm{core}_k(a), \mathrm{core}_k(b), d(a,b)\}$$

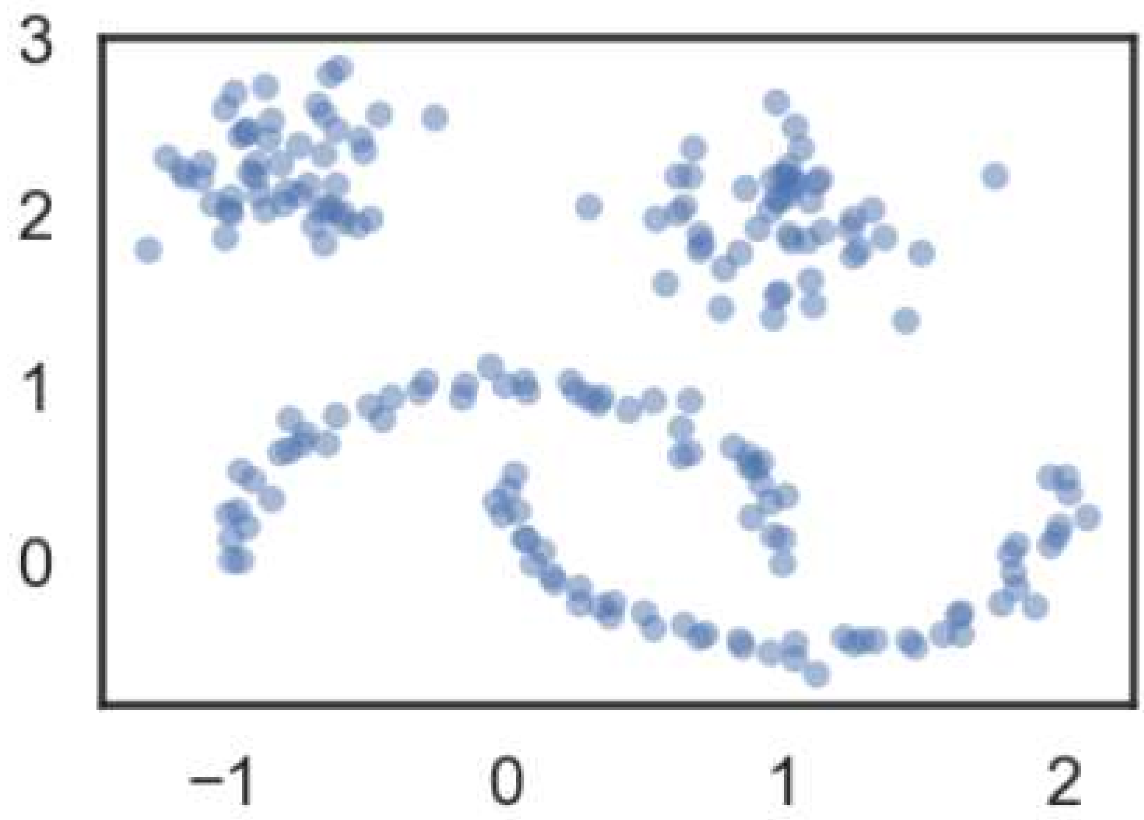
艾鍗學院 ITTRAINING



$$d_{\mathrm{mreach}-k}(a,b)=\max\{\mathrm{core}_k(a),\mathrm{core}_k(b),d(a,b)\}$$

