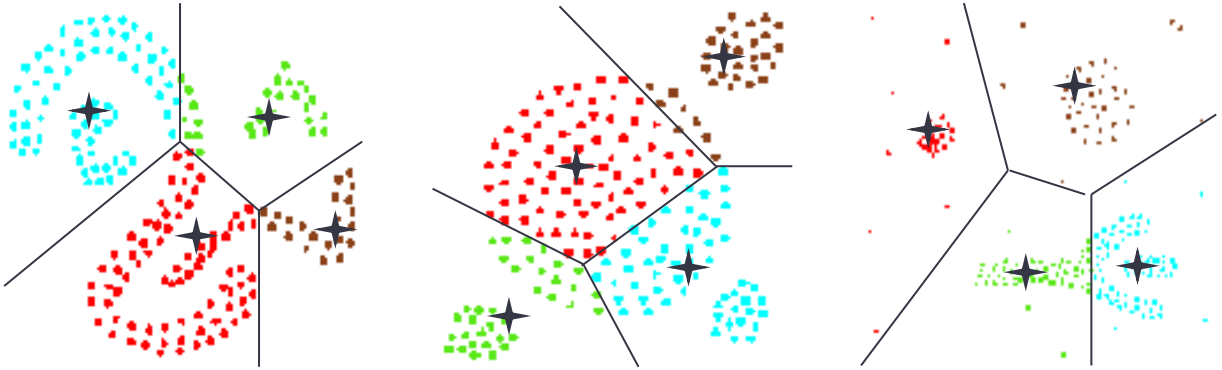
DBSCAN: Density-Based Spatial Clustering of Applications with Noise

Presented by Wondong Lee

Written by M.Ester, H.P.Kriegel, J.Sander and Xu.

Why Density-Based Clustering?

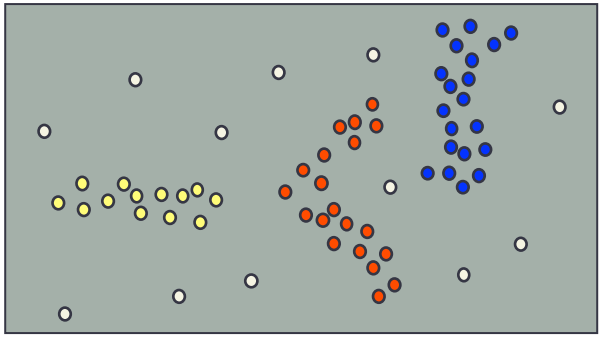
- Results of k-means algorithm for k = 4



➔The result is not satisfiable!!

DBSCAN

- Relies on a density-based notion of cluster
- Discovers clusters of arbitrary shape in spatial databases with noise
- Basic Idea
  - Group together points in high-density
  - Mark as outliers ➔ points that lie alone in low-density regions



DBSCAN

- Local point density at a point *p* defined by two parameters

(1) ε ➔ radius for the neighborhood of point p:

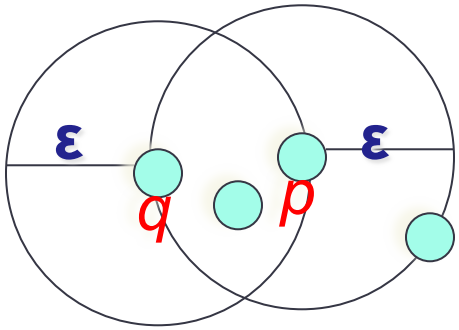
- ε-Neighborhood: all points within a radius of ε from the point p

$$N_{\varepsilon}(p) := \{q \text{ in data set } D \mid dist(p, q) \leq \varepsilon\}$$

(2) ***MinPts*** ➔ minimum number of points in the given neighborhood *N*(*p*)

High Density?

