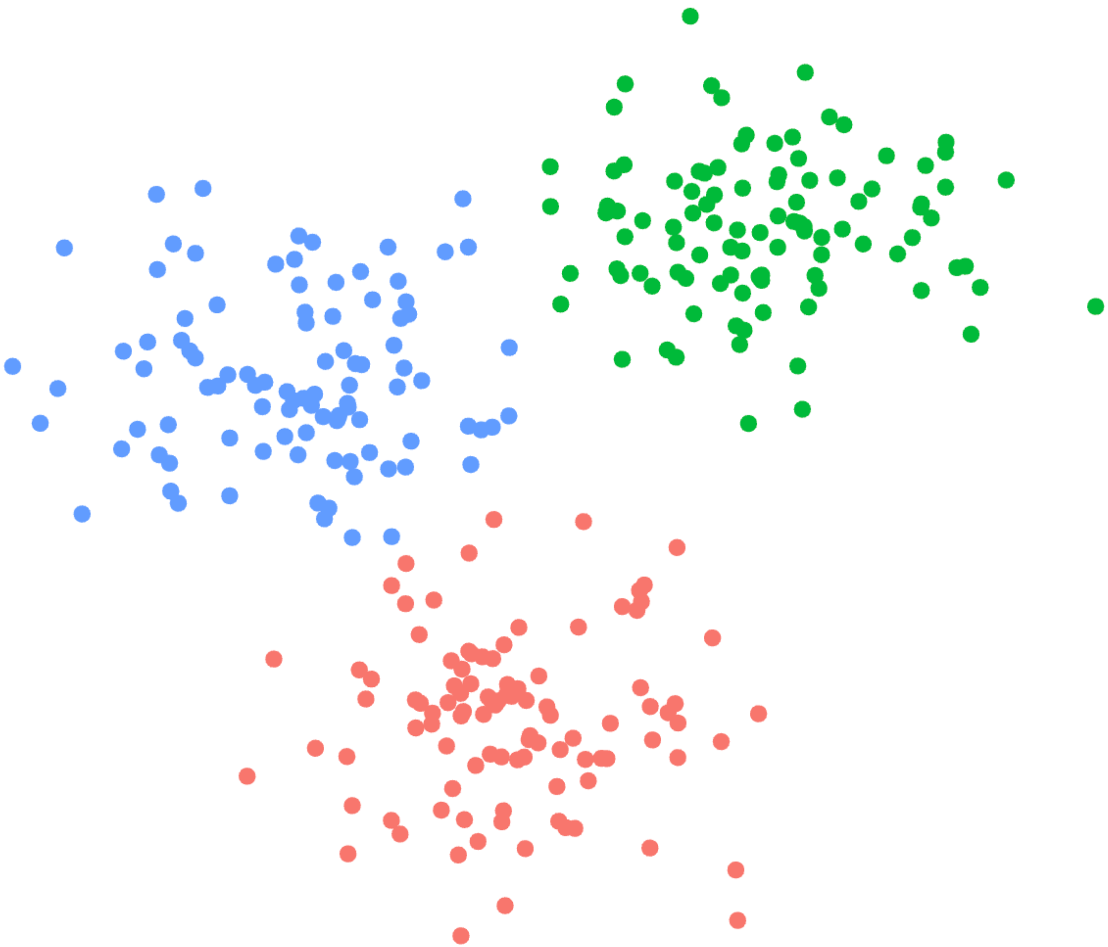
# CSCI 550: Adv. Data Mining

## 09- Density Based Clustering

# Announcement

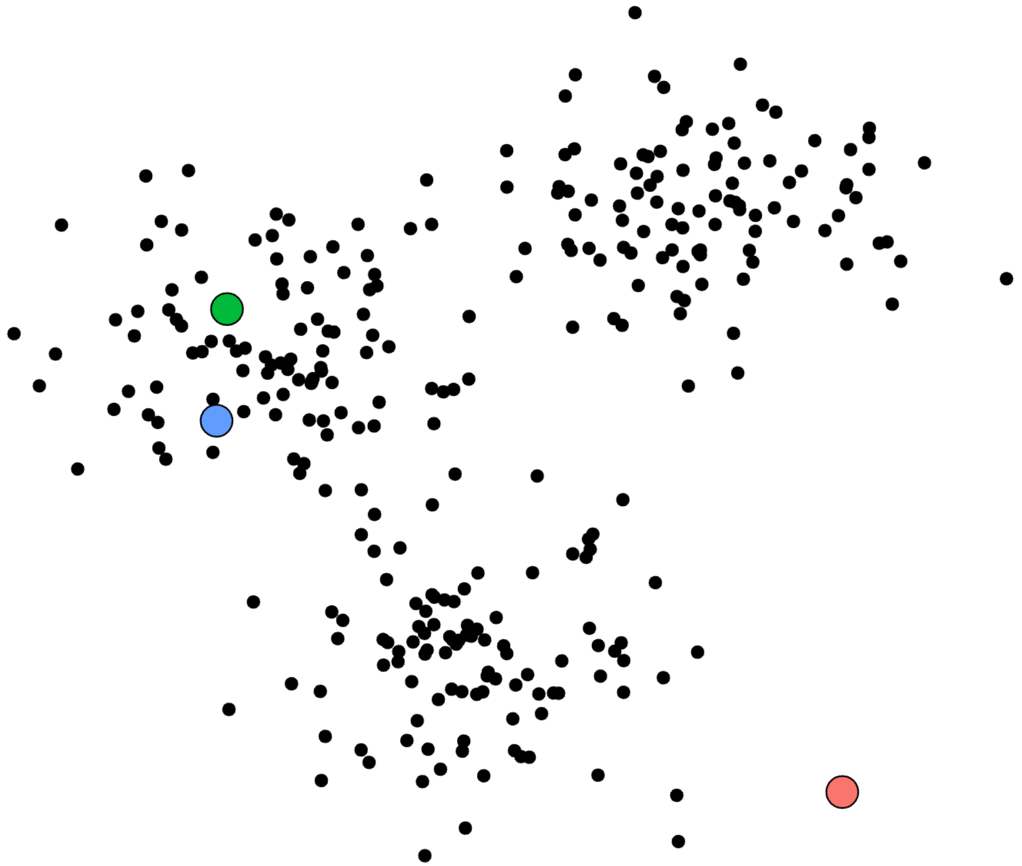