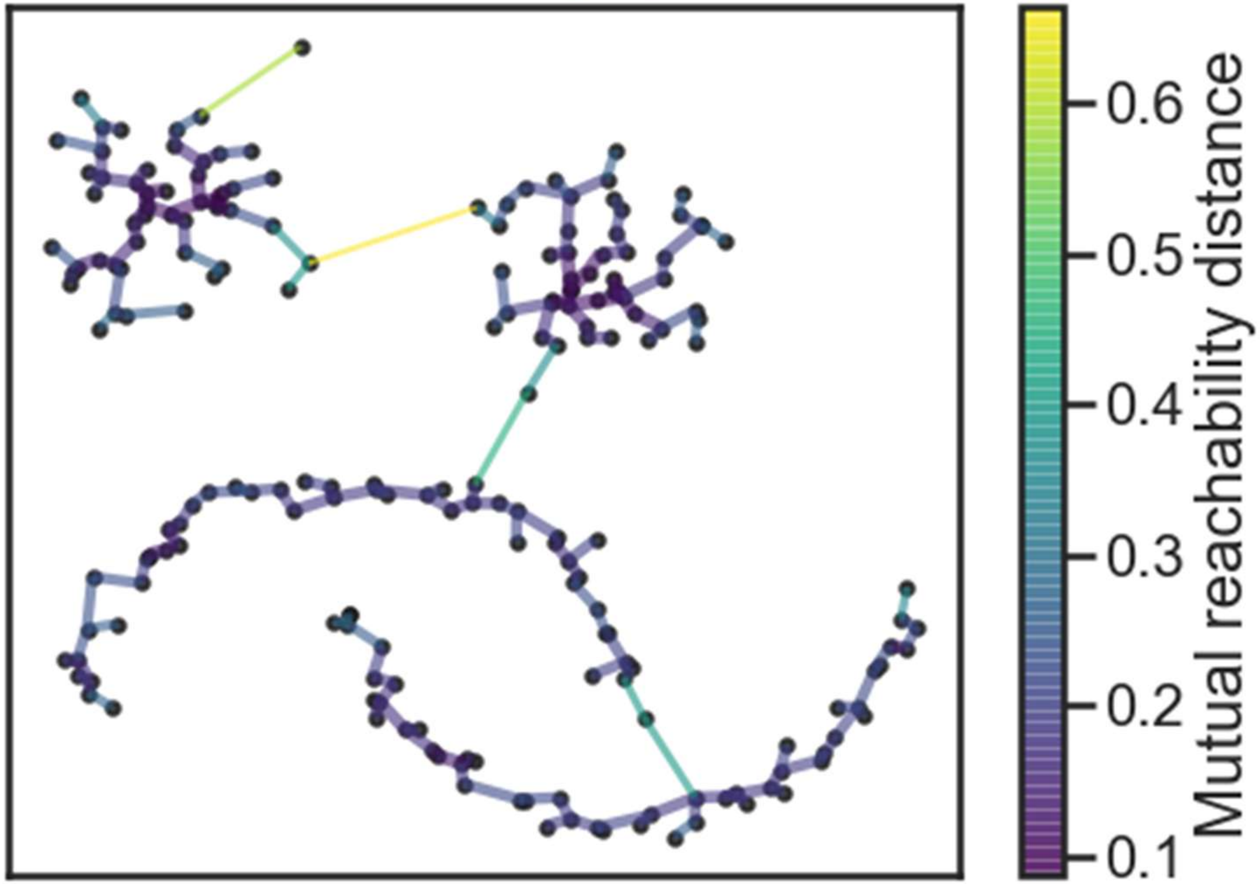
一個點落在偏僻之處,和別人的距離$d_{mreach-k}(a,b)$ 多半會是d(a,b),
而鬧區會是$core_k(a)$或$core_k(b)$