- ε-Neighborhood of an point contains at least *MinPts*



ε-Neighborhood of *p*
ε-Neighborhood of *q*

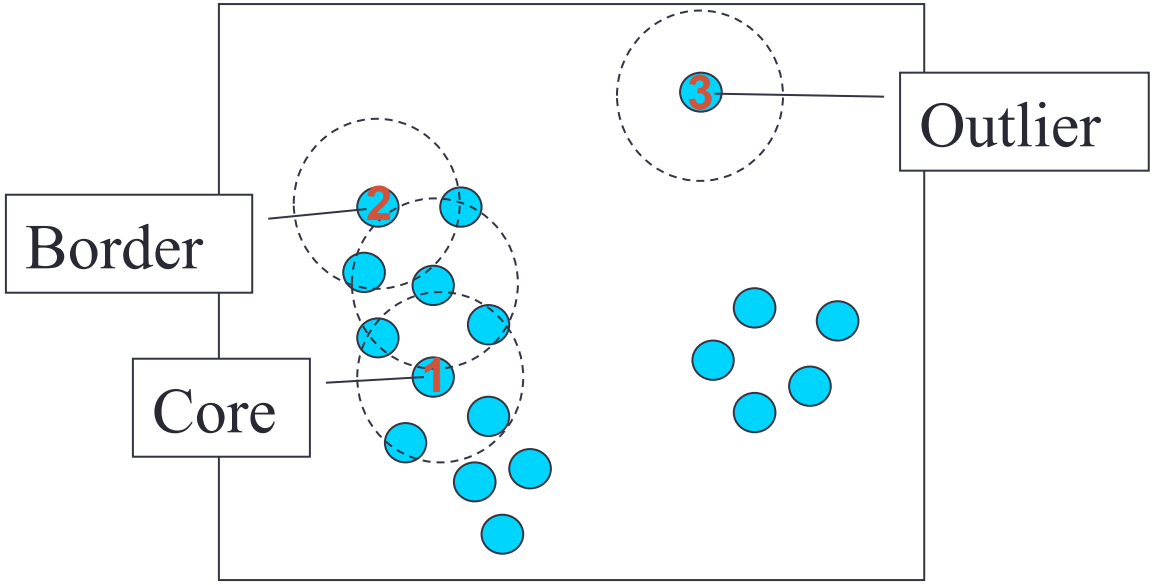
Q. When MinPts = 4 ?

Density of p is "high"

Density of q is "low"

Core, Border & Outlier

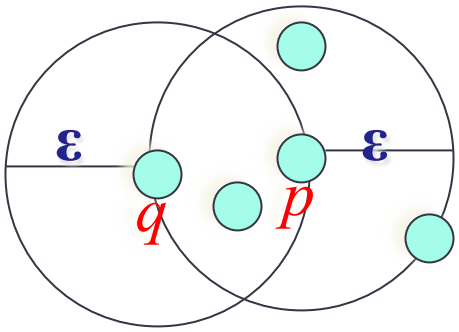
- Three category for each point
  - Core point: if its density is high
  - Border point: density is low (but in the neighborhood of a core point)
  - Noise point: any point that is not a core point nor a border point



Density-Reachability

- Directly density-reachable
  - A point q is **directly density-reachable** from a point p:

  -If p is a core point and q is in p's ε-neighborhood



Minpts = 4

Q. p is directly density-reachable from q?

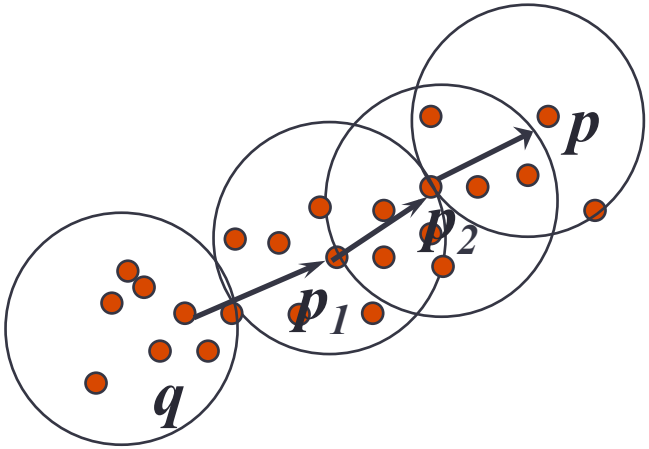
No, why?

