DBSCAN — Density-based spatial clustering of applications with noise

HDBSCAN — Hierarchical DBSCAN

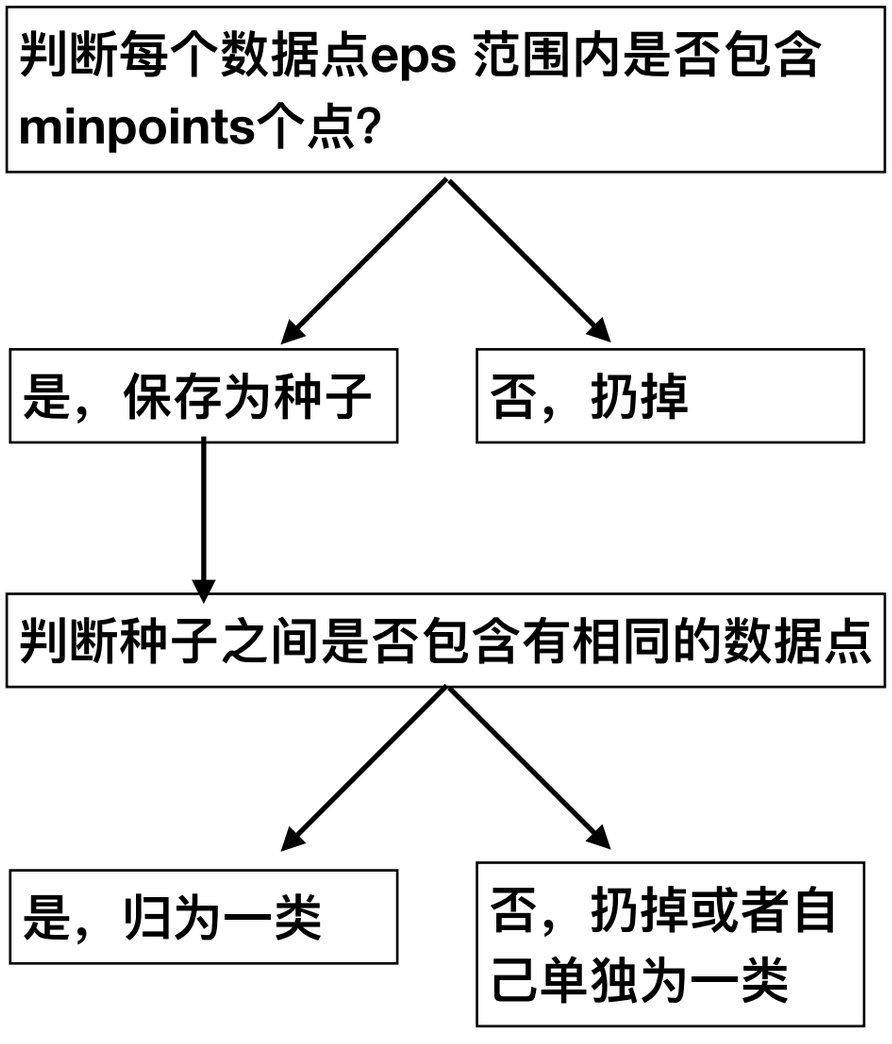
Lu Li
20210416

https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

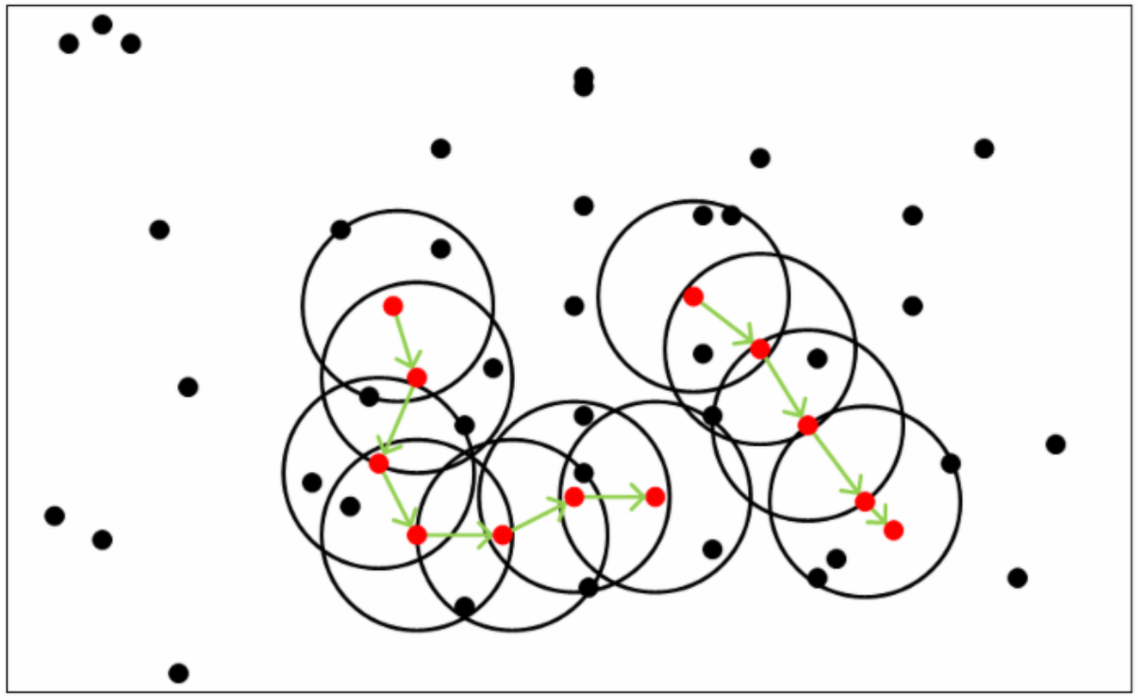
Clustering

DBSCAN

参数： eps (邻域的半径)
minPoints(邻域半径内包含的点数)

