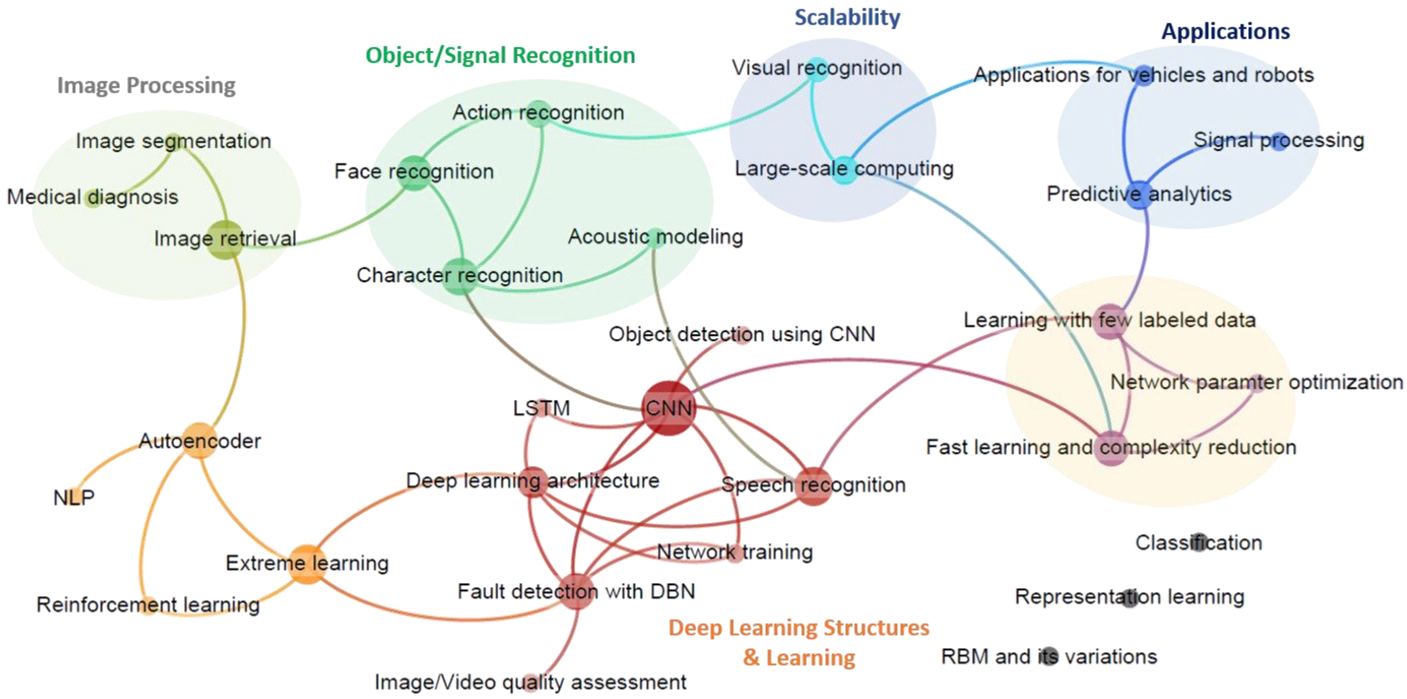


# Lecture 9: Clustering

Pilsung Kang

School of Industrial Management Engineering

Korea University

# AGENDA

01 Clustering: Overview

02 K-Means Clustering

03 Hierarchical Clustering

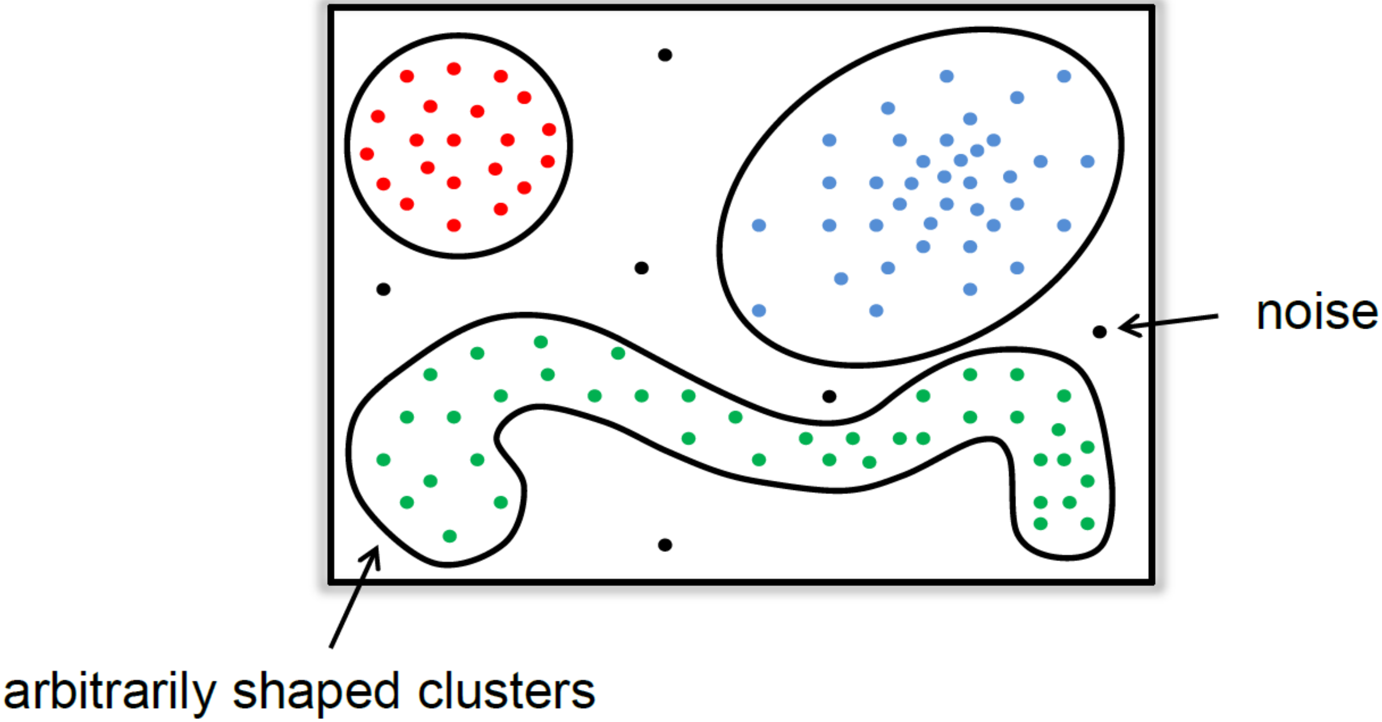
04 Density-based Clustering: DBSCAN

04 R Exercise

# Density-based Clustering

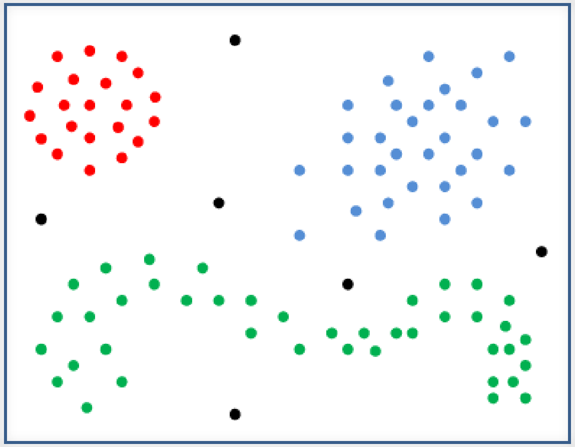
Ester et al. (1996)

- Density-based clustering
  - ✓ Conduct a clustering by considering the density of data points
    - Can find an arbitrary shape of cluster
    - Can remove noise from clustering result