- Student lecture proposals are due by Oct 2$^{nd}$

# Look at this sample labeled data set



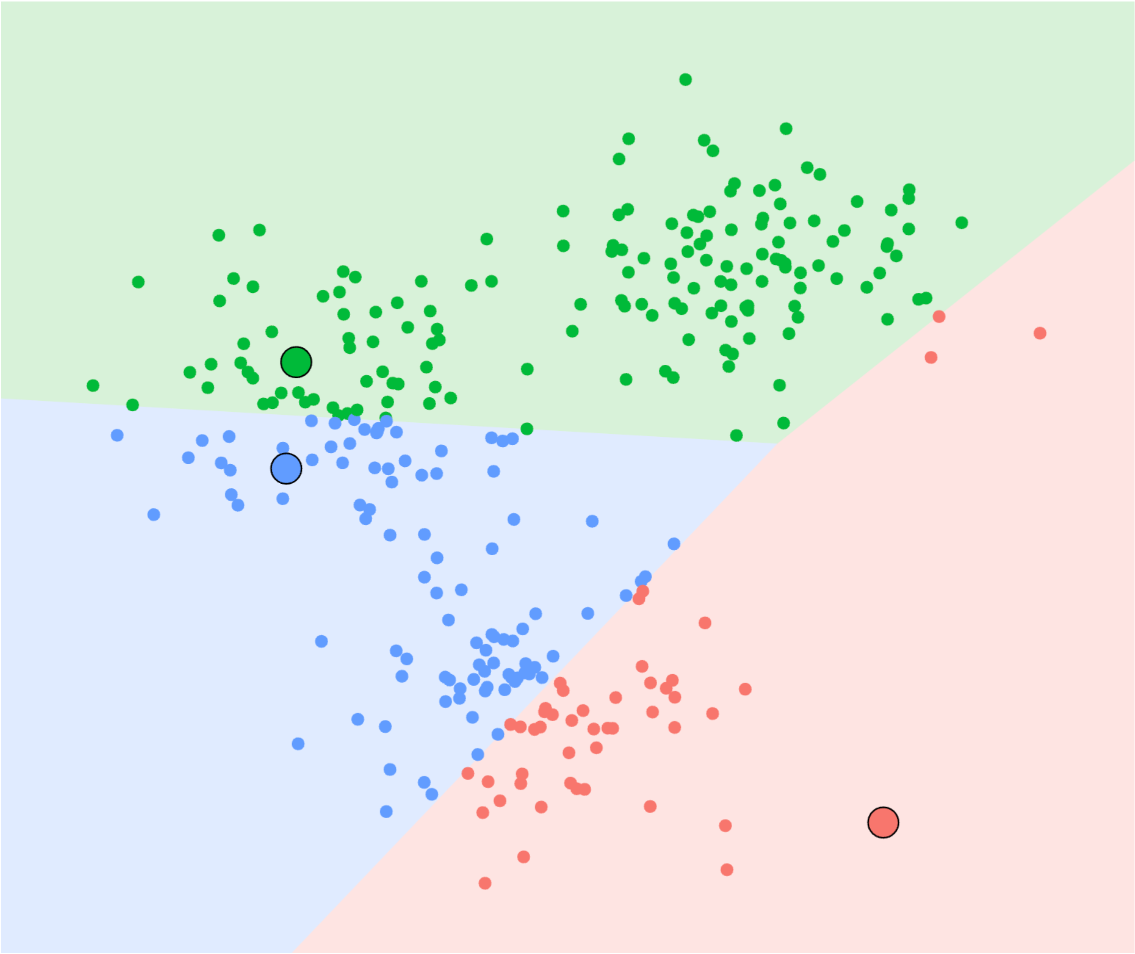
# How K-means cluster this data with K=3?

