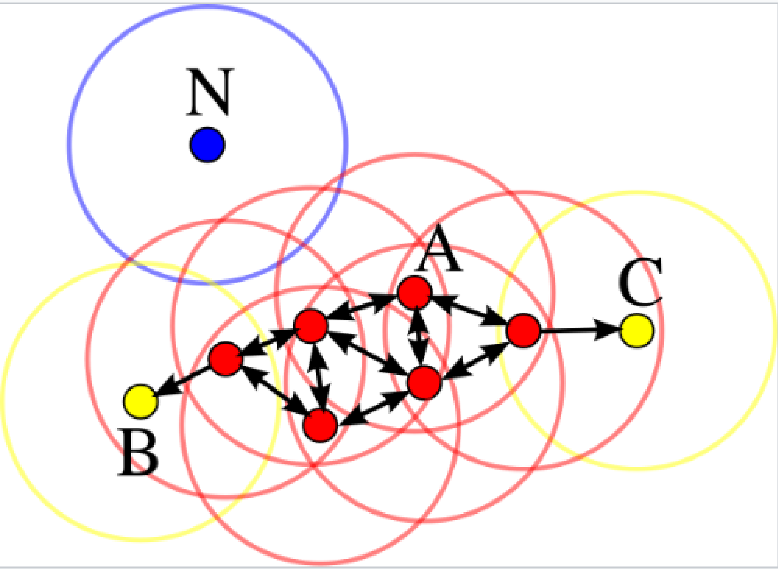
# DB SCAN



In this diagram, $\mathrm{minPts} = 4$. Point A and the other red points are core points, because the area surrounding these points in an $\varepsilon$ radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable.

● core point

Border point

● Noise / outlier

It is helpful for non lineer clustering.

## Hyperparameter

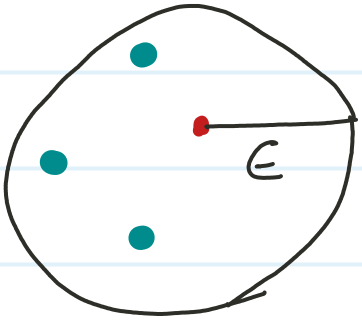
① minpoints = 4

② $\epsilon$ = Radius

($\epsilon$ = epslon)

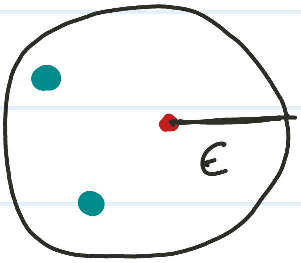
☆ Core point –

No. of points within the $\epsilon$ should be $\geq$ min point



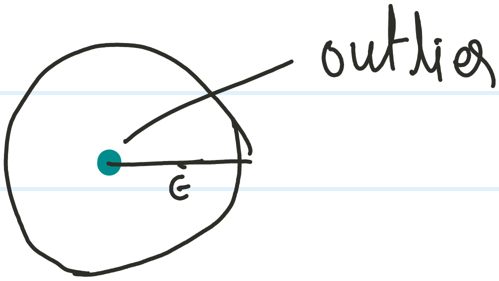
☆ Border point –

No. of data point within the radius will be less the min point 4.