# Density-based Clustering

- DBSCAN
  - ✓ Most popular density-based clustering algorithm
- Idea
  - ✓ Clusters are the collections of data points with high density
  - ✓ Density around a noise point is very low
- Purpose
  - ✓ Quantify the features of clusters and noise points to find a set of valid clusters



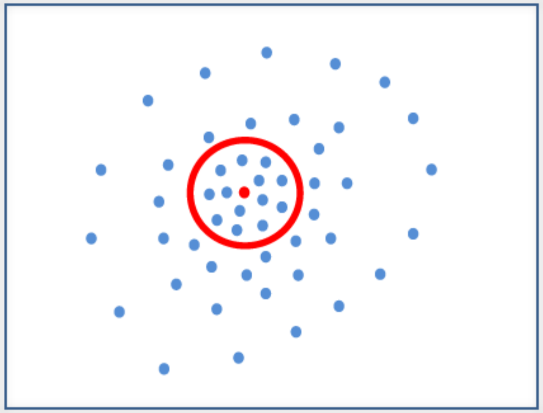
# Density-based Clustering

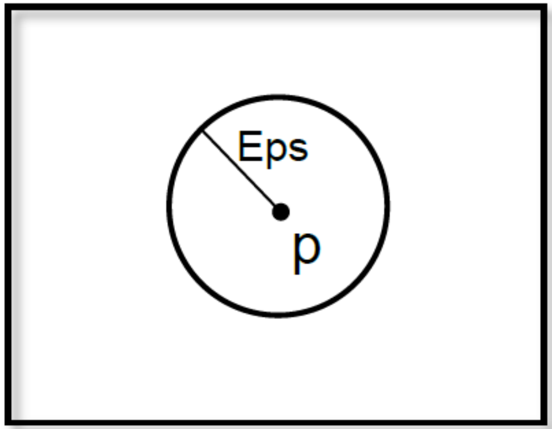
- DBSCAN

  ✓ Definition 1: ε-neighborhood of a point

  - The ε-neighborhood of a point, denoted by $N_\varepsilon(p)$, is defined by

$$N_\epsilon(p) = \{q \in D \mid dist(p, q) \leq \epsilon\}$$

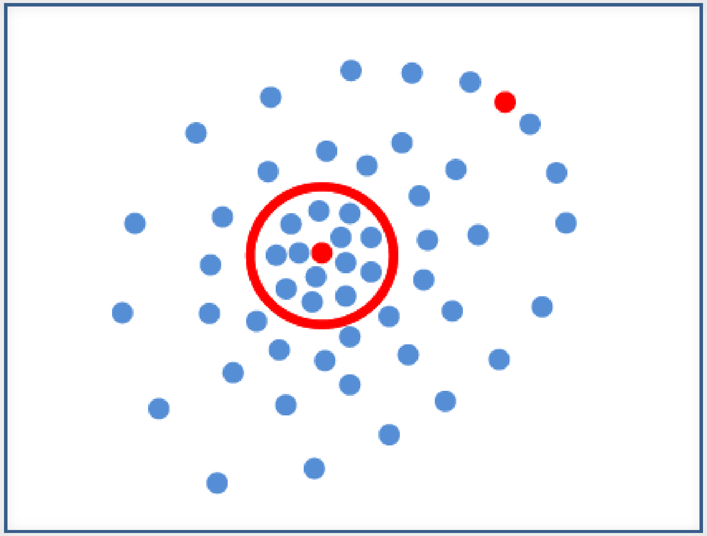




✓ Naïve Approach: require for each point in a cluster that there are at least a minimum number (MinPts) of points in an ε-neighborhood of that point