Q. Density-reachability is **asymmetric**

Density-Reachability

- Density-reachable
  - A point p is **density-reachable** from a point q

  -If there is a chain of points $p_1, \ldots, p_n$, with $p_1 = q$, $p_n = p$ such that p_{i+1} is directly density-reachable from p_i



MinPts = 7

- p1 is directly density-reachable from q
- p2 is directly density-reachable from p1
- p is directly density-reachable from p2
- There is a chain from q to p (q→p1→p2→q)

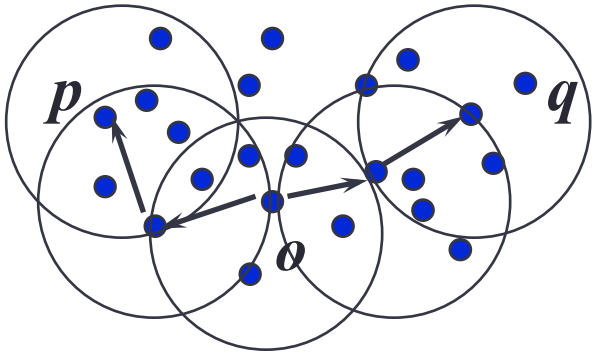
Q. q is density-reachable from p?

No, why?

Density-Connectivity

- Density-connected
  - A pair of points p and q are density-connected

  -If they are commonly density-reachable from a point o



MinPts = 7

Q. o is density-reachable from p?

Yes, why?

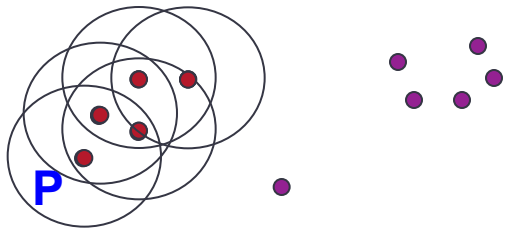
Q. Density-connectivity is symmetric

Formal Description of Cluster

- Given a data set D, parameter ε and *MinPts,*
- A cluster C is a subset of D satisfying two criteria:
  - **Maximality**
    - $\forall$ *p*, *q* if $p \in \boldsymbol{C}$ and if *q* is density-reachable from *p*, then also $q \in \boldsymbol{C}$
  - **Connectivity**
    - $\forall$ $p, q \in \boldsymbol{C}$, *p* and *q* are density-connected
- ***Note:*** cluster contains *core points* as well as *border points*

Parameter

- ε = 2, *MinPts* = 3

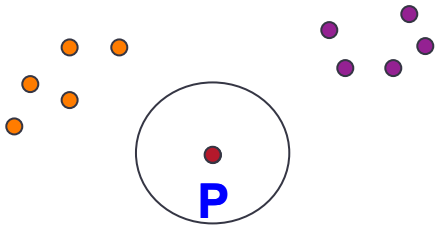


if *p* is not yet classified **then**
  if *p* is a core-point **then**
    collect all points density-reachable from *p*
    and assign them to a new cluster.
  else
    assign *p* to NOISE

exmaple)

Parameter

- ε = 2, *MinPts* = 3

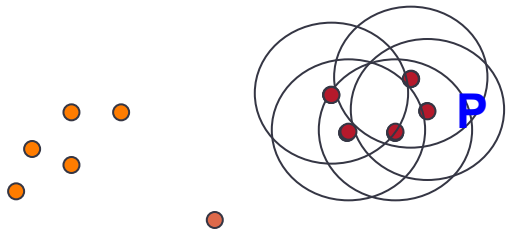


$\in D$ **do**
if p is not yet classified **then**
if p is a core-point **then**
collect all points density-reachable from p
and assign them to a new cluster.
else
assign p to NOISE

exmaple)

Parameter

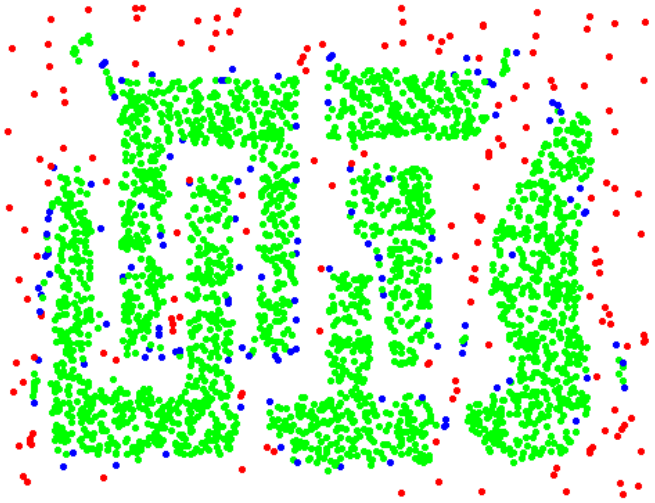
- ε = 2, *MinPts* = 3



$\in D$ **do**
if p is not yet classified **then**
if p is a core-point **then**
collect all points density-reachable from p
and assign them to a new cluster.
else
assign p to NOISE

