DBSCAN

Density Based Spatial Clustering of Applications with Noise

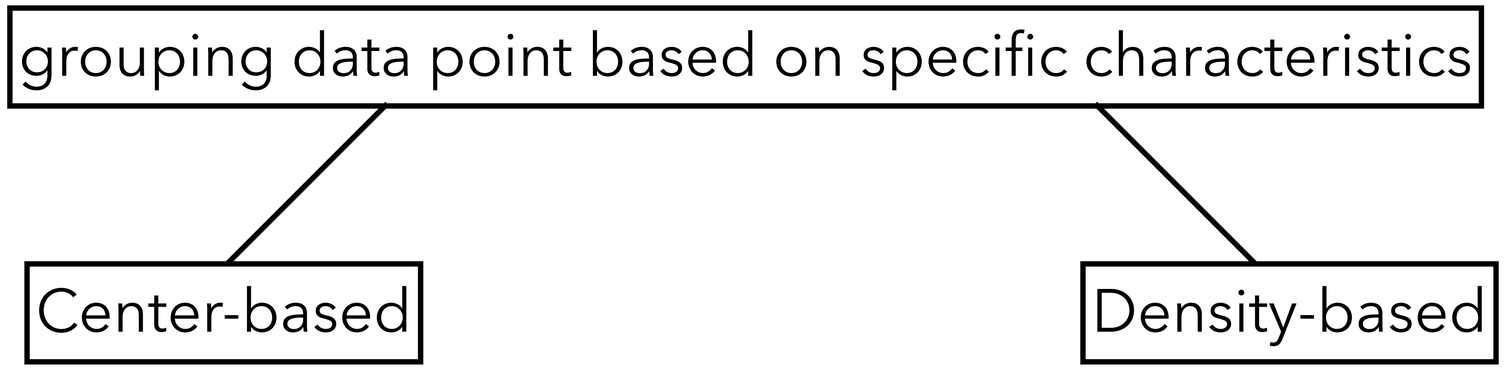
AI 05 Hyungjin Kim

What is DBSCAN

Clustering

No labels

Clustering = unsupervised ML technique



K-means Clustering

DBSCAN

What is DBSCAN

Density Based Clustering Algorithm

Key Assumption:

Clusters are dense regions in space separated by regions of lower density

Use **density** to gather points in space to form clusters

useful in the data that have a high density of observations

How it works

What we need

ϵ

eps (epsilon)