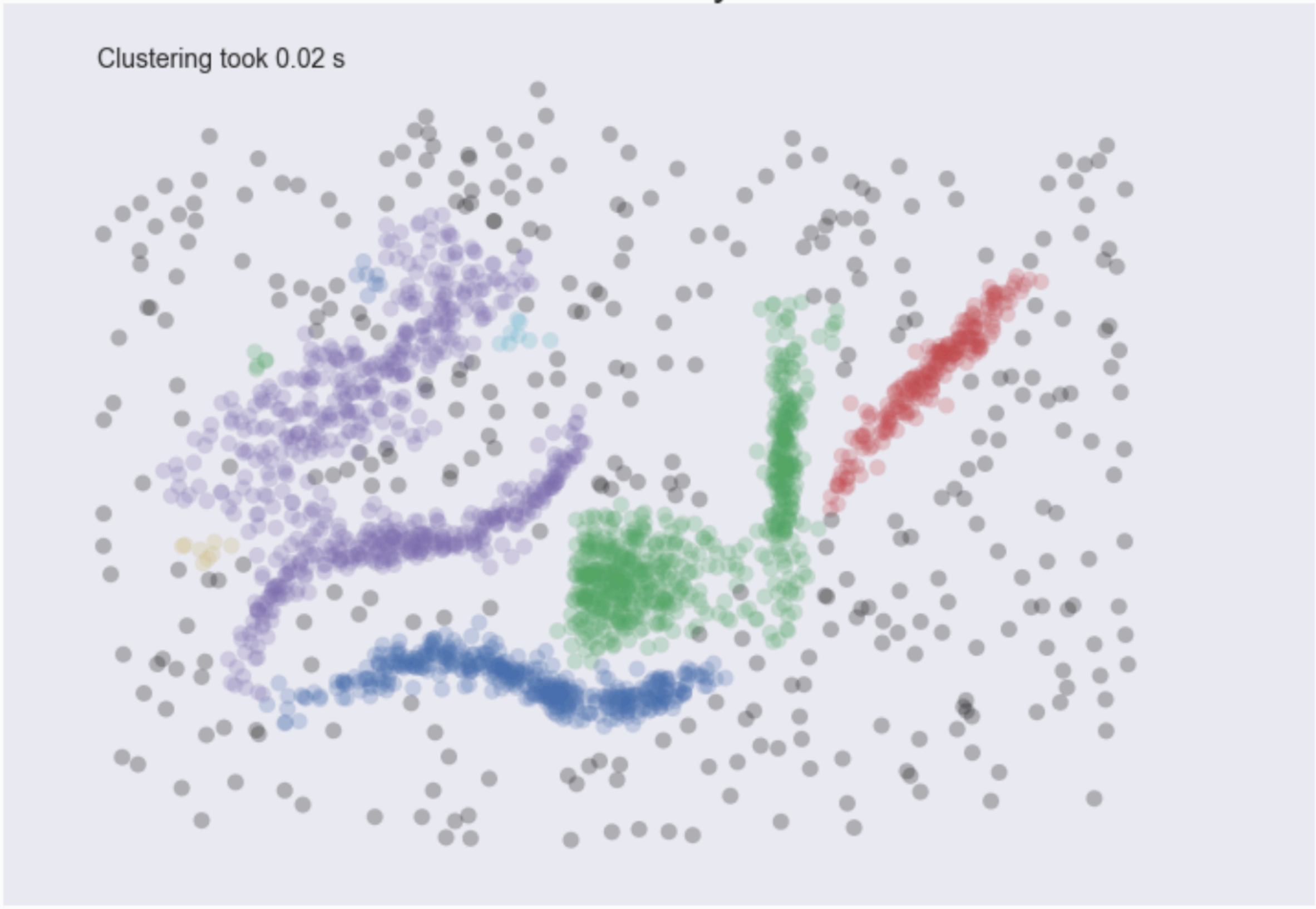


FOF： minPoints = 1



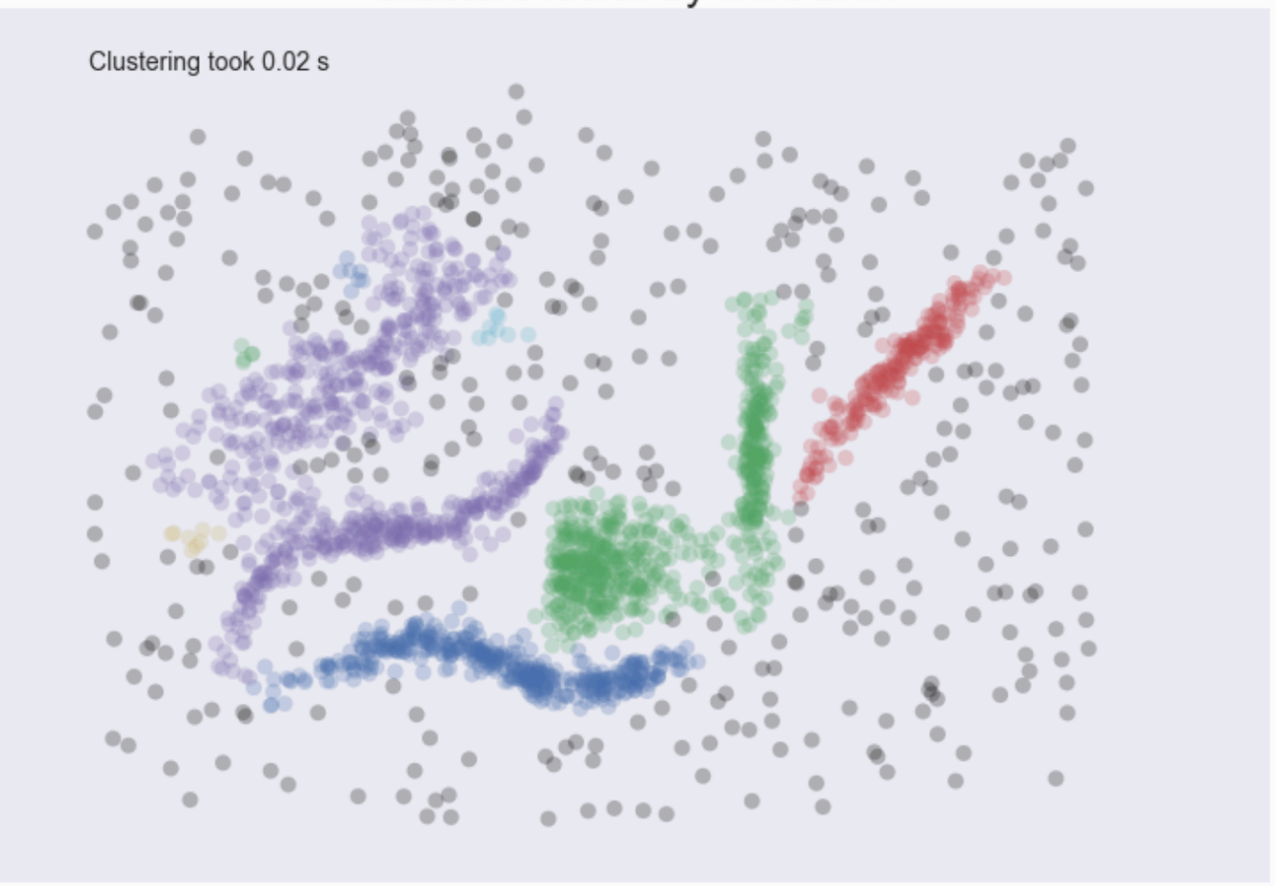
定义了临界密度一刀切

Clusters found by DBSCAN

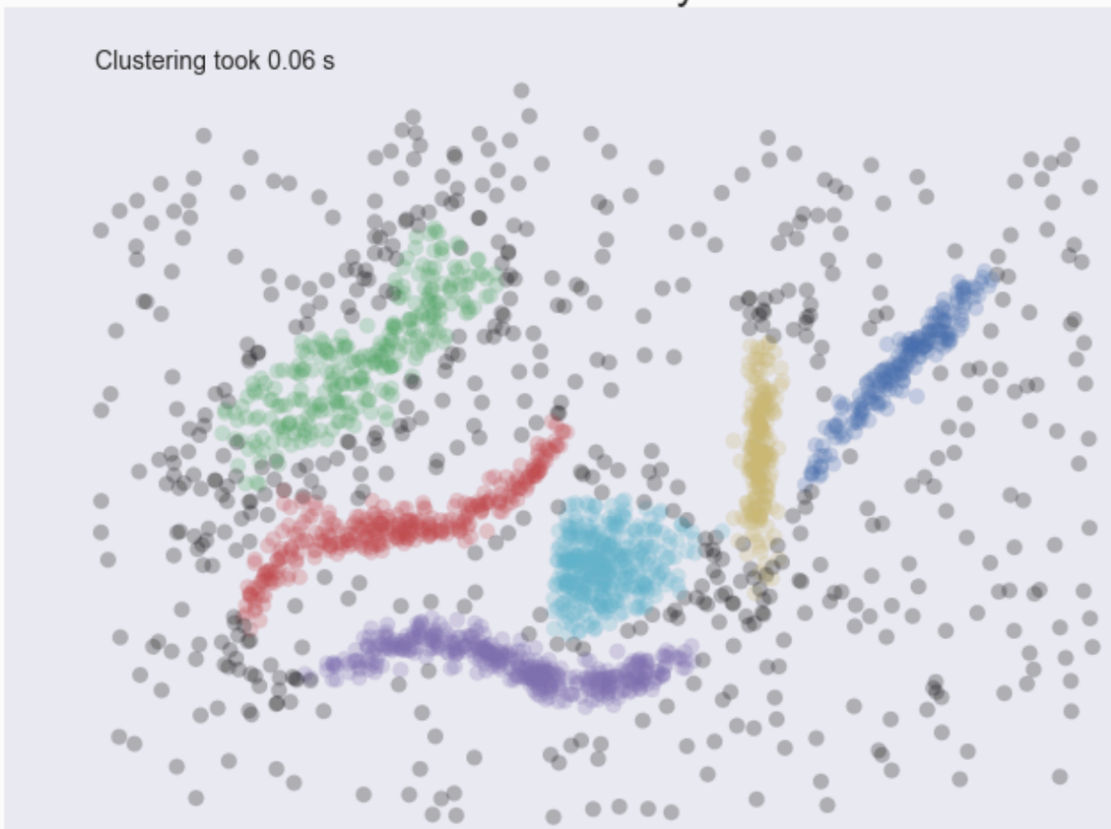


HDBSCAN

Clusters found by DBSCAN

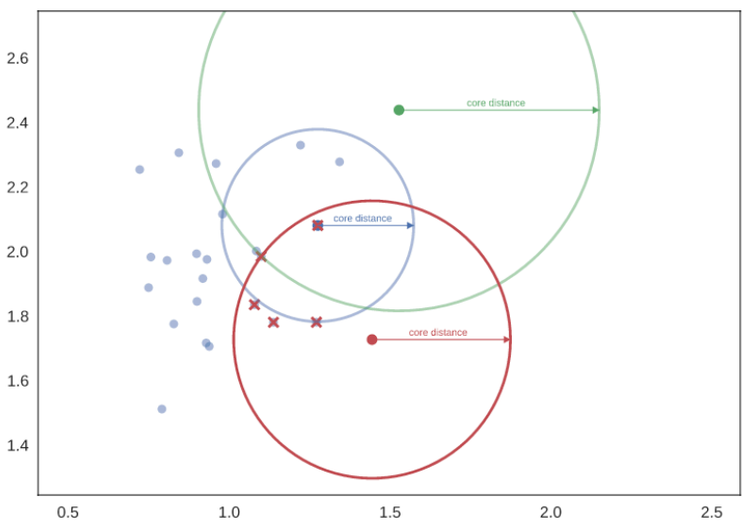