1. Start with *k* randomly chosen Centroids



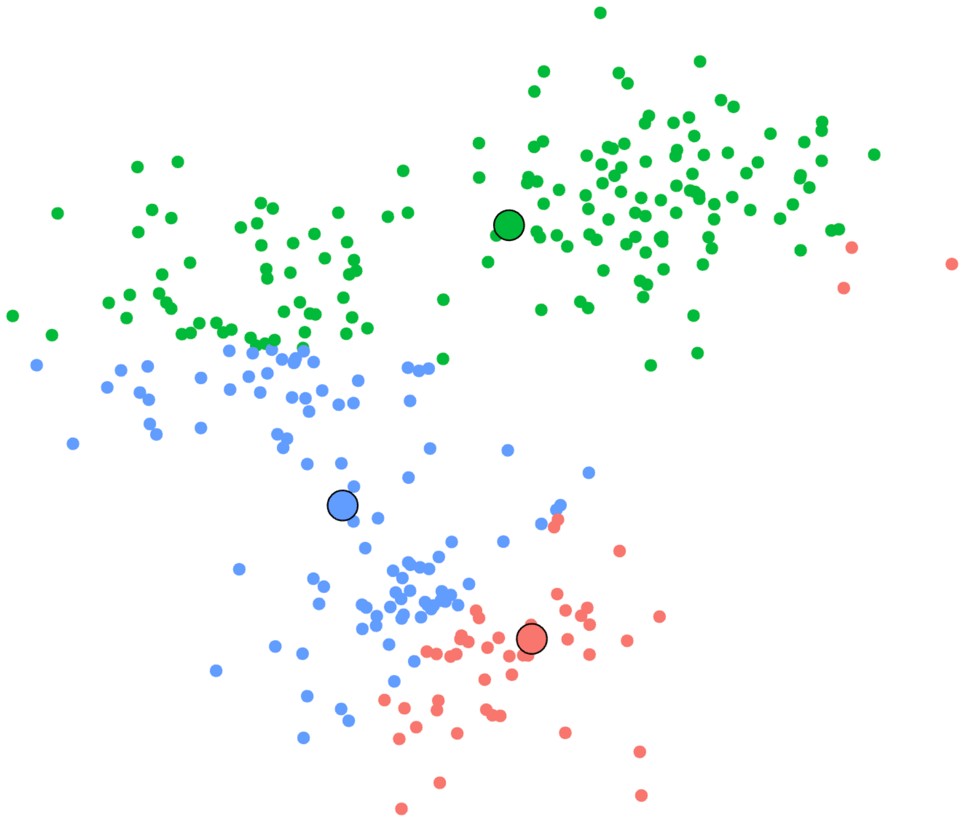
# How K-means cluster this data with K=3?