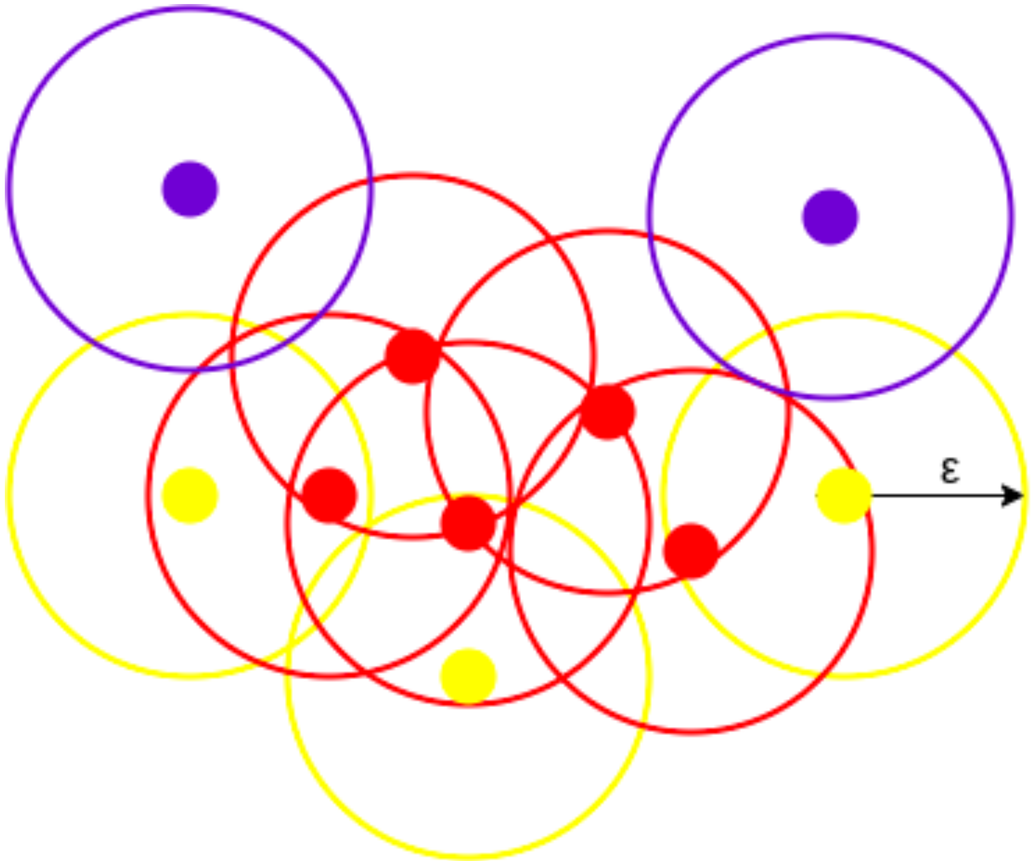
n_c

min_samples

(min_pts)

How it works

Hyper-parameter



ϵ

eps (epsilon)

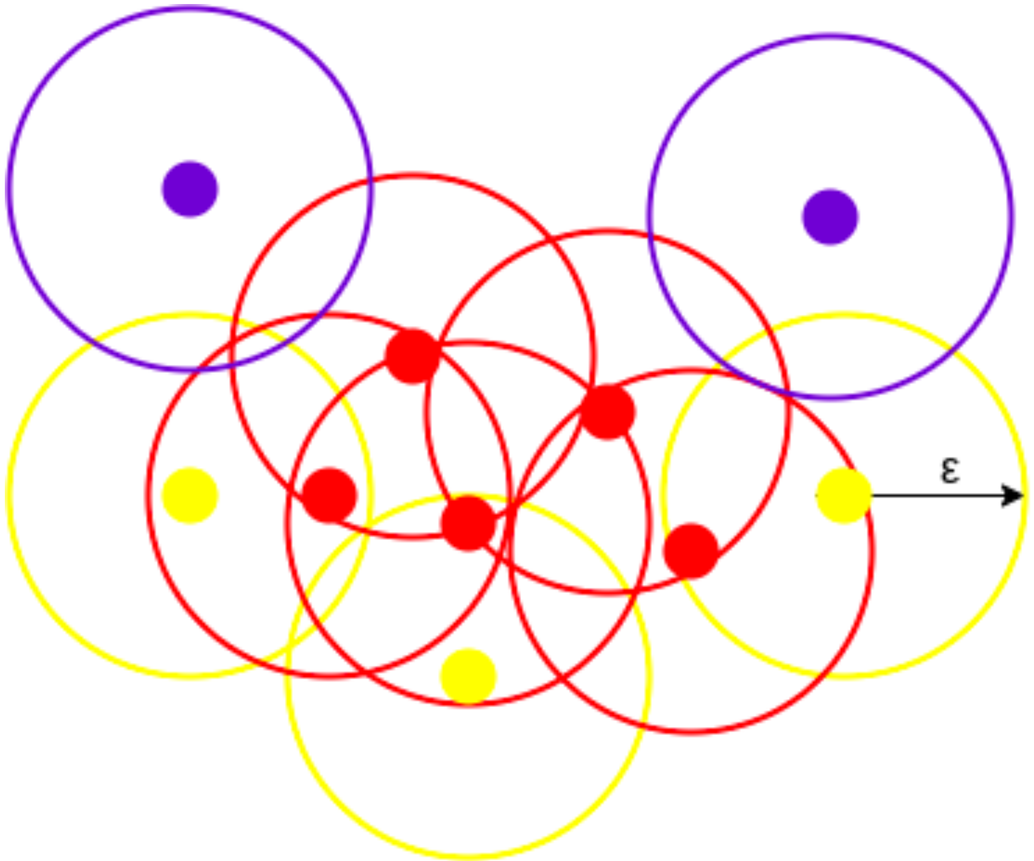
the radius of the circle which around each data point