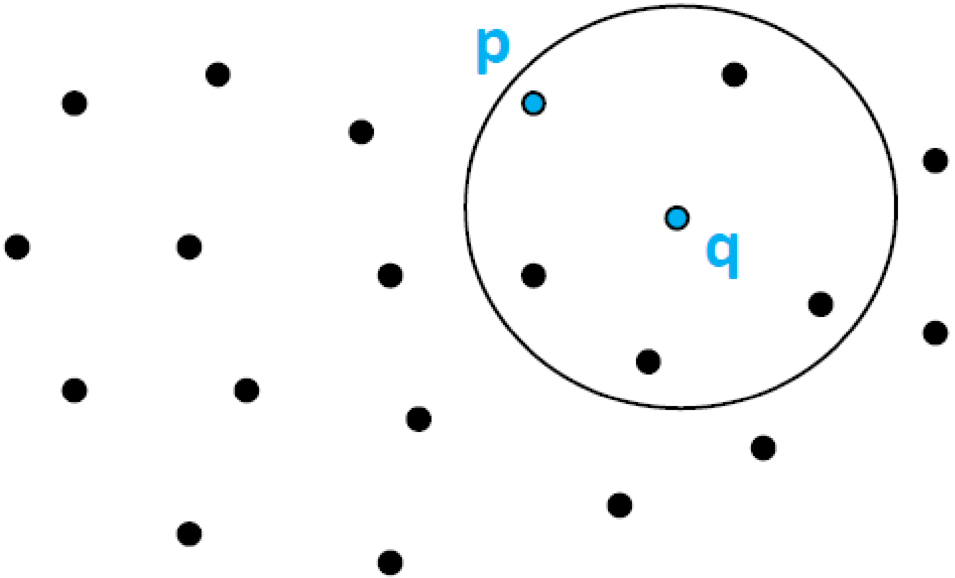
# Density-based Clustering

- DBSCAN
  - ✓ Problem of Naïve Approach
    - There are two kinds of points in a cluster
      - Points inside of the cluster (core points)
      - Points on the border of the cluster (border points)
    - An ε-neighborhood of a border point contains significantly less points than an ε-neighborhood of a core point



# Density-based Clustering

- DBSCAN
  - ✓ Better idea
    - For every point p in a cluster C, there is a point q in C so that p is inside of the ε-neighborhood of q (Border points are connected to core points)
    - $N_\varepsilon(q)$ contains at least MinPts points (Core points = high density)

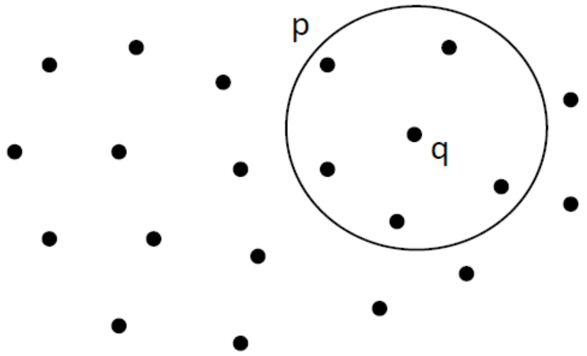


# Density-based Clustering

- DBSCAN
  - ✓ Definition 2: directly density-reachable
    - A point p is directly density-reachable from a point q with regard to the parameters ε and MinPts, if

$$1)\quad p \in N_\epsilon(q) \quad (reachability)$$

$$2)\quad |N_\epsilon(q)| \geq MinPts \quad (core\ point\ condition)$$

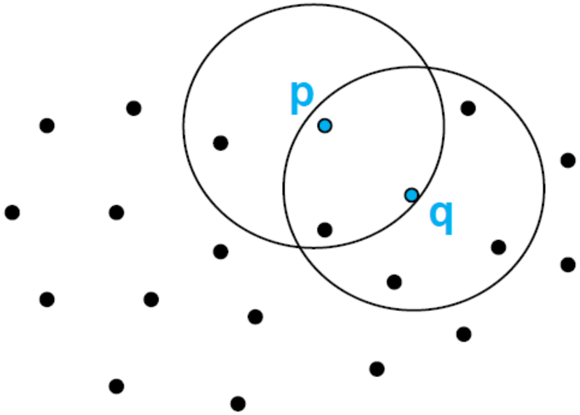


MinPts = 5

$| N_{Eps}(q) | = 6 \geq 5 = MinPts$ (core point condition)