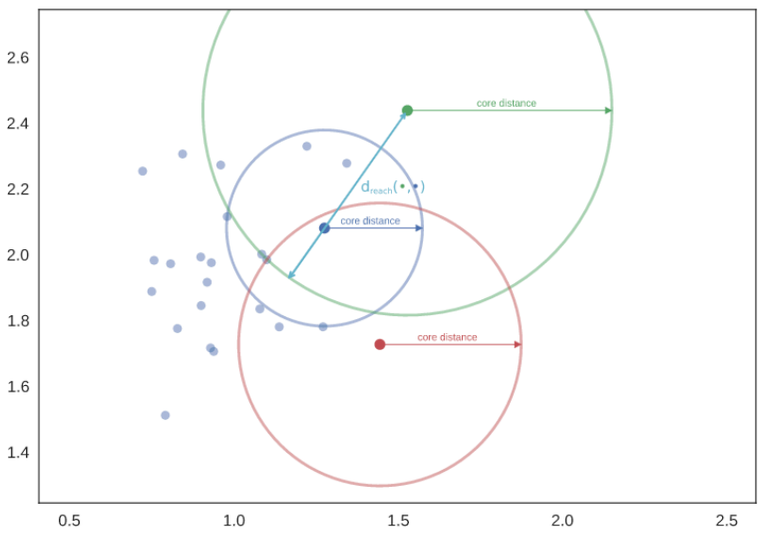
Clusters found by HDBSCAN

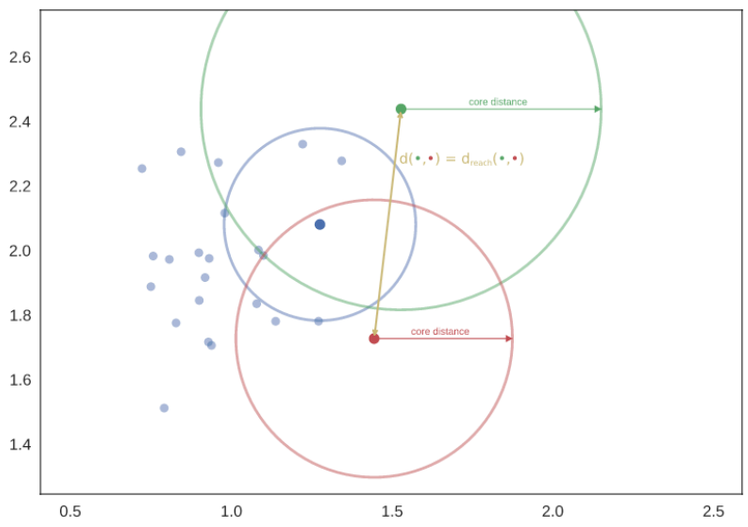


mutual reachability distance

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

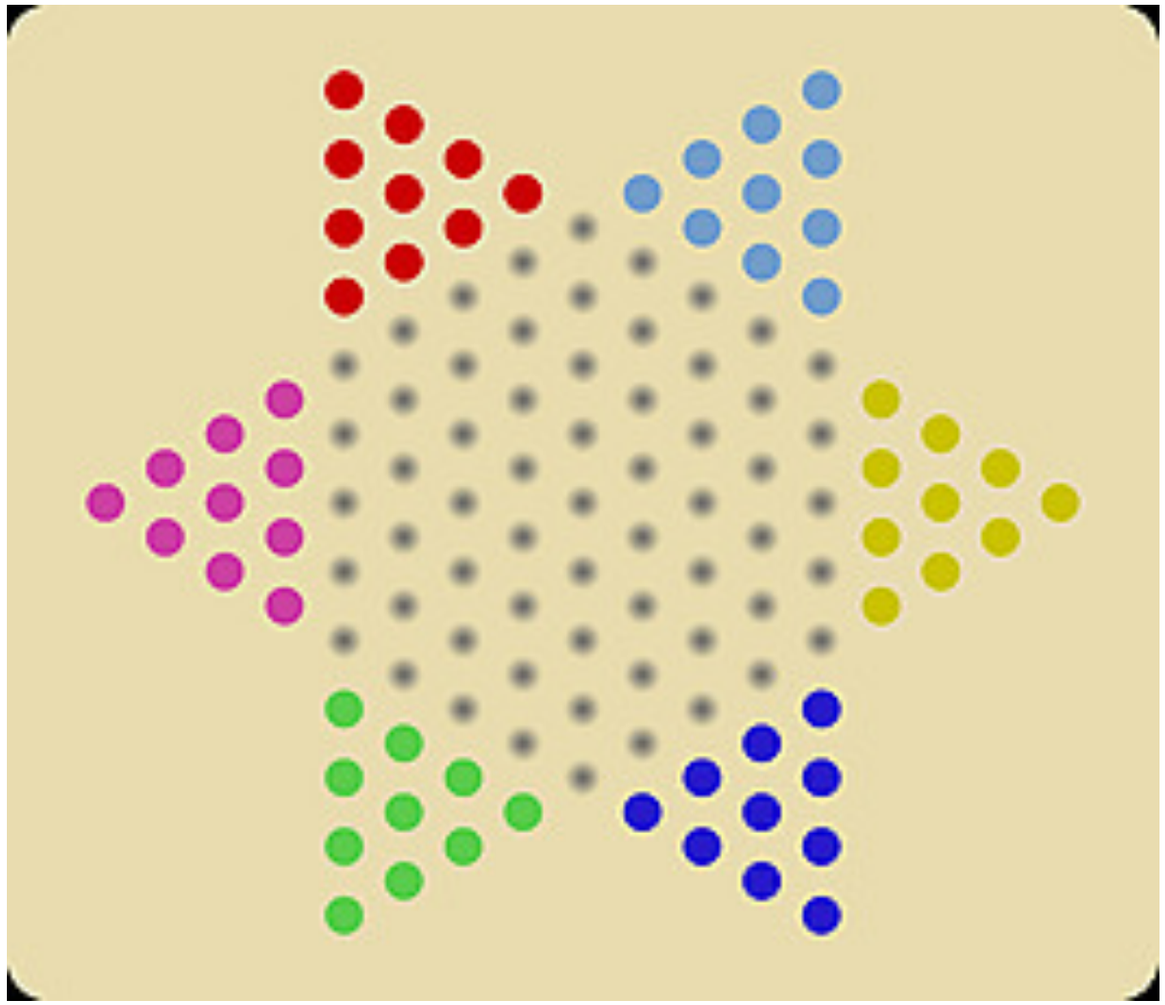






mutual reachability distance