2. Assign data points to clusters by the shortest distance to any mean



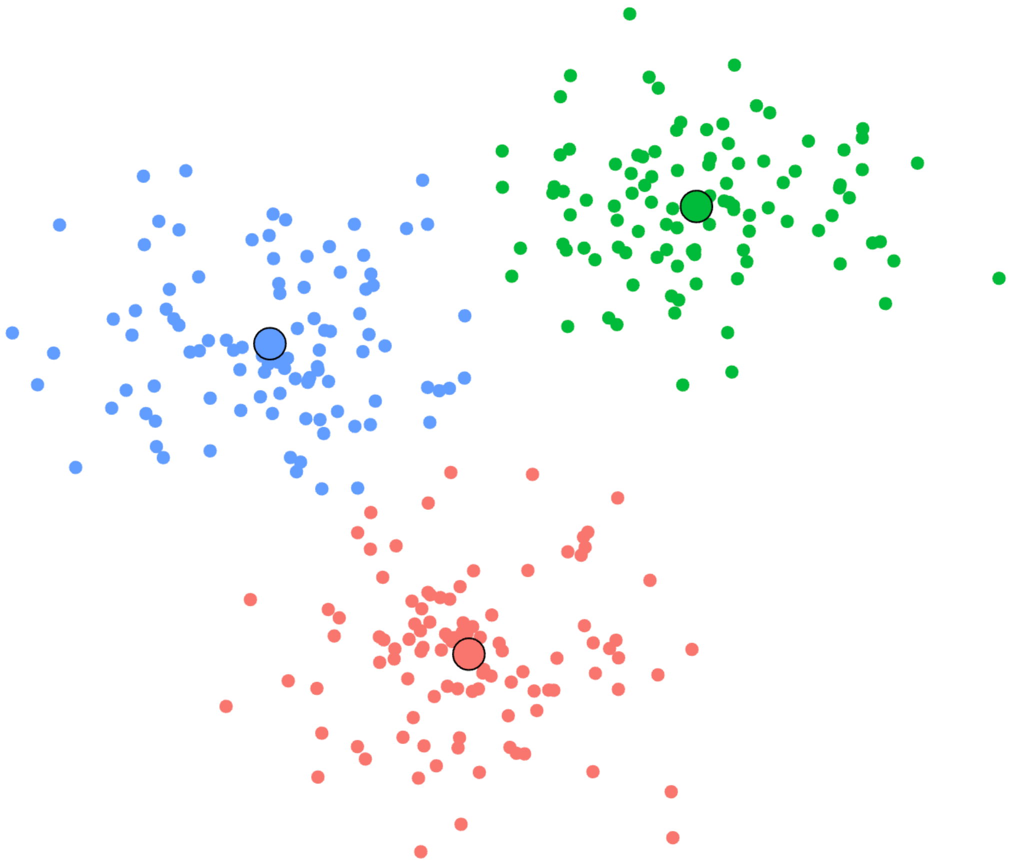
# How K-means cluster this data with K=3?