# Density-based Clustering

- DBSCAN
  - ✓ Property
    - Directly density-reachable is symmetric for pairs of core points
    - It is not symmetric if one core point and one border point are involved



**Parameter: MinPts = 5**

p directly density reachable from q

$p \in N_{Eps}(q)$

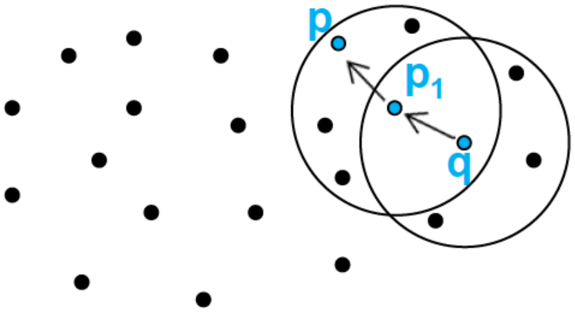
$| N_{Eps}(q) | = 6 \geq 5 = MinPts$ (core point condition)

q **not** directly density reachable from p

$| N_{Eps}(p) | = 4 < 5 = MinPts$ (core point condition)

# Density-based Clustering

- DBSCAN
  - ✓ Definition 3: density-reachable
    - A point p is density-reachable from a point q with regard to the parameters ε and MinPts, if there is a chain of points $p_1, p_2, \ldots, p_s$ with $p_1 = q$ and $p_s = p$ such that $p_{i+1}$ is directly density-reachable from $p_i$ for all $1 < I < s-1$



MinPts = 5

$|N_{Eps}(q)| = 5 = MinPts$ (core point condition)

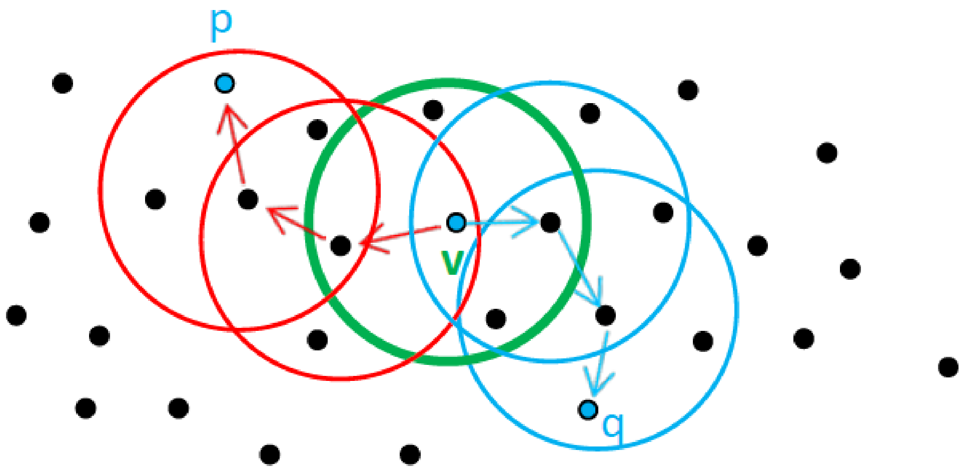
$|N_{Eps}(p_1)| = 6 \geq 5 = MinPts$ (core point condition)