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

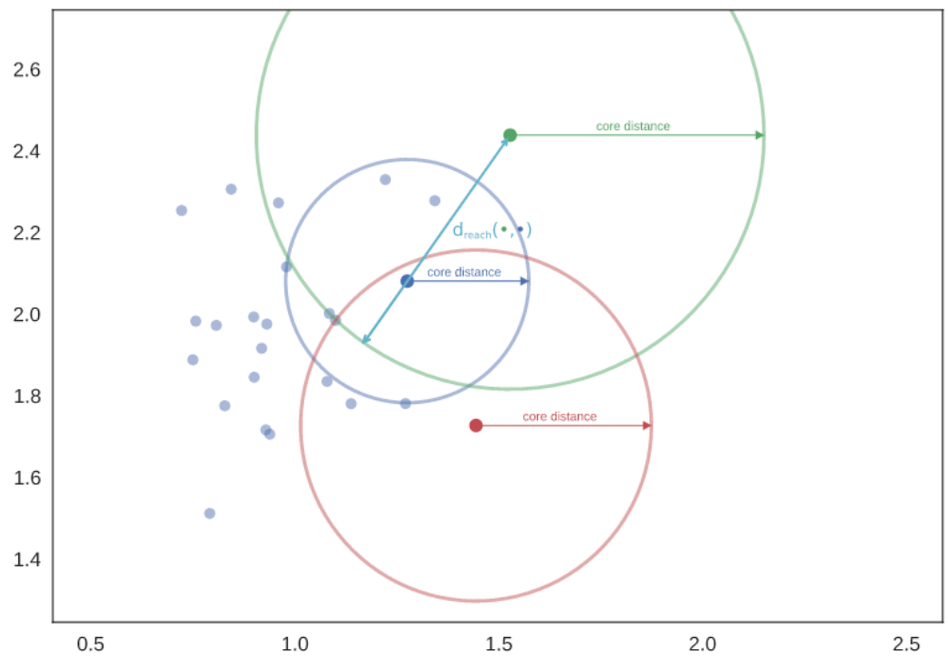


空间密度均匀：

点与点之间互相可达距离一致

mutual reachability distance

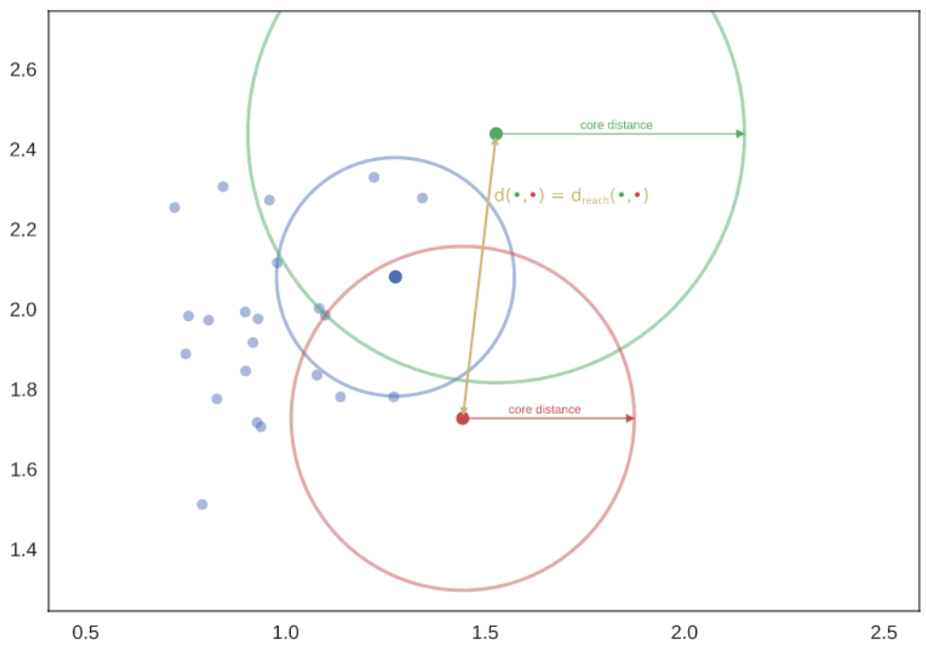
$$d_{mreach-k}(a,b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a,b)\}$$