Example

Original Points

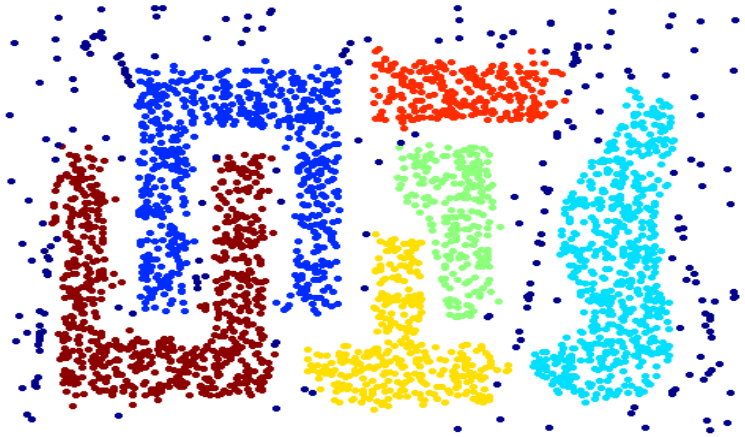


Point types: core, border and outliers

ε = 10, MinPts = 4

When DBSCAN Works Well

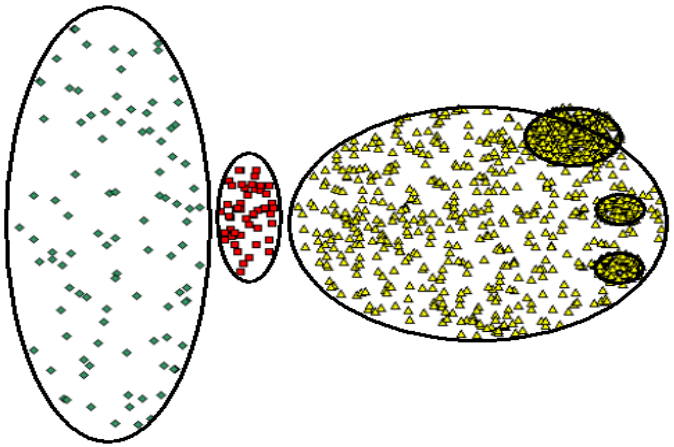
- Resistant to Noise
- Can handle clusters of different shapes and sizes



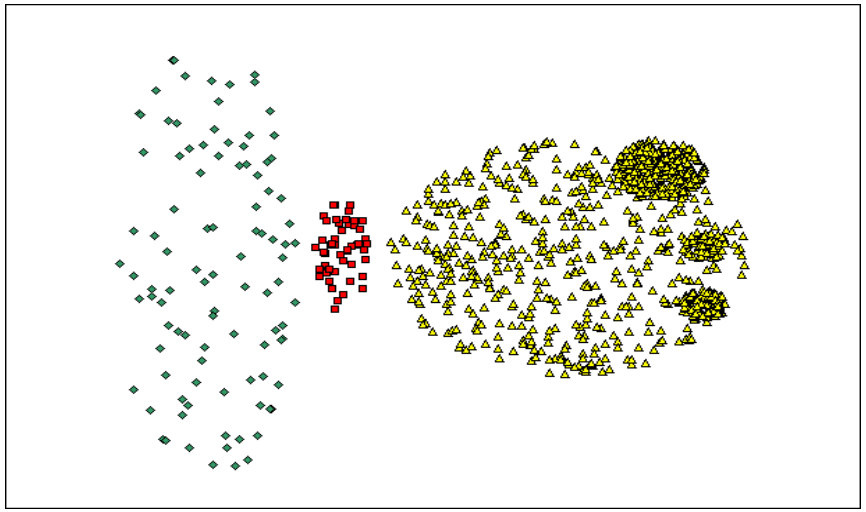
Clusters

When DBSCAN Does Not Work Well

- Cannot handle varying densities



Original Points



(ε =9.92, MinPts=4)