# Density-based Clustering

- DBSCAN

  ✓ Definition 4: density-connected

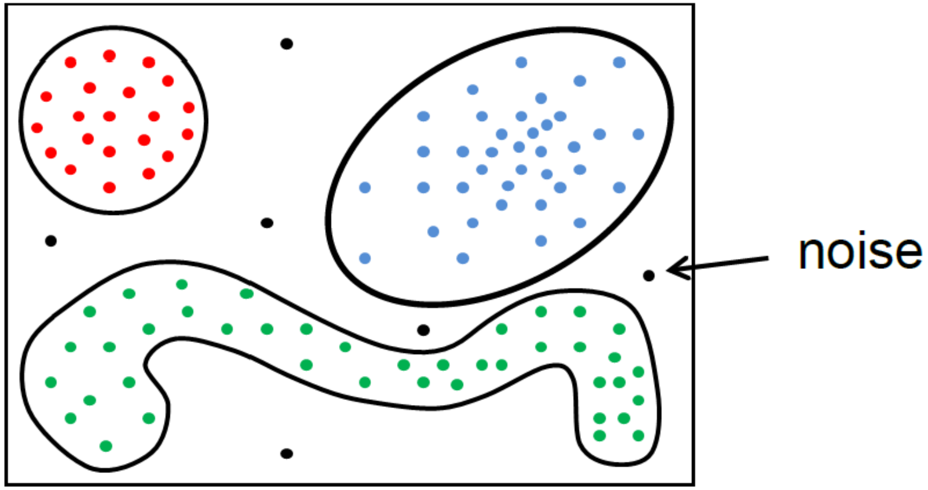
  - A point p is density-connected to a point q with regard to the parameters ε and MinPts, if there is a point v such that both p and q are density-reachable from v



MinPts = 5

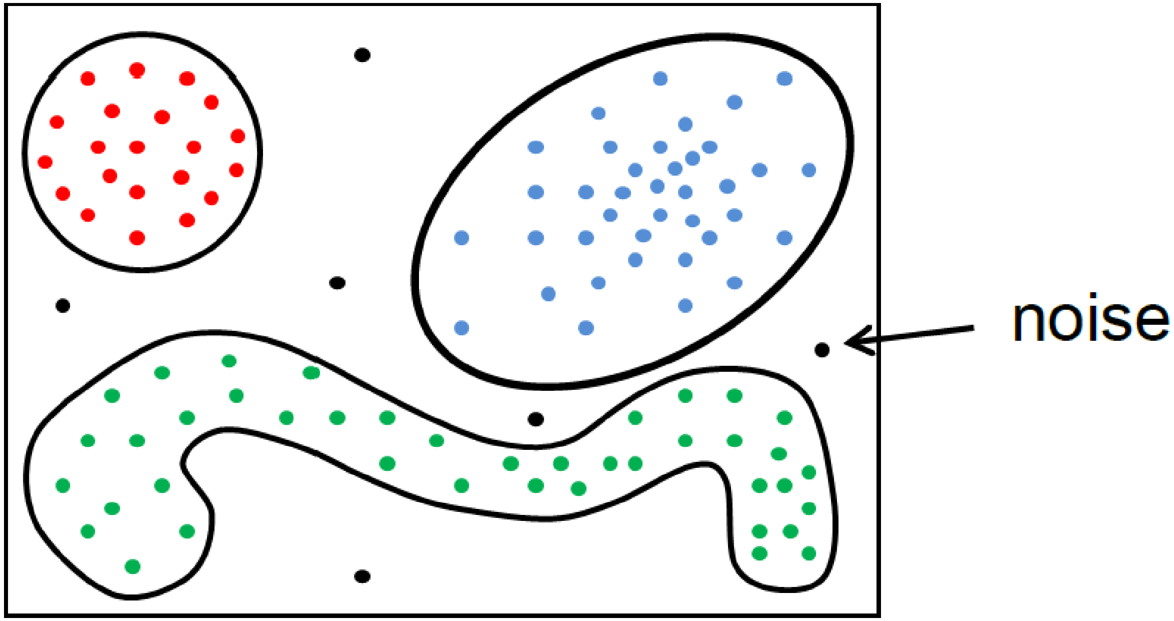
# Density-based Clustering

- DBSCAN
  - ✓ Definition 5: Cluster
    - A cluster with regard to the parameters ε and MinPts is a non-empty subset C of the database D with
      - (1) For all p, q ∈ D: If p ∈ C and q is density-reachable from p with regard to the parameters ε and MinPts, then q ∈ C (Maximality)
      - (2) For all p, q ∈ C: The point p is density-connected to q with regard to the parameters ε and MinPts (Connectivity)