- 建立一個complete graph G(V,E), 其中兩點間的 mutual reachability distance 作為 edge weight



2. Build the minimum cost spanning tree

- **對此complete graph 建立 minimum cost spanning tree，即 |V|=n，|E|=n-1**

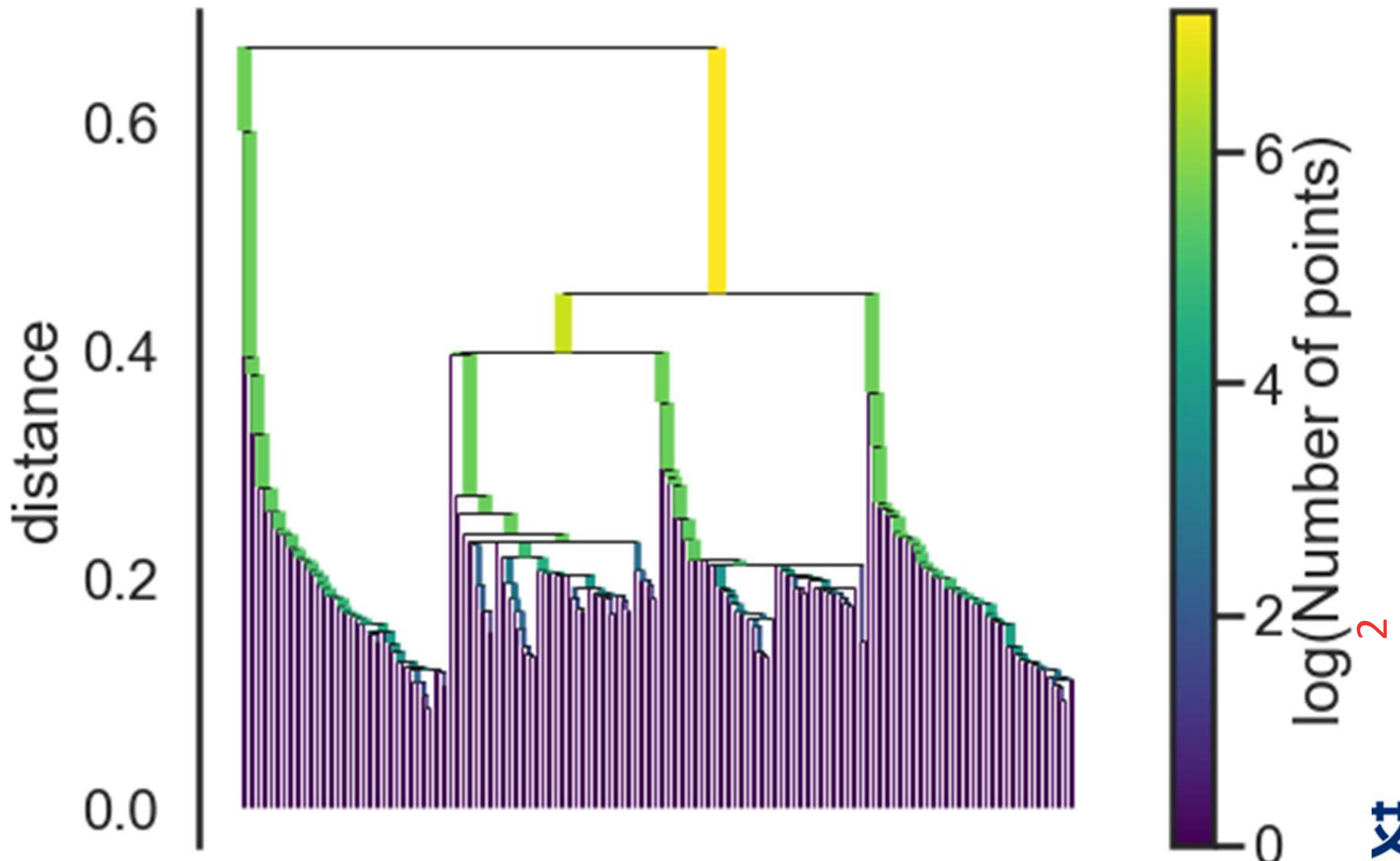


"一個點落在偏僻之處,和別人的距離$d_{mreach-k}(a,b)$ 多半會是d(a,b)……"
因而此graph所建立minimum cost spanning tree 會排除偏僻的點

3. Construct a cluster hierarchy

- Construct a cluster hierarchy of *connected components* → "Hierarchy clustering" (最小的兩邊先合併)