3. Update centroids

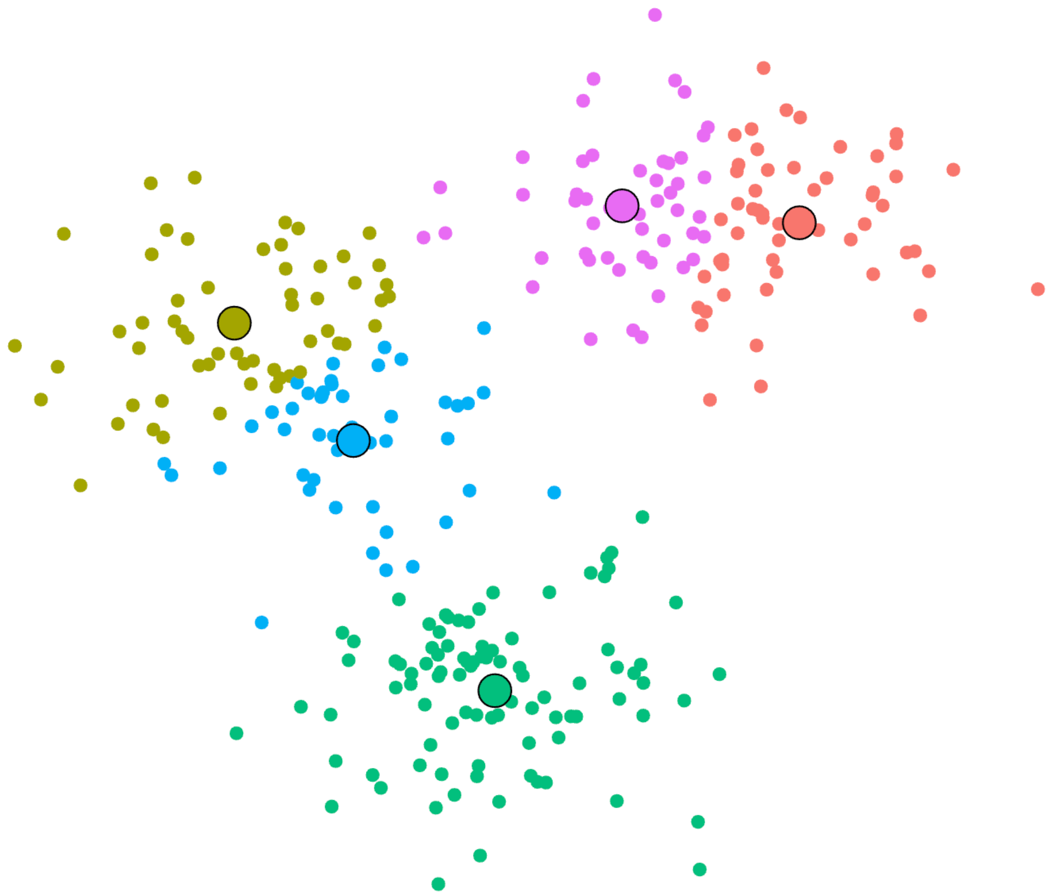


# How K-means cluster this data with K=3?

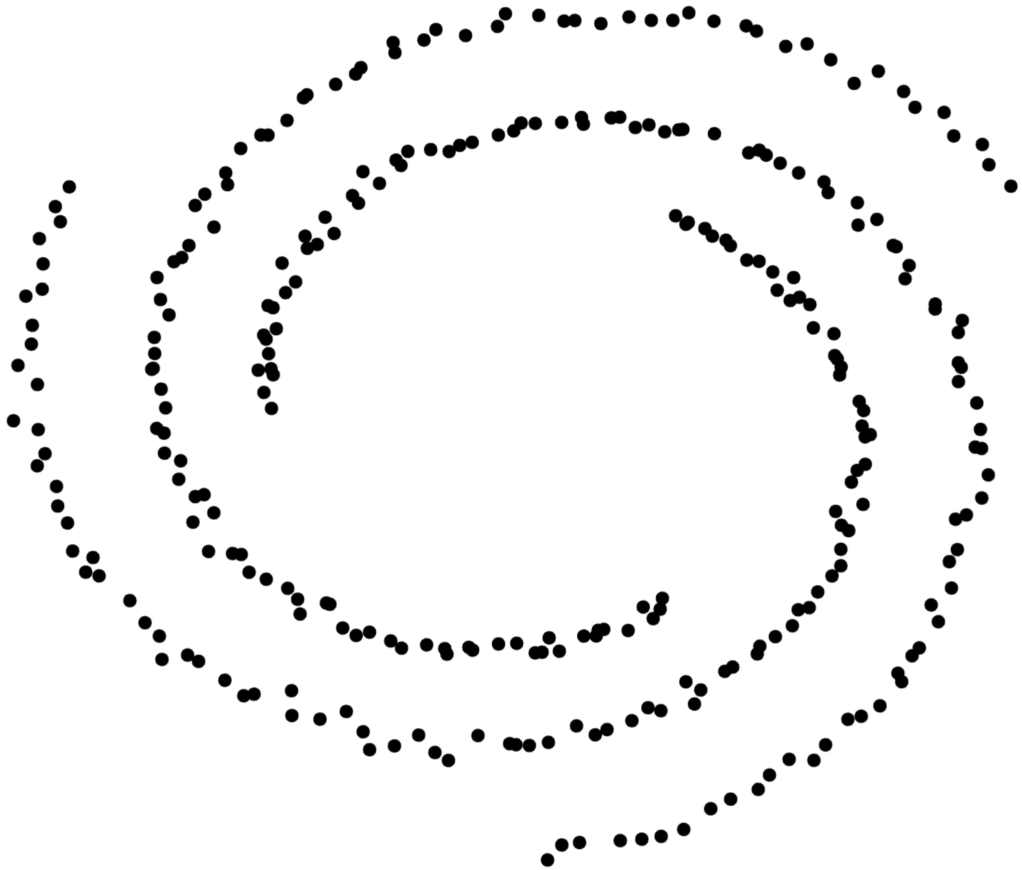