# Density-based Clustering

- DBSCAN
  - ✓ Definition 6: Noise
    - Let $C_1, \ldots, C_k$ be the clusters of the database D with regard to the parameters ε and MinPts
    - The set of points in the database D not belonging to any cluster $C_1, \ldots, C_k$ is called noise



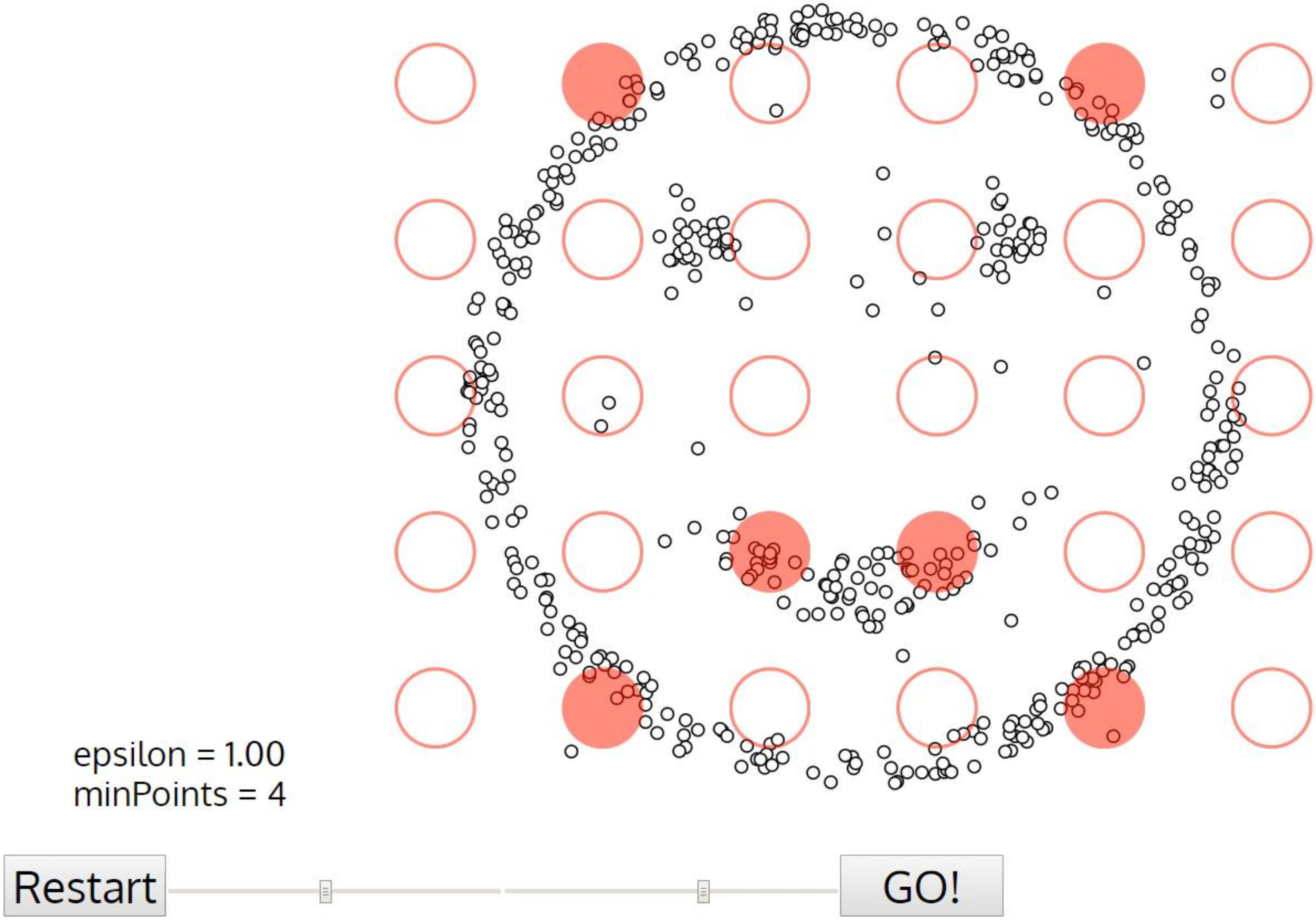
# Density-based Clustering

- DBSCAN: Algorithm
  - ✓ Input: N objects to be clustered and global parameter, ε and MinPts
  - ✓ Output: Cluster of objects