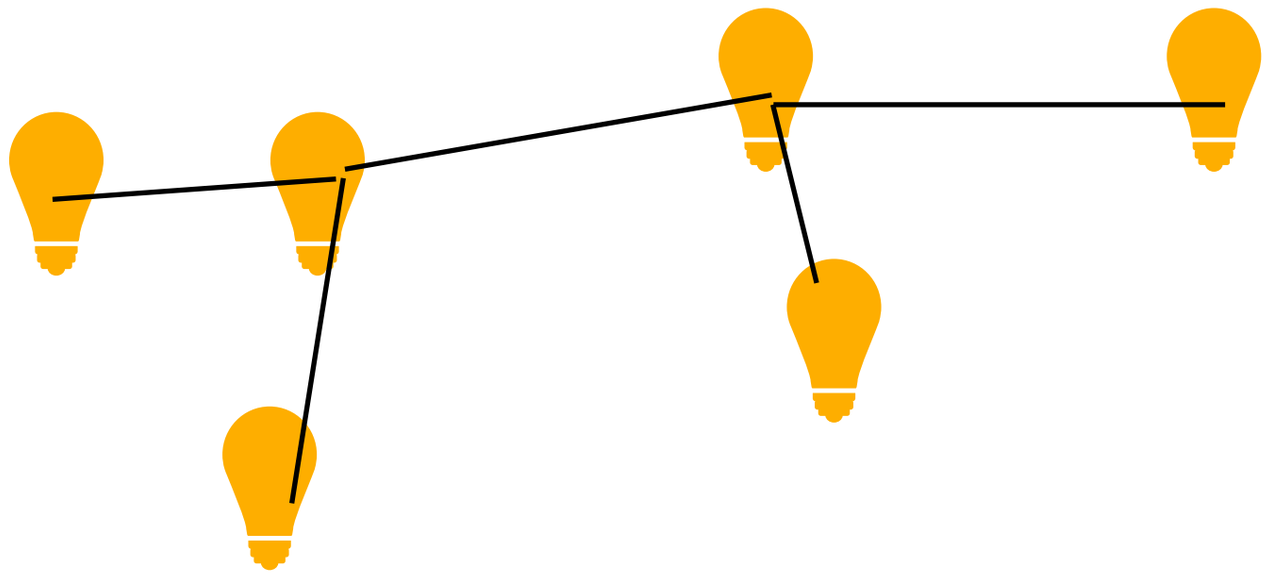
同一个簇内的点，相互距离没有影响；

噪声与噪声点之间的距离也没有影响。

但是簇和簇之间的噪声点被“推开”。



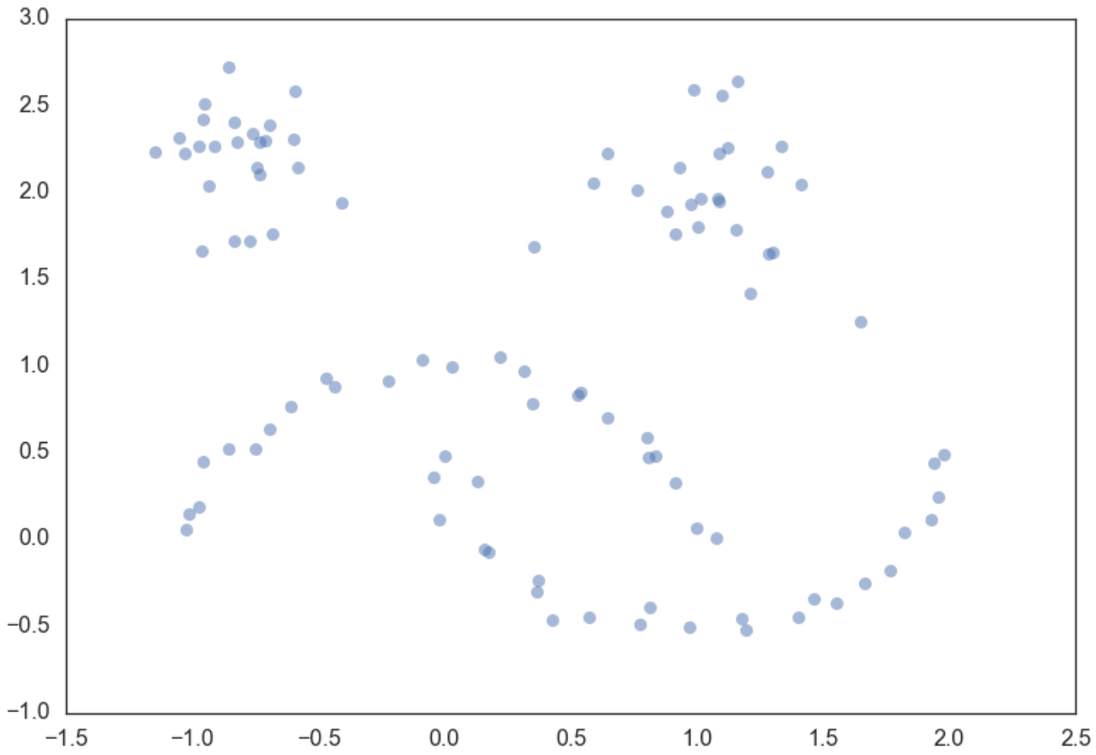
Minimum Spanning Tree

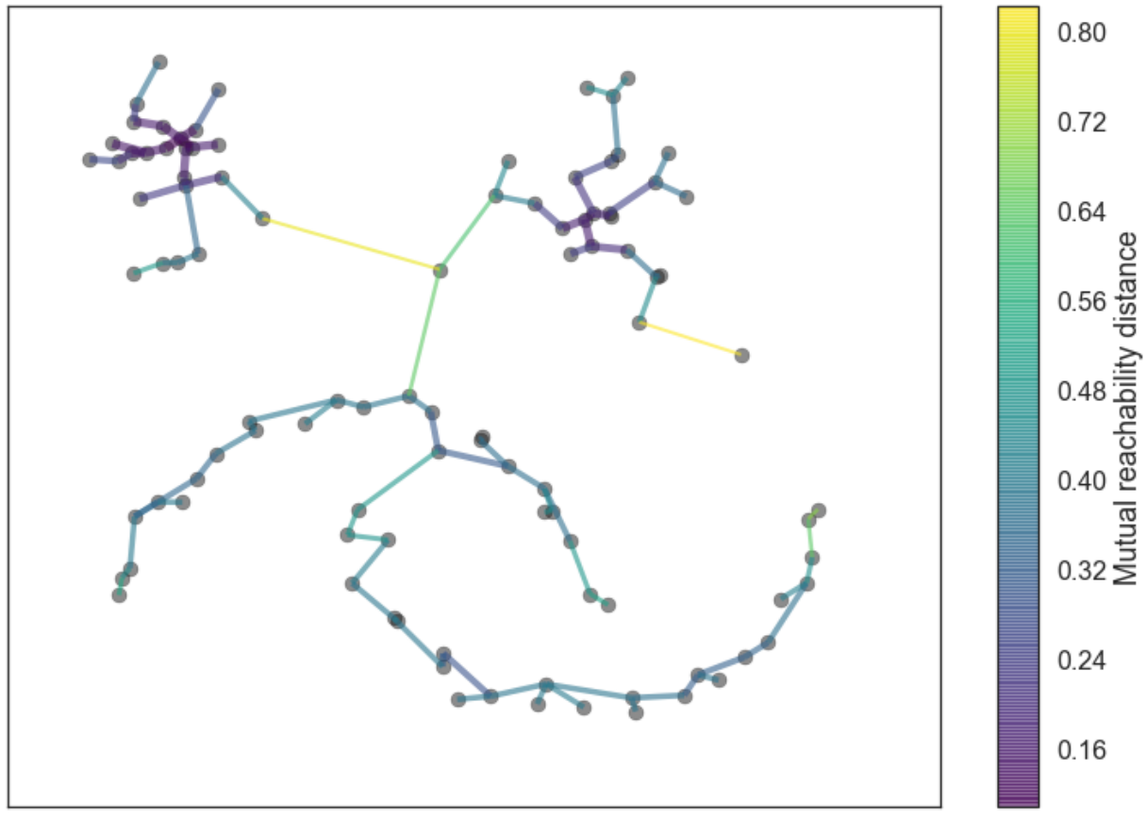


最小生成树

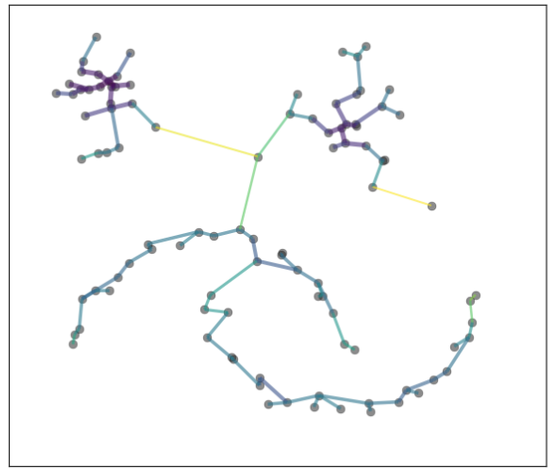
要在n个城市之间铺设光缆，主要目标是要使这 n 个城市的任意两个之间都可以通信，使铺设光缆的总费用最低。