check density

calculate distance
based on **Euclidean distance**

How it works

Hyper-parameter



n_c

min_samples (min_pts)

minimum number of data points required inside the circle

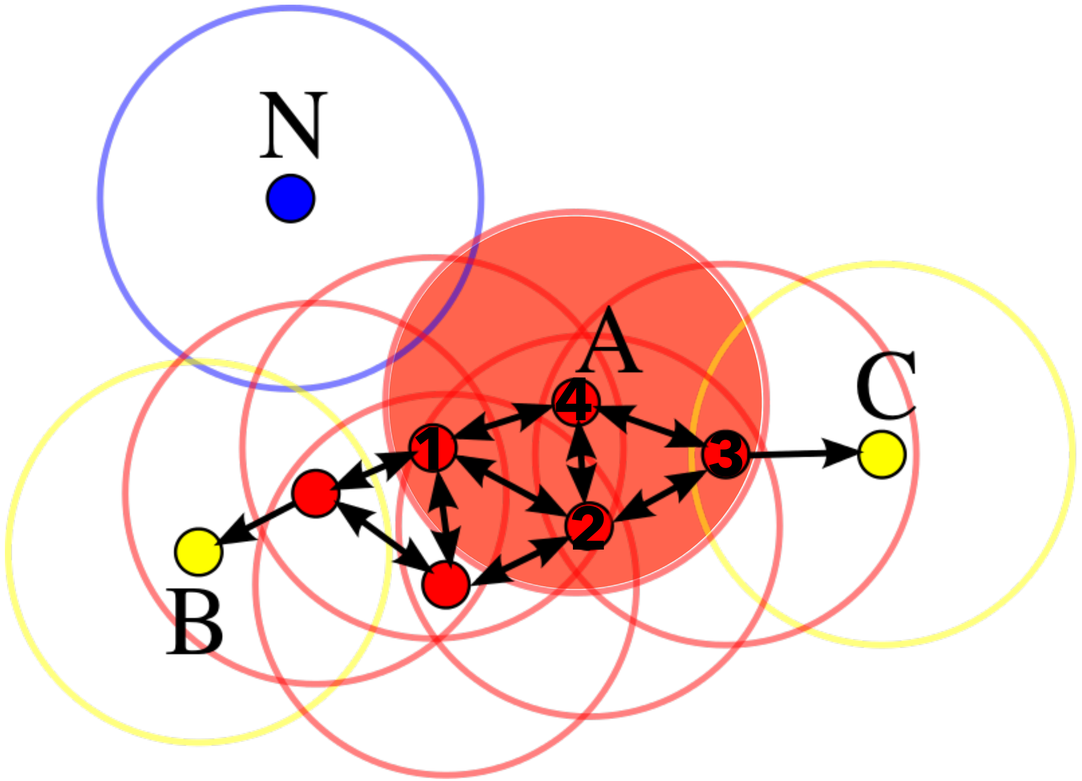
Satisfied = **Core** point

Not enough = **Border** point

Not satisfied at all = **Noise**

How it works

Core - Border - Noise



Case:

min_samples = 4

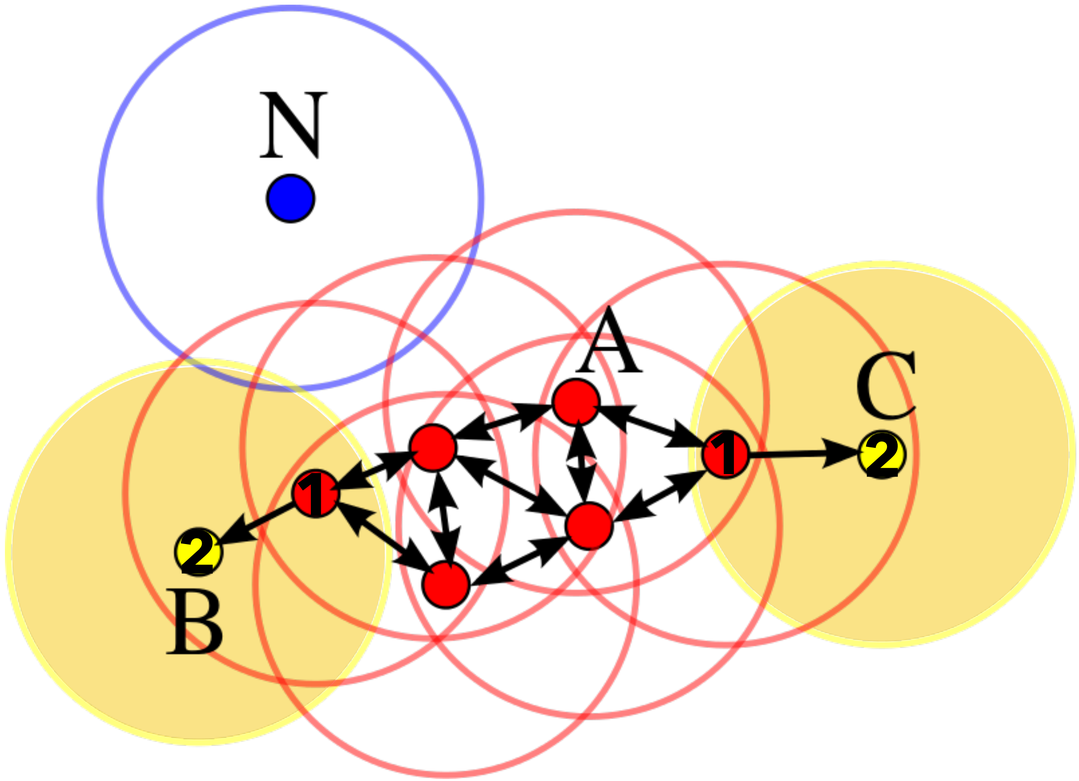
A = Core point

B, C = Border (Reachable) point

N = Noise point

How it works

Core - Border - Noise



Case:

min_samples = 4

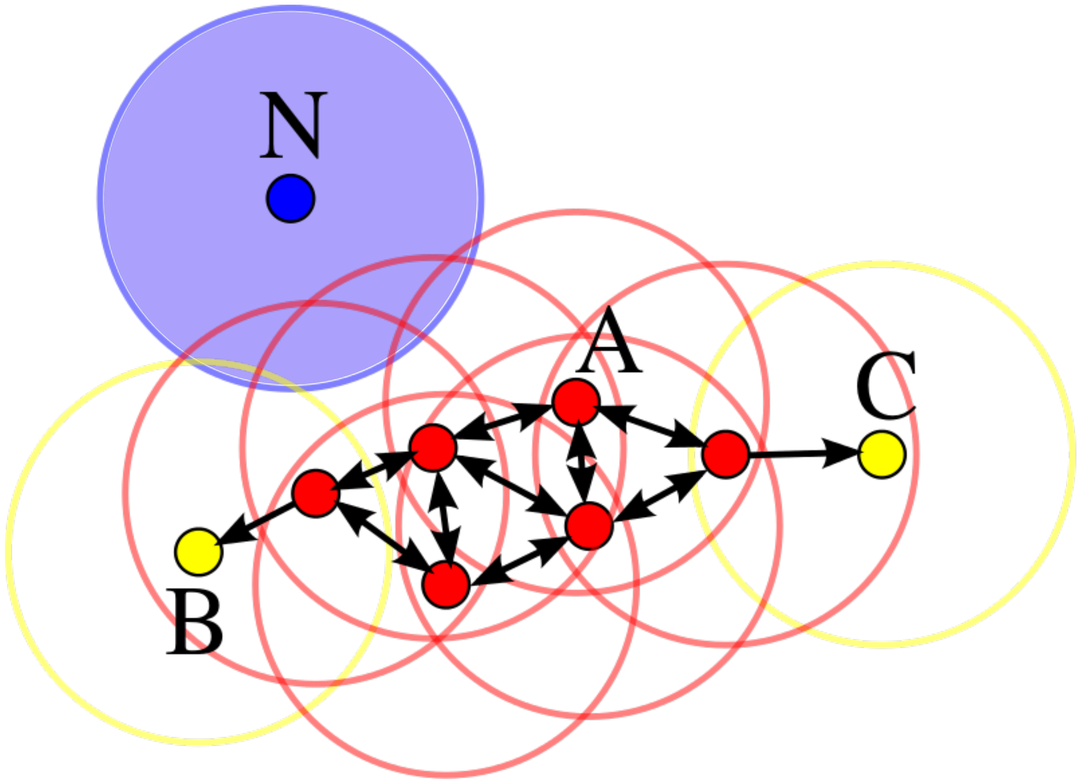
A = Core point

B, C = Border (Reachable) point

N = Noise point

How it works

Core - Border - Noise



Case:

min_samples = 4

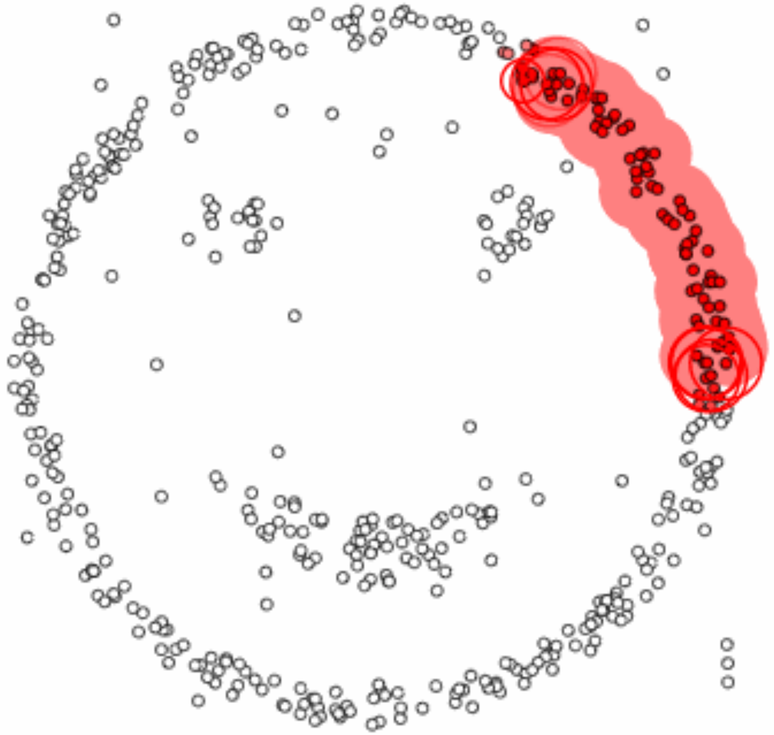
A = Core point

B, C = Border (Reachable) point

N = Noise point

How it works

Step by Step Procedure



Picking up a point in dense region

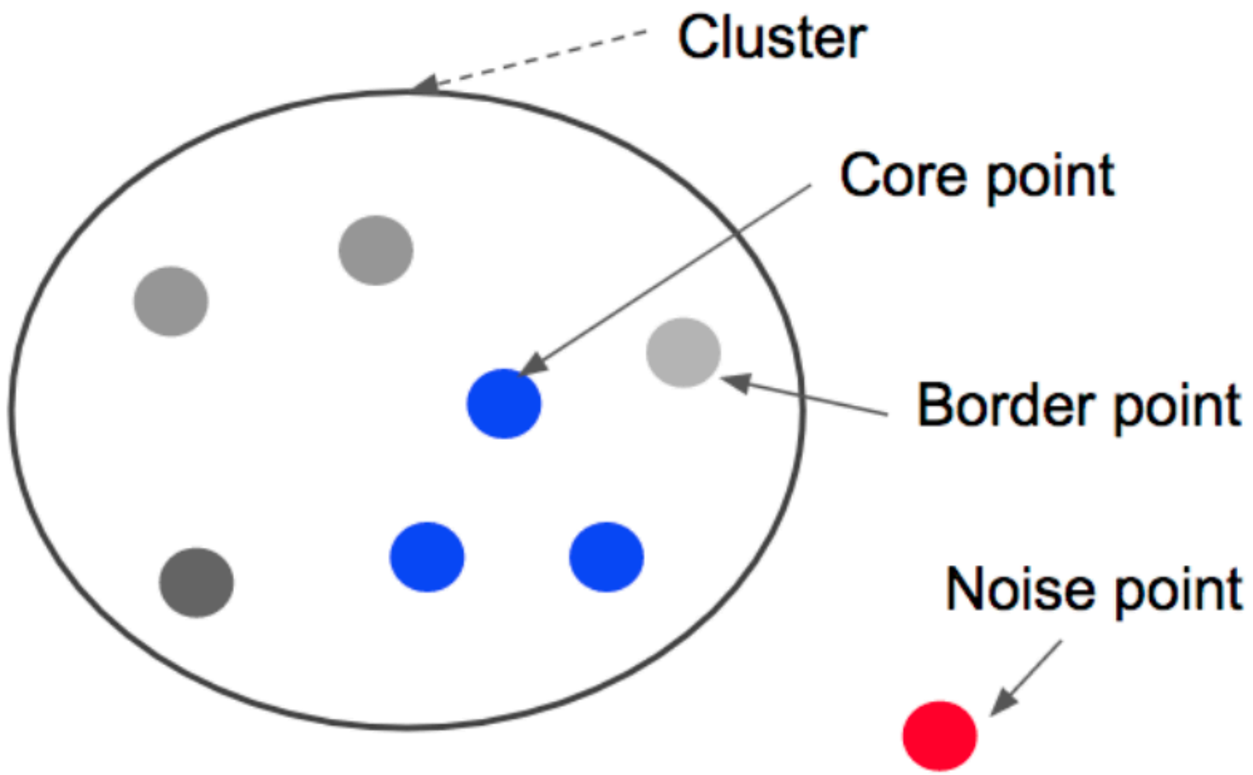
Find out how many points are inside the point's circle

Satisfied the required num -> cluster

Repeating until visiting every point

How it works

Result of DBSCAN



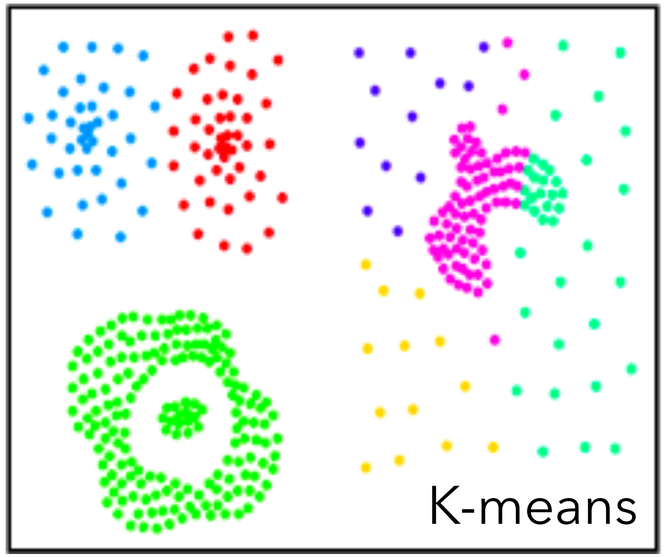
Do we have to use DBSCAN?