- Algorithm
  - ✓ Arbitrary select a point p
  - ✓ Retrieve all points density-reachable from p w.r.t. ε and MinPts
  - ✓ If p is a core points, a cluster is formed
  - ✓ If p is a border point, no points are density reachable from p and DBSCAN visits the next point of the database
  - ✓ Continue the process until all of the points have been processed

# Density-based Clustering

- DBSCAN example

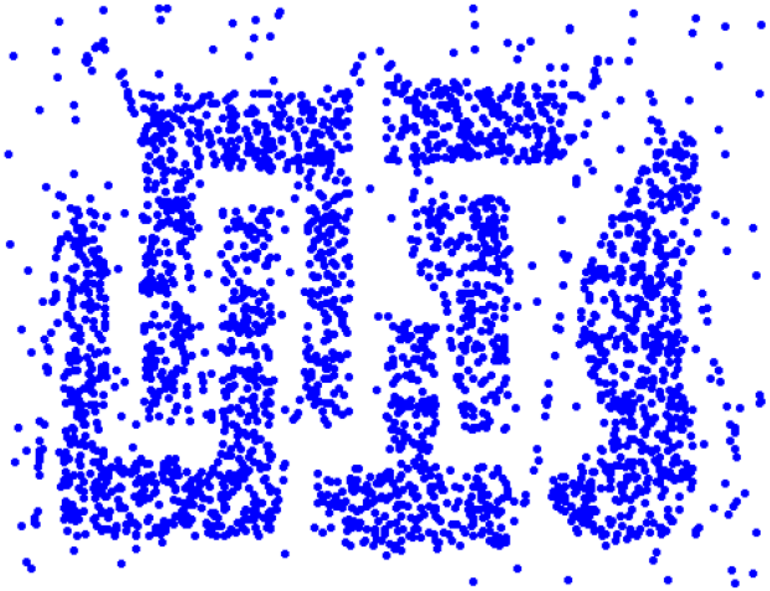


epsilon = 1.00
minPoints = 4

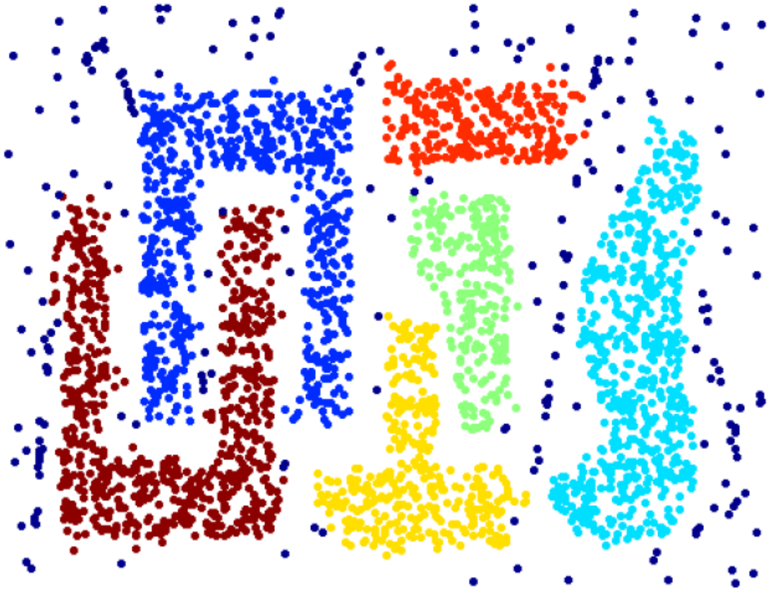
https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

# Density-based Clustering

- DBSCAN example



Original Points



Clusters

ANY
questions?