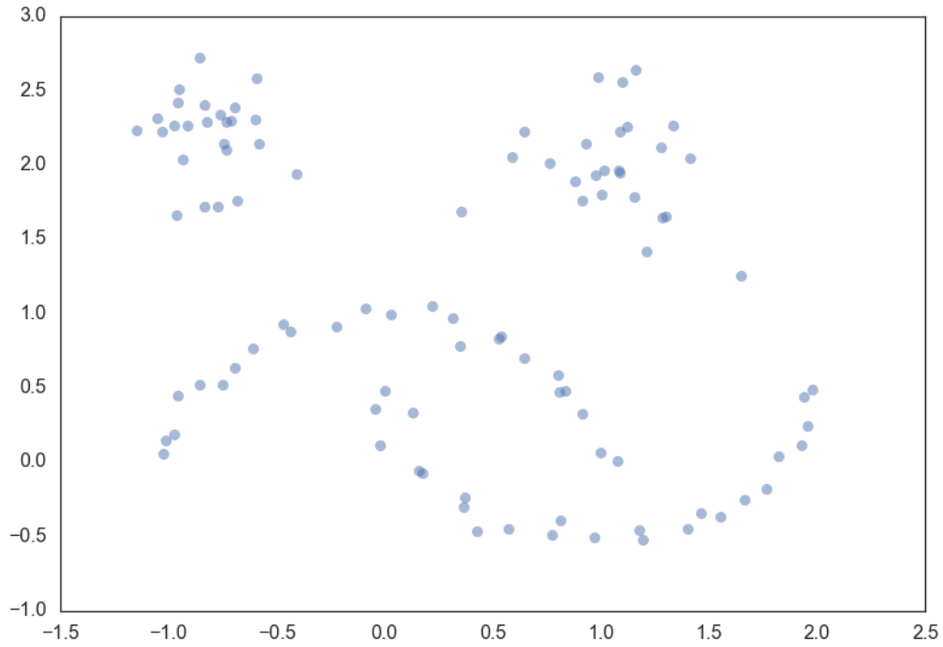
Minimum Spanning Tree



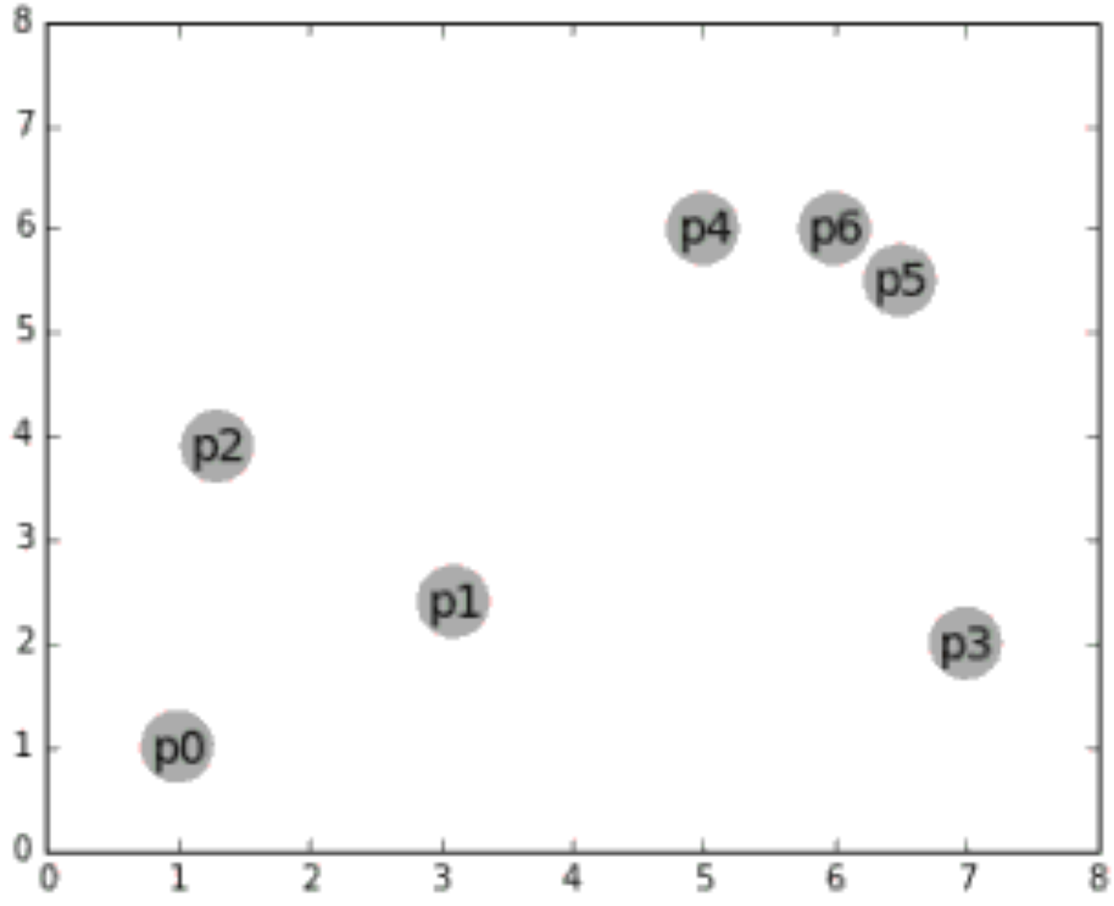


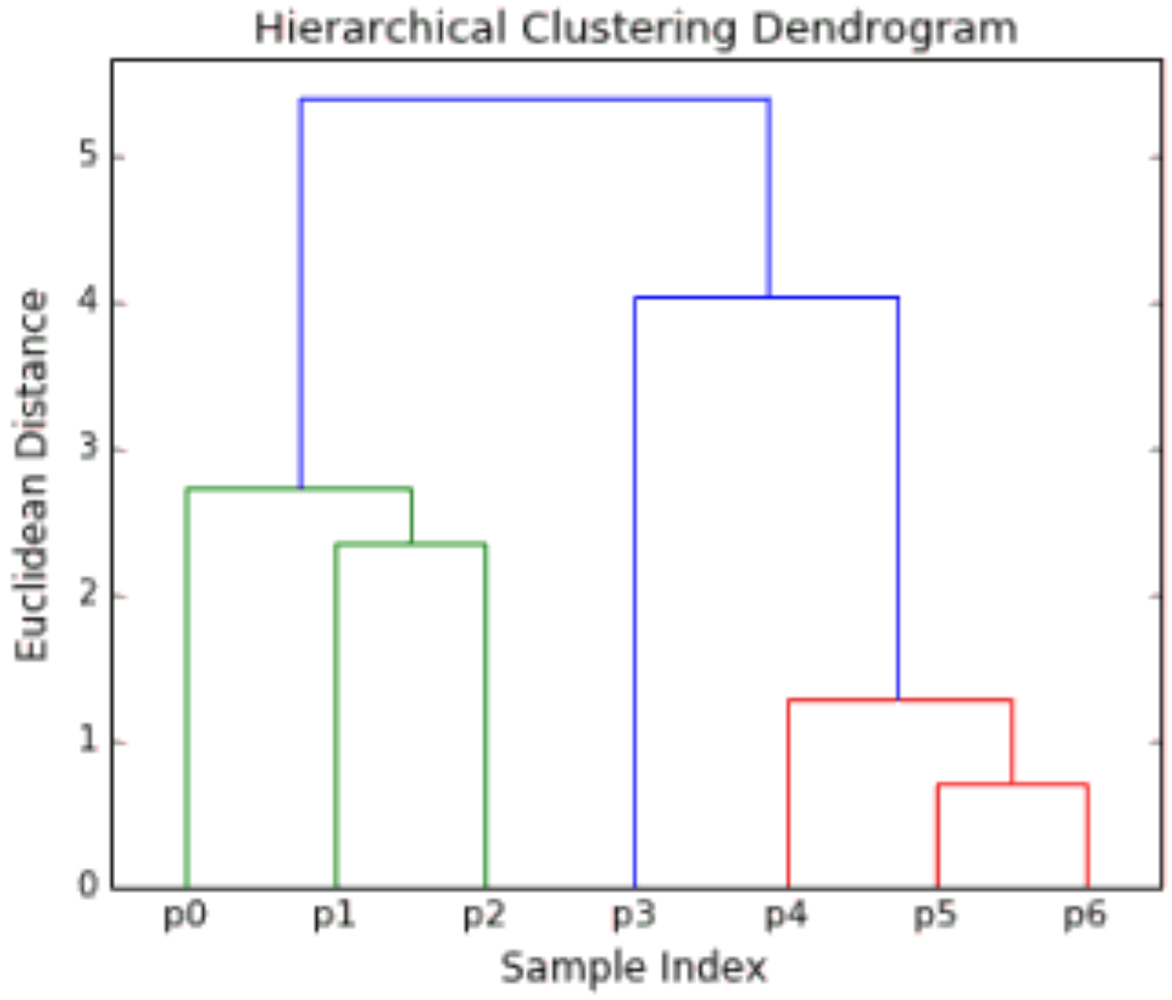
Hierarchical Clustering



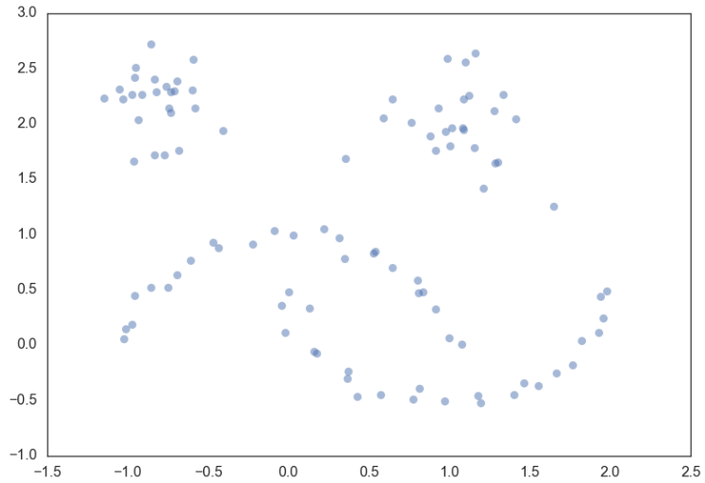


Hierarchical Clustering

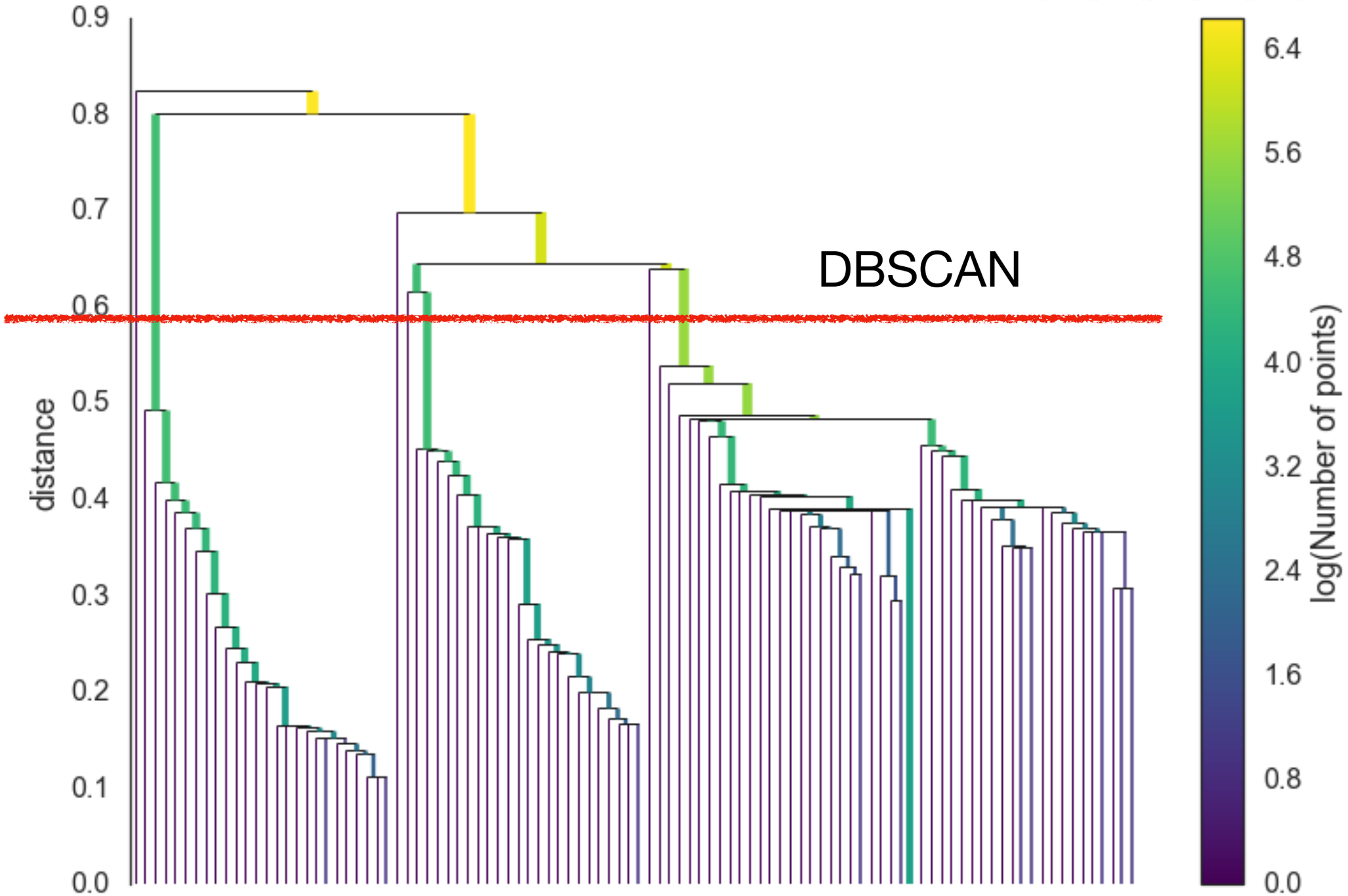




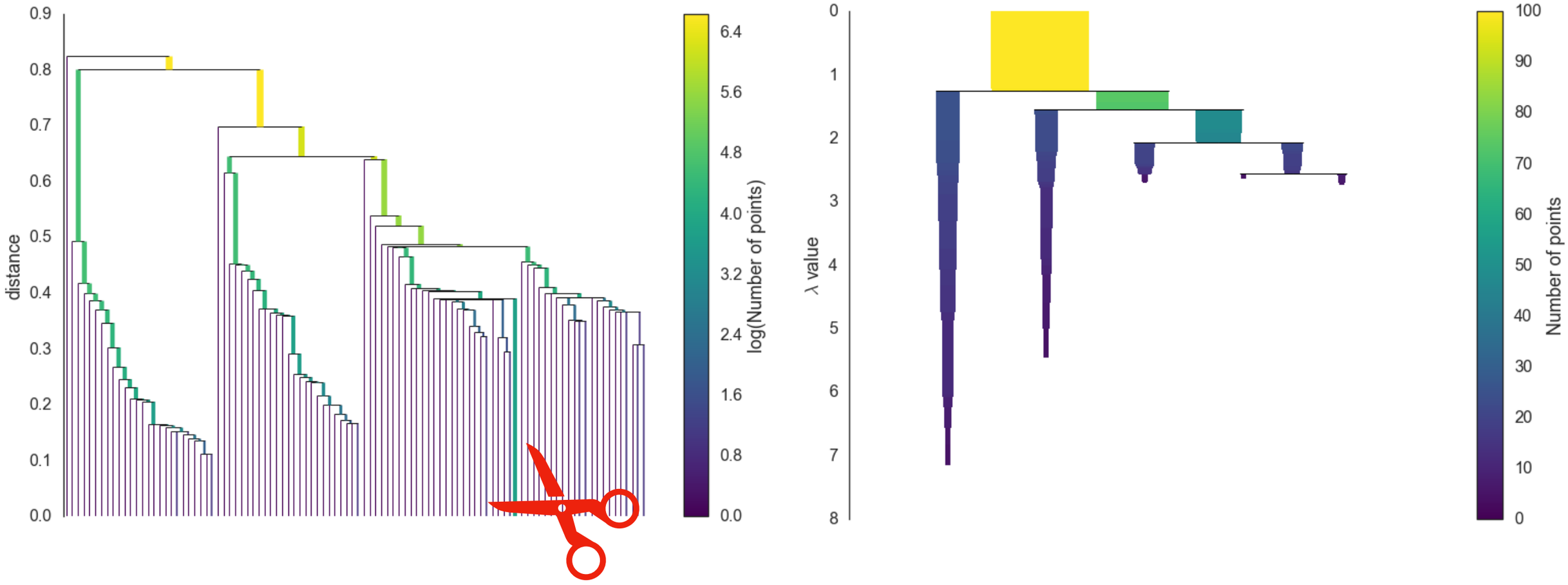
Hierarchical Clustering