4. Repeat from step 2 until convergence



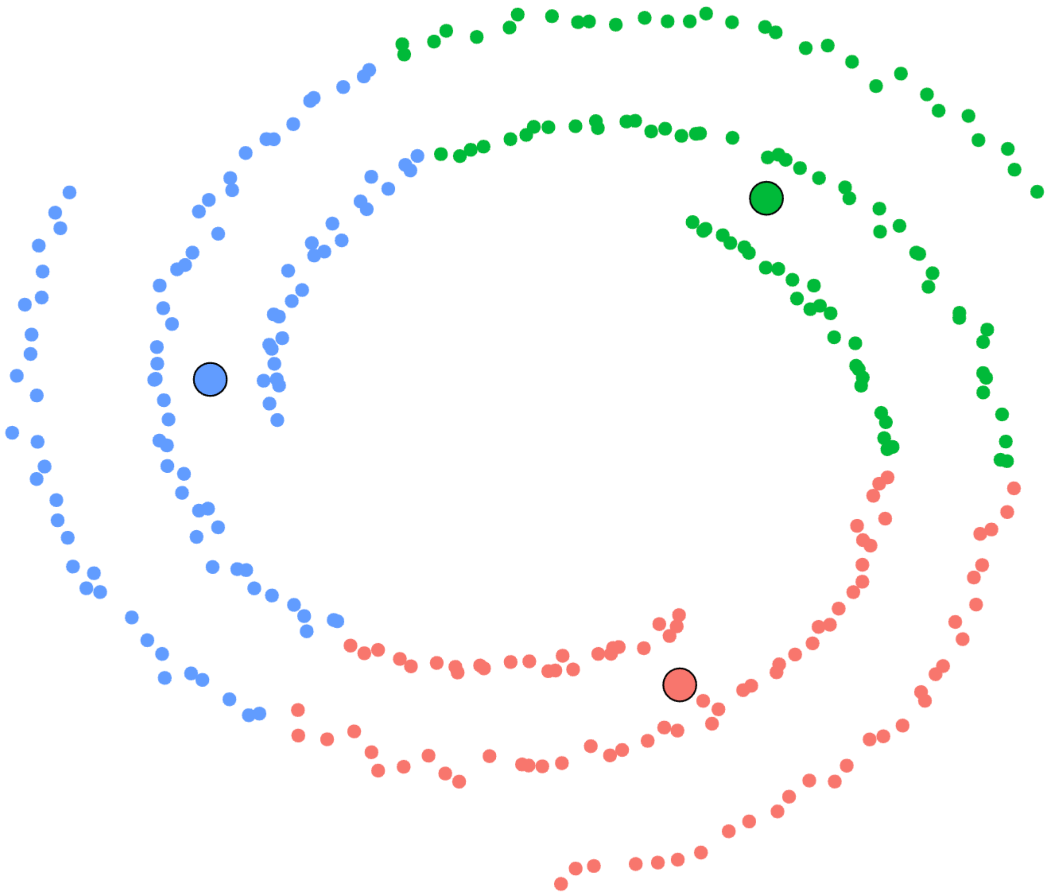
# How K-means cluster this data with K=5?