線的寬度代表集群中的點數

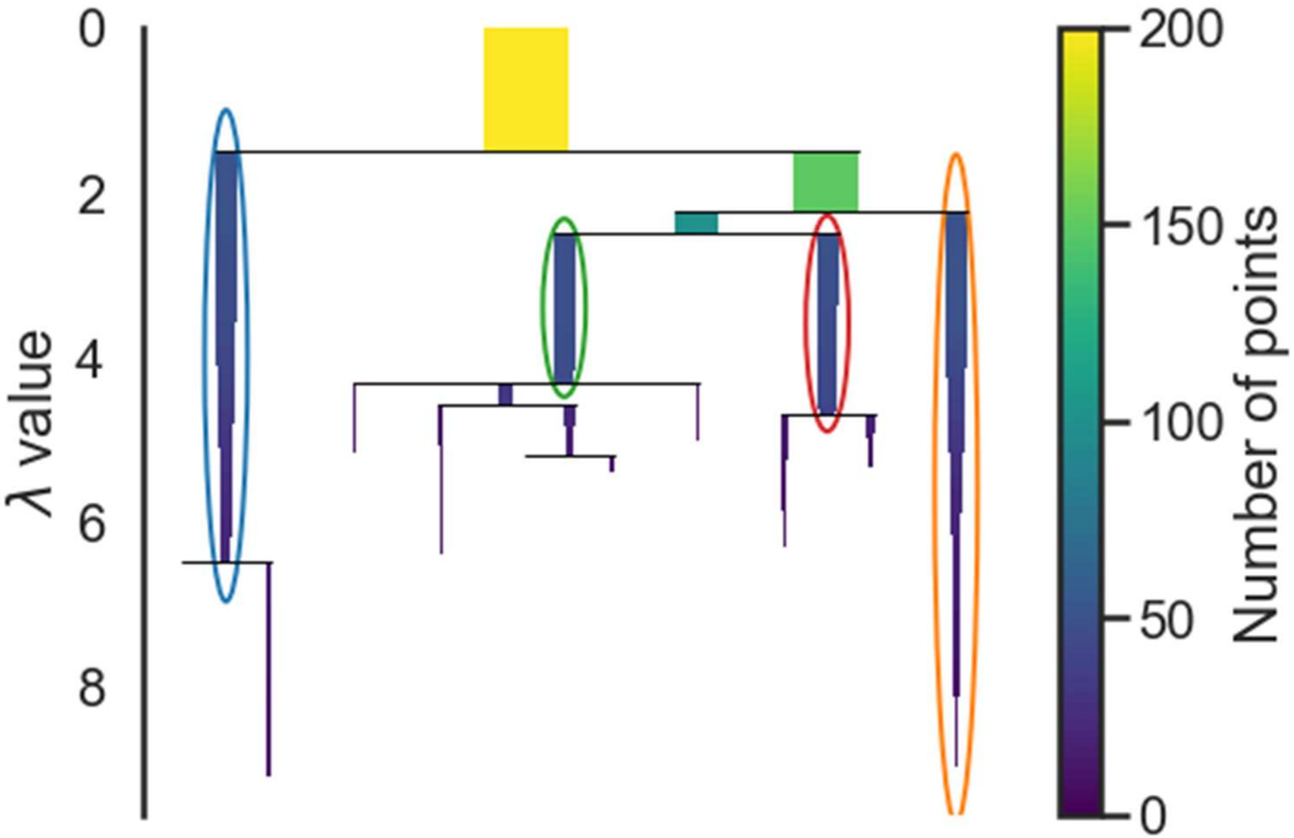


4.Condense the cluster hierarchy

- 濃縮**cluster hierarchy**圖, 將不足**min_cluster_size** 的**Cluster** 剔除, 則樹狀圖看起來比較簡潔 , 容易看出應該分幾群