When DBSCAN Does Not Work Well

- Sensitive to parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

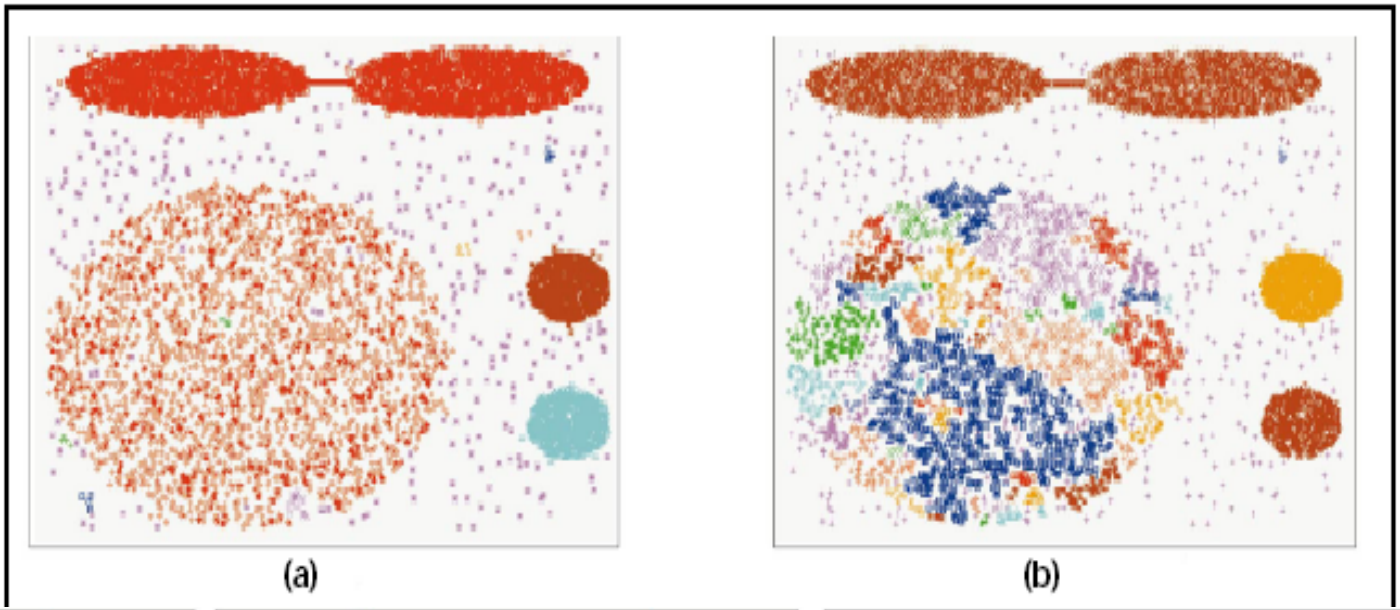
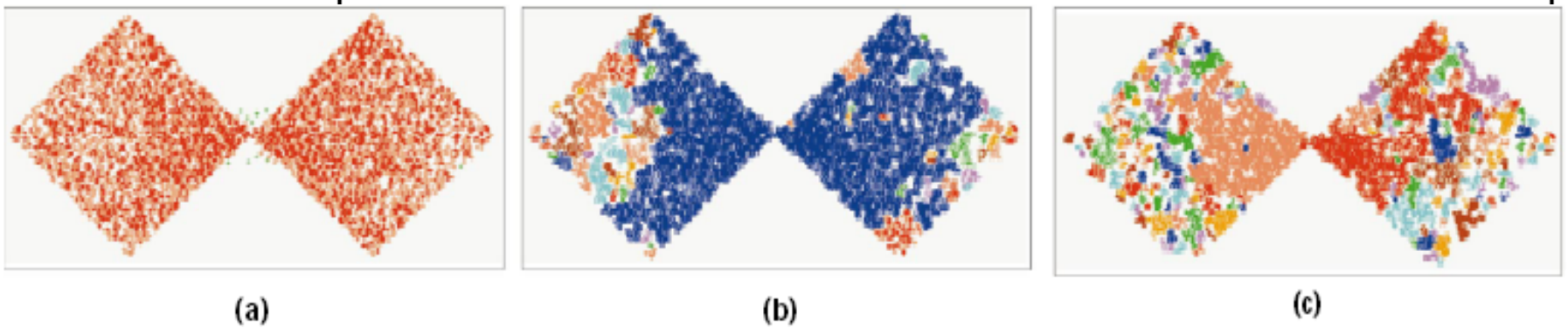
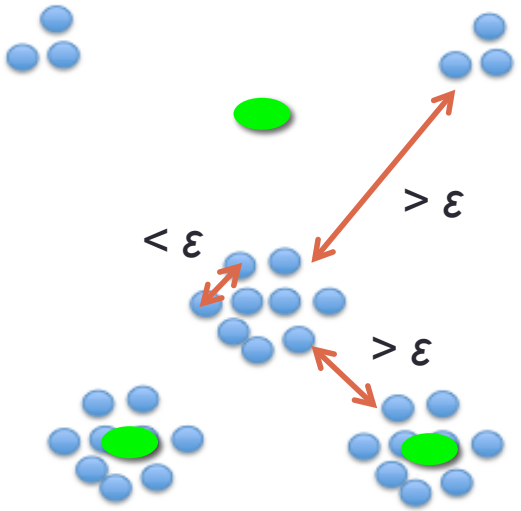