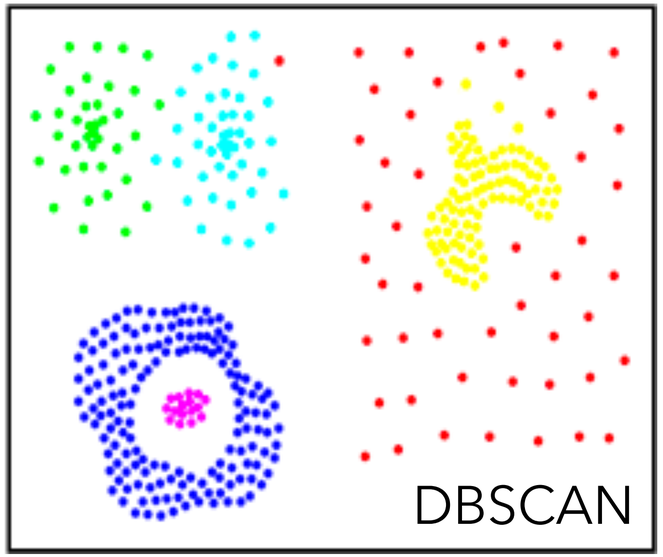
Advantages of DBSCAN (compare to K-means)

No need to **predefine the number of clusters**

Can be adaptable to **any random cluster shape**

Be able to **identify noise data** (a.k.a. outliers)

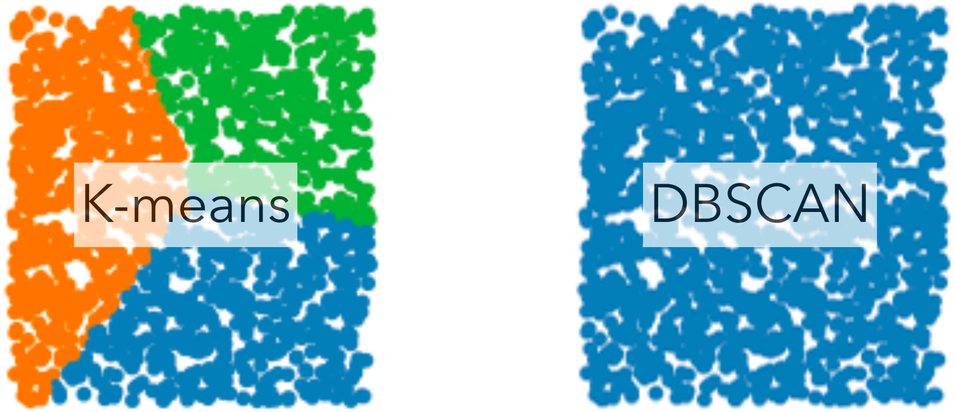




Do we have to use DBSCAN?

Disadvantages of DBSCAN

Struggles with clusters of **similar densities**



Curse of Dimensionality - hard to find optimal eps

Not entirely **deterministic** - how border point designated can be varies

References

- https://bcho.tistory.com/1205
- https://en.wikipedia.org/wiki/DBSCAN
- https://deep-eye.tistory.com/36
- https://untitledtblog.tistory.com/146
- https://jhryu1208.github.io/data/2020/12/26/ML_DBSCAN/
- https://gentlej90.tistory.com/29
- https://www.geeksforgeeks.org/difference-between-k-means-and-dbscan-clustering/
- https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80
- https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/
- https://www.analyticsvidhya.com/blog/2021/05/20-questions-to-test-your-skills-on-dbscan-clustering-algorithm/
- https://elutins.medium.com/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818
- https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html

Thank You 💚