$$\lambda = \frac{1}{distance}$$

愈後面合併, 一定是距離較遠, 故λ愈小; 反之.,λ 愈大



in this example, min_cluster_size is 5
線的寬度代表集群中的點數

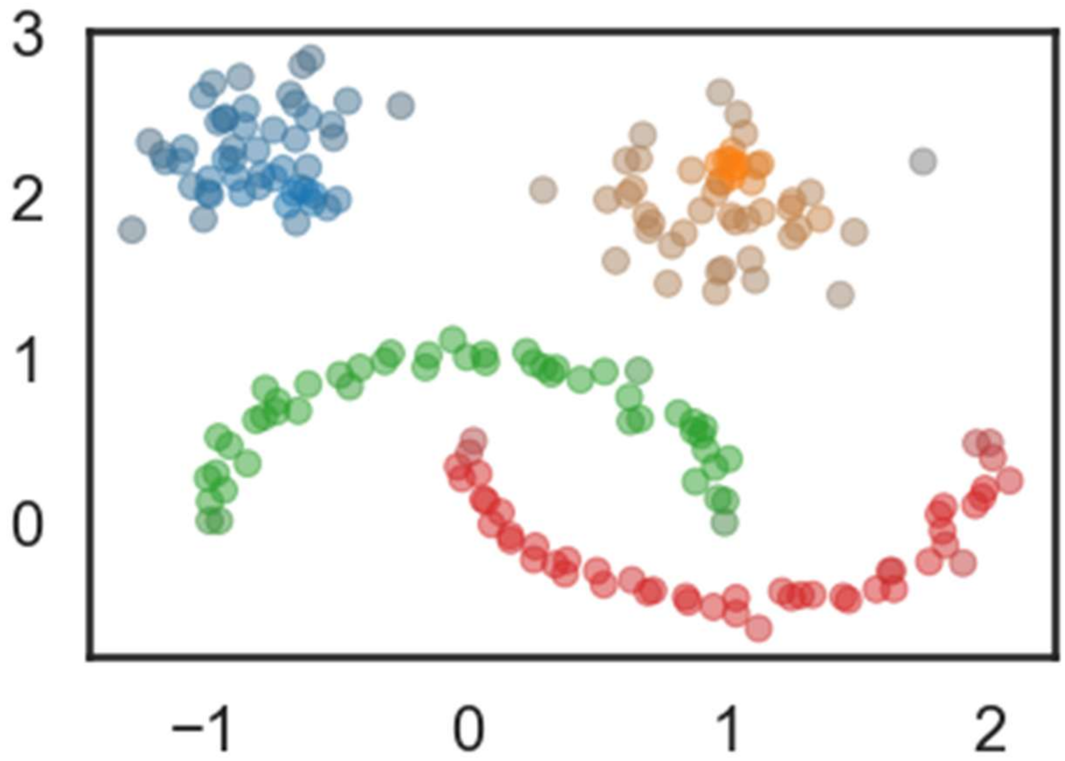
Key Parameters of HDBSCAN

- **min_cluster_size:** Minimum size of clusters. Controls granularity.
- **min_samples:** Similar to DBSCAN; affects core distance calculation. Can be set equal to min_cluster_size for simplicity.

Summary

Algorithm	Strengths	Weaknesses
DBSCAN	• Faster than the HDBSCAN algorithm. • Discovers the clusters in a dataset • Identifies outlier points.	• The algorithm requires an obscure, data dependent, distance parameter. • Not effective at identifying clusters of varying density.
HDBSCAN	• Identifies clusters of varying density (only one parameter) • Discovers the clusters in a dataset. • Identifies outlier points.	• The algorithm has higher complexity compared to DBSCAN.

See Example




```python
#用分好的cluster lable 作為顏色,而機率值作為色彩飽和度
plot_kwds = {'alpha' : 0.5, 's' : 80, 'linewidths':1}
palette = sns.color_palette()
cluster_colors = [sns.desaturate(palette[col], sat)
                  if col >= 0 else (0.5, 0.5, 0.5) for col, sat in
                  zip(clusterer.labels_, clusterer.probabilities_)]
plt.scatter(test_data.T[0], test_data.T[1], c=cluster_colors, **plot_kwds)
```