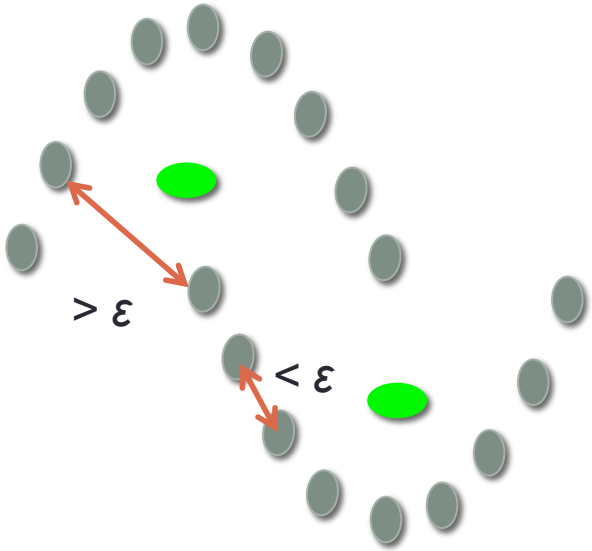
Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



K-means VS DBSCAN



: Initial center

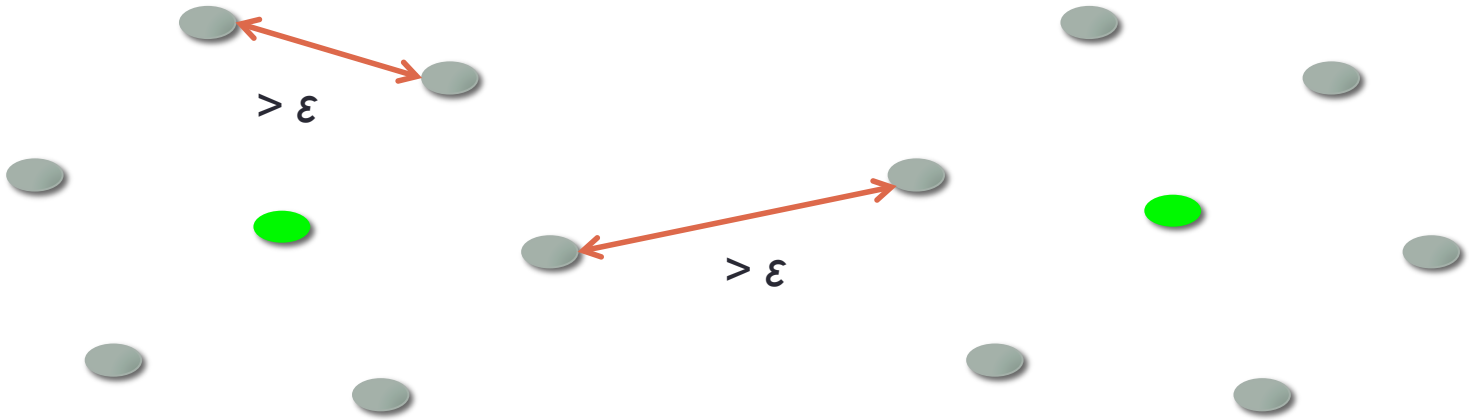


(1) When k = 3
MinPts = 4

(2) When k = 2
MinPts = 3

Winner is DBSCAN

K-means VS DBSCAN



: Initial center

(1) When k = 2
MinPts = 3

Winner is K-means

Thank you for attention

Any Questions?

Reference

- Comparing Clustering Algorithm
  - http://www.cise.ufl.edu/~jmishra/clustering/DataMiningPresentation.ppt
- Density-Based Clustering
  - http://www.cse.buffalo.edu/faculty/azhang/cse601/density-based.ppt
- Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." *Kdd*. Vol. 96. No. 34. 1996.