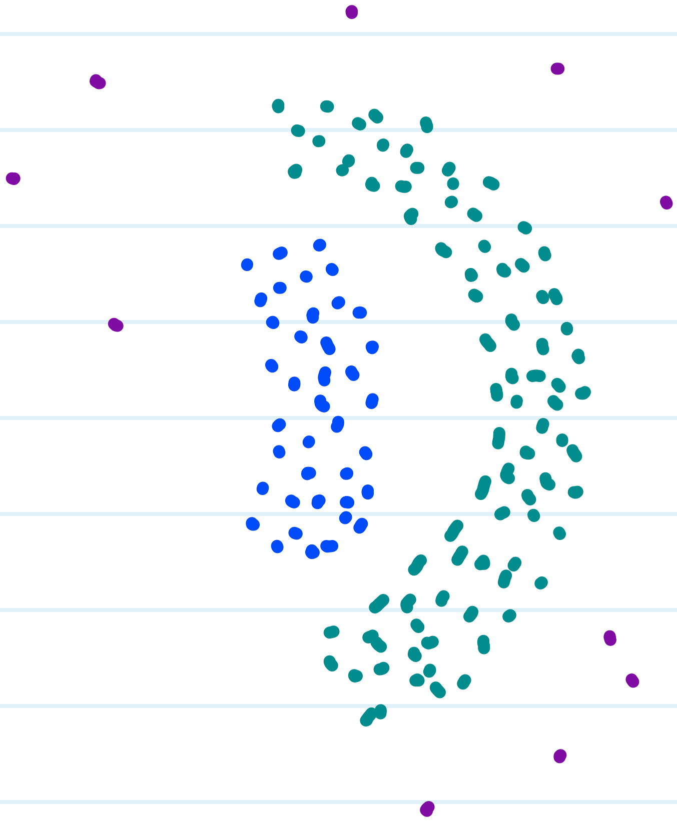


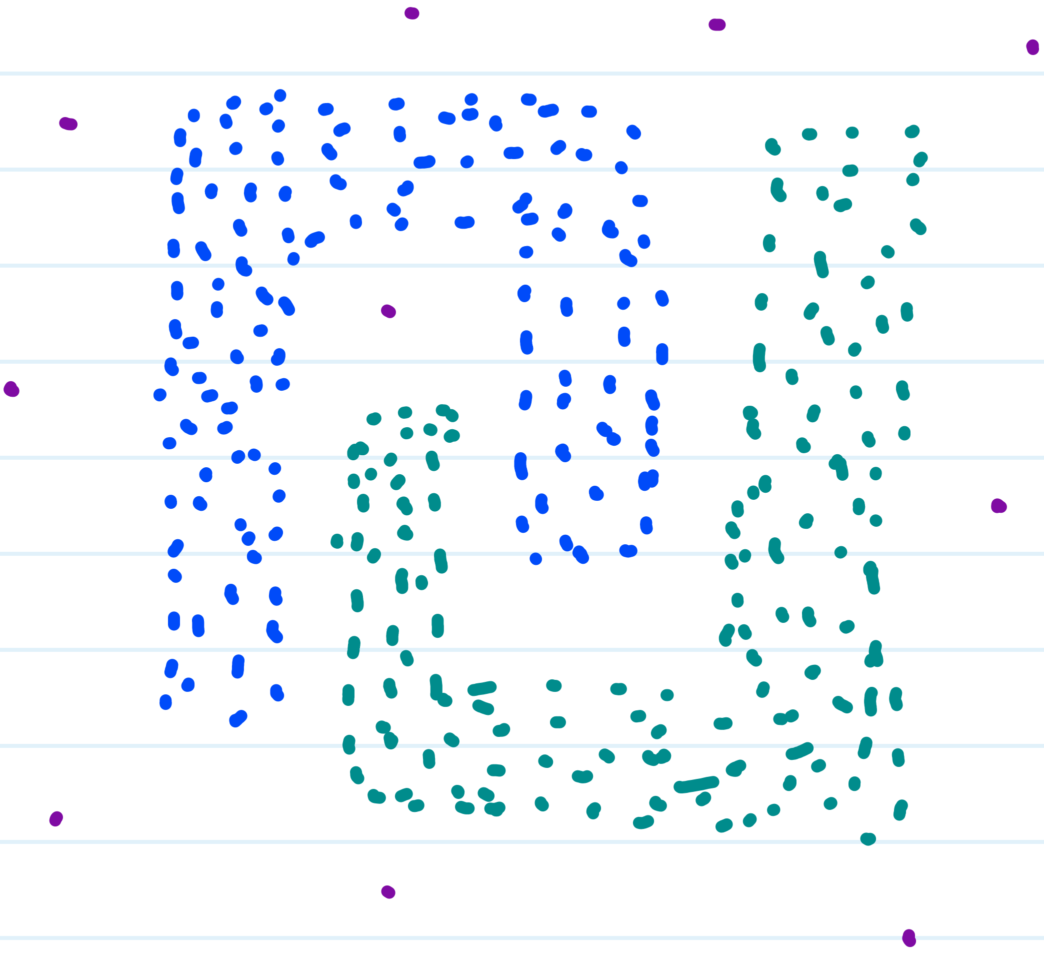
# Noise/ outlier - (DBSCAN robust to outlier)

No. of



Some example after applying DBSCAN

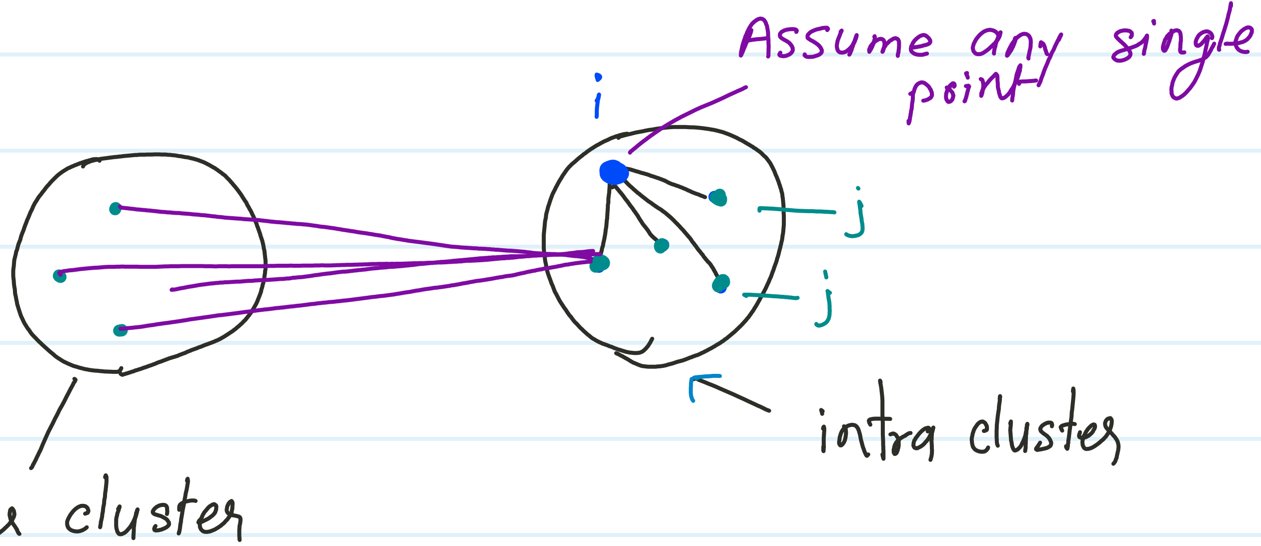




For any clustering method, validation method we will use, that is silhouette score.

# ☆ Important hyperparameter in the clustering

## ☆ Silhouette score



① $a(i) = \frac{\text{Sum of distance of all the point in intra cluster}}{\text{No. of data point}}$

Formula -

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

② $b_i = \dfrac{\text{Distance of all the point of any other clustes}}{\text{No of point}}$

Formula :-

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j)$$

$b_i = 20$

$a_i = 15$

③ silhouette formula -

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_I| > 1$$

Constraint

$$S(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

Range of Silhouette score is 1 to -1

1 -> Good clustering

-1 -> Bad clustering