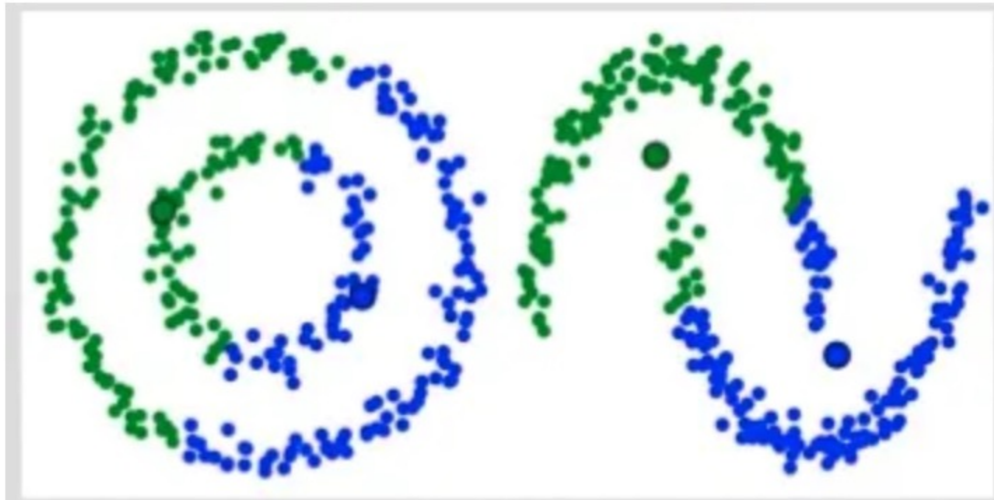
# How K-means cluster this dataset?

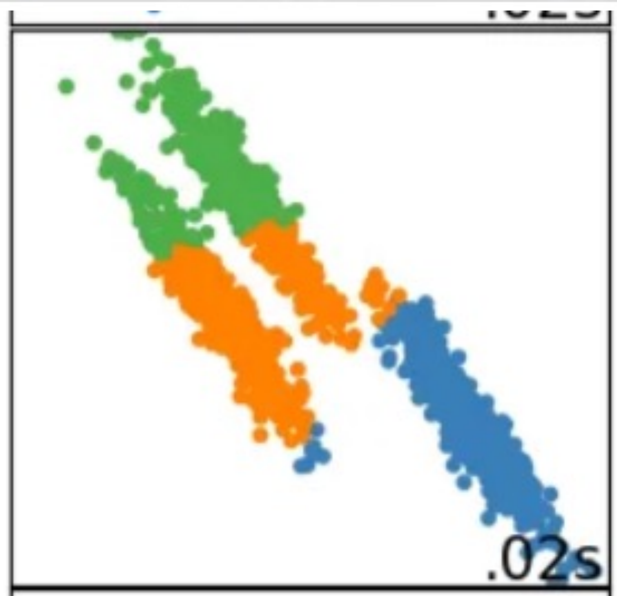


# How K-means cluster this dataset?



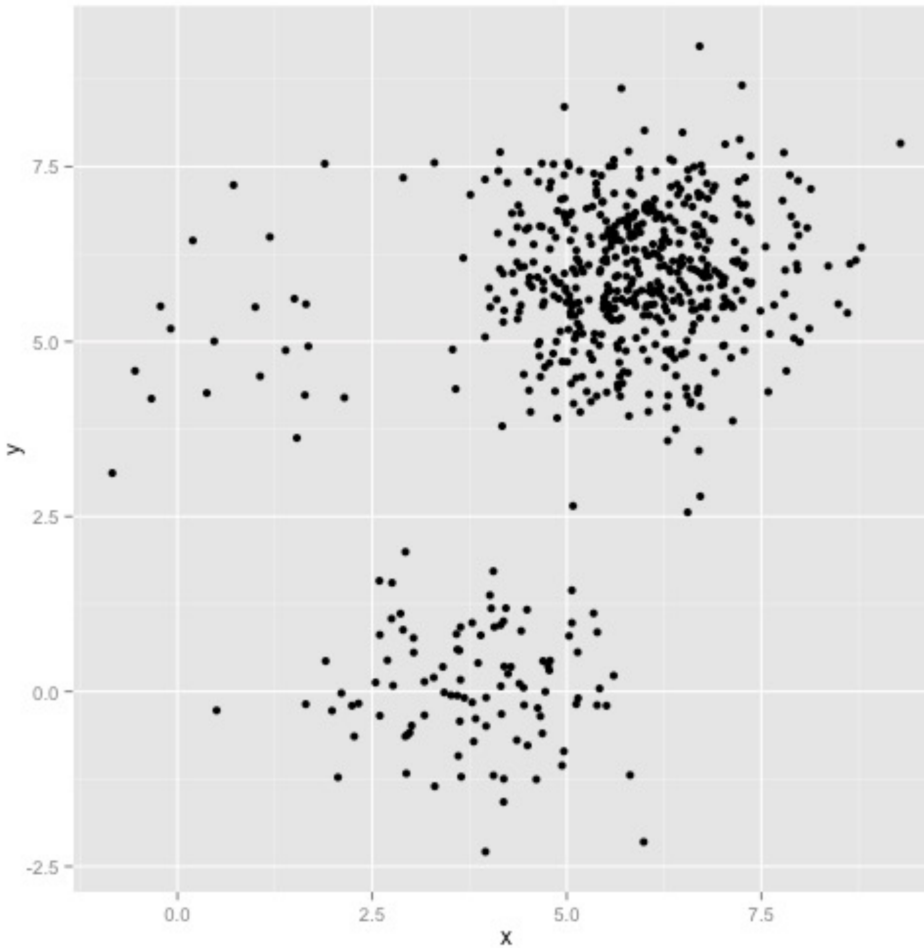
# How about these cases?