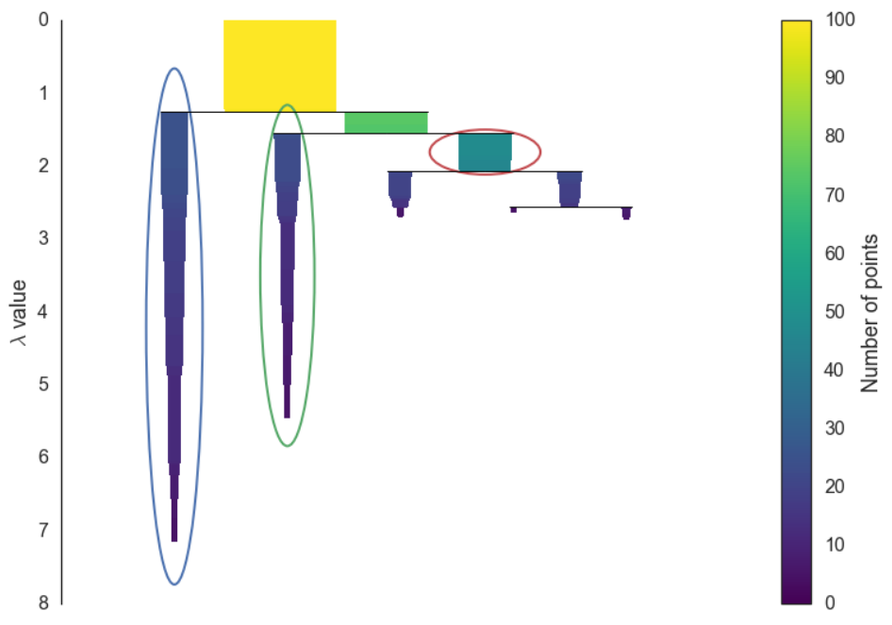
DBSCAN

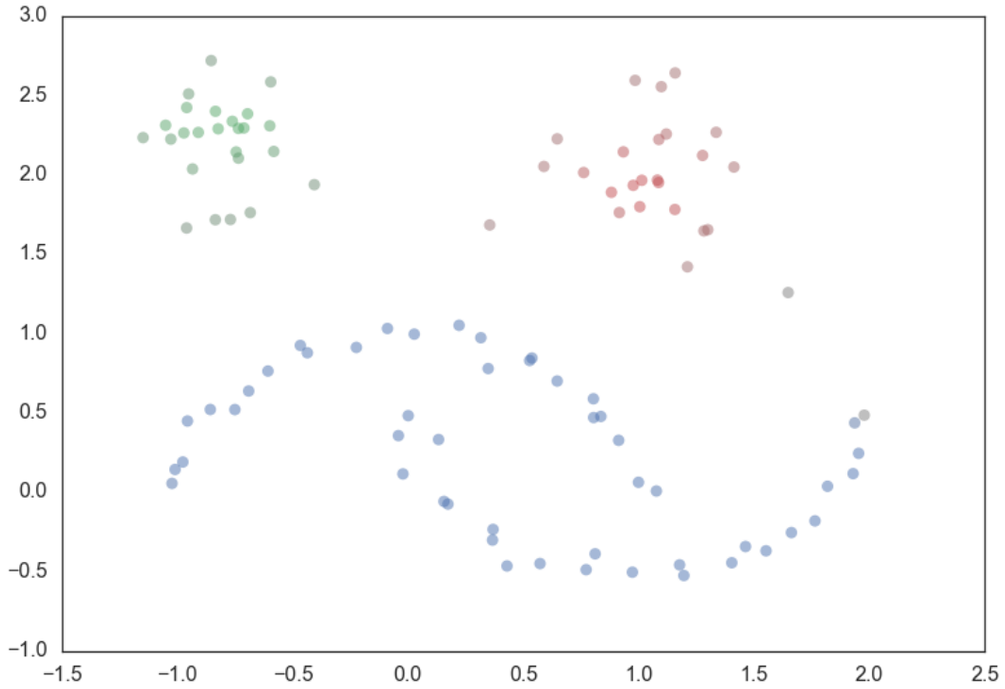


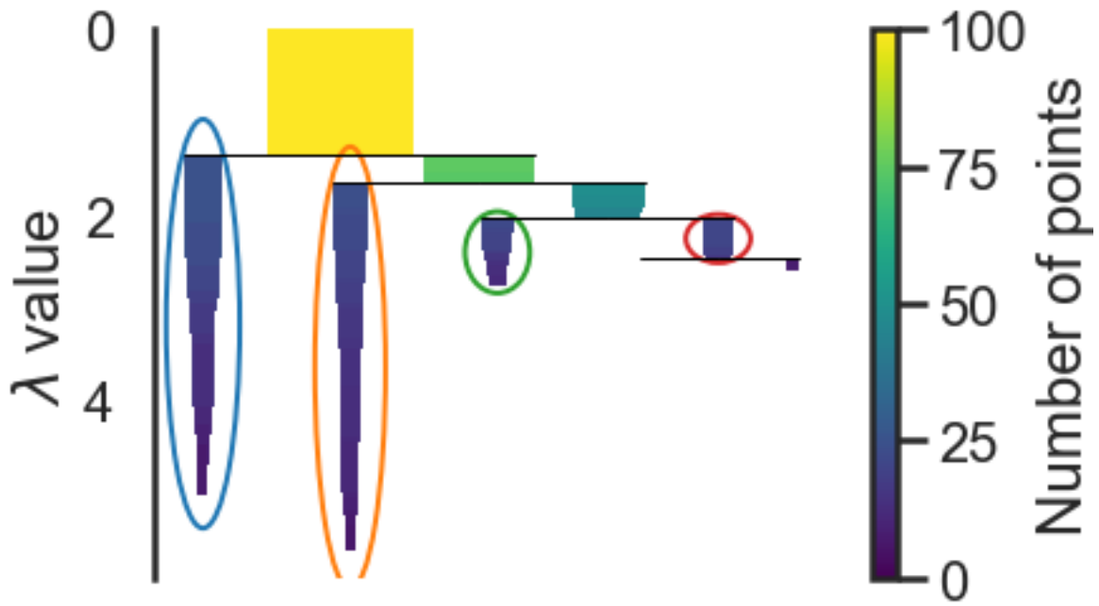
Condense the cluster tree

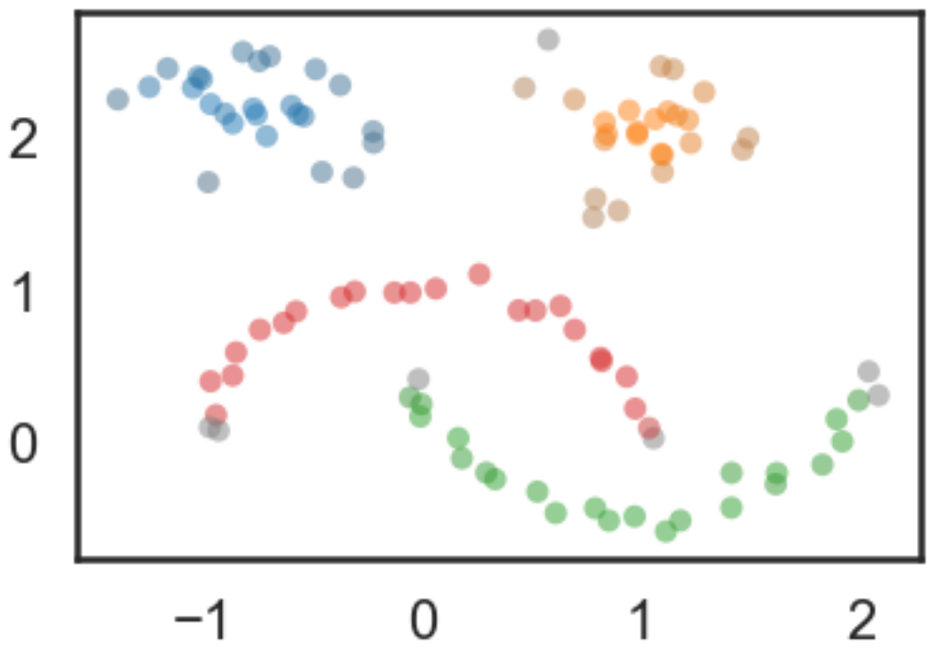


修剪枝丫：从上至下，历遍节点，剪掉小于 min cluster members 的节点

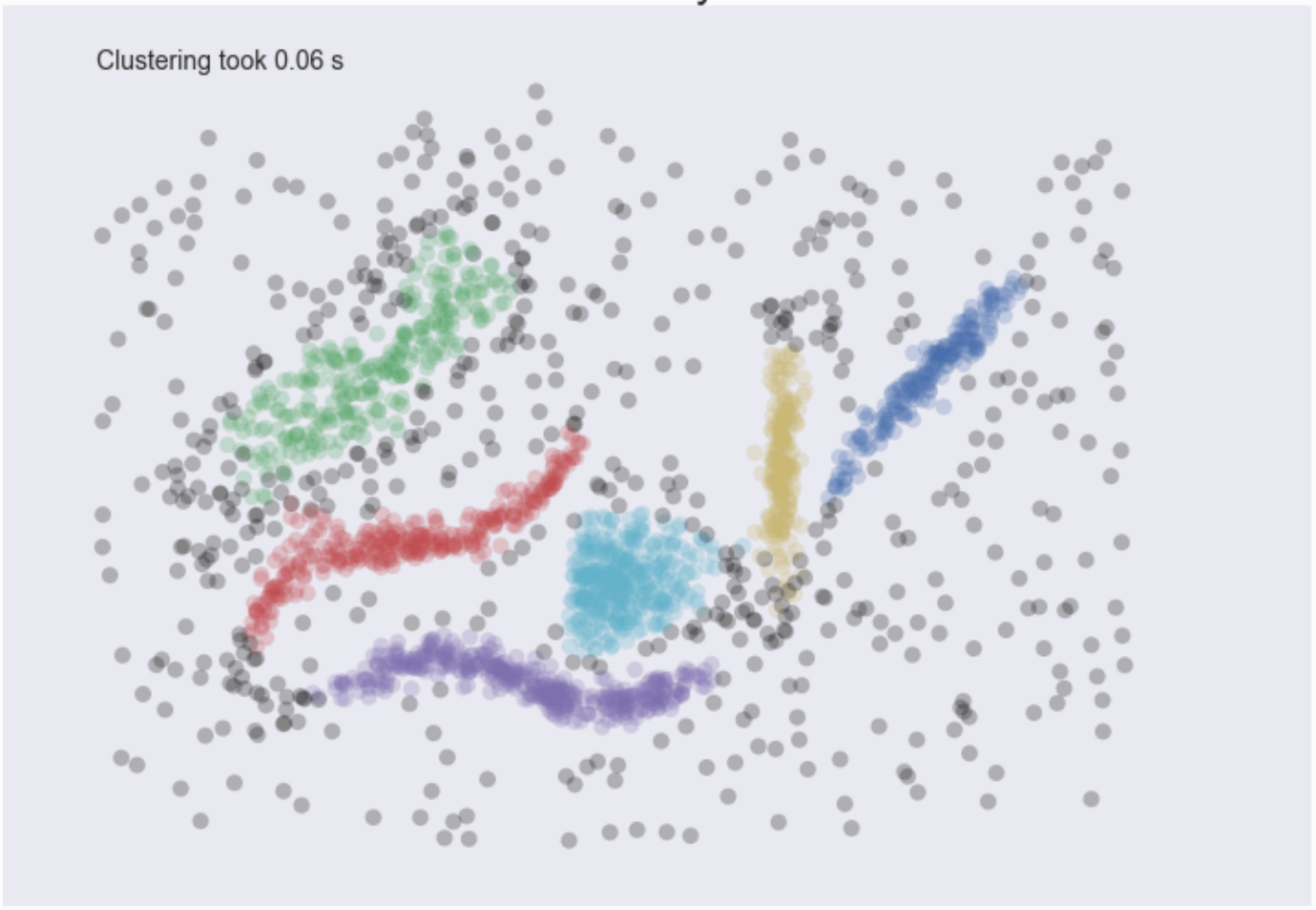

λ value
0
1
2
3
4
5
6
7
8
Number of points
100
90
80
70
60
50
40
30
20
10
0


3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5


λ value
0
2
4
Number of points
100
75
50
25
0


2
1
0
−1
0
1
2

Clusters found by HDBSCAN



使用**相互可达距离**替换欧氏距离，该距离可以使得密度低的点离密度高的区域更远，减少dbscan对Eps阈值的依赖性
使用最小生成树构建层次聚类模型，引入**层次聚类**思想
对最小生成树的最小子树做了限制，减少计算量，同时保证生成的类簇不要过小
使用“**簇稳定性**”的度量方式自动划分类簇，不需要自行设定阈值