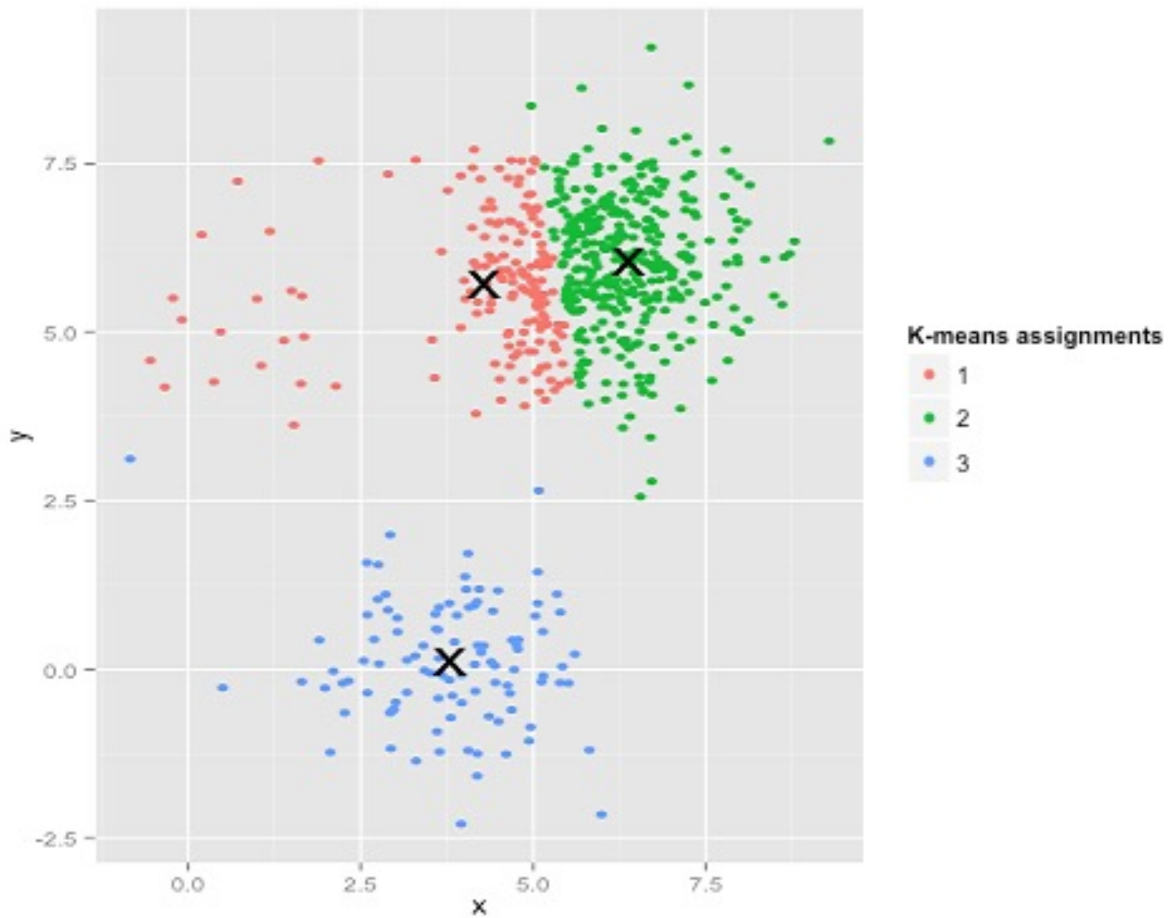
# Look at this example

- How many clusters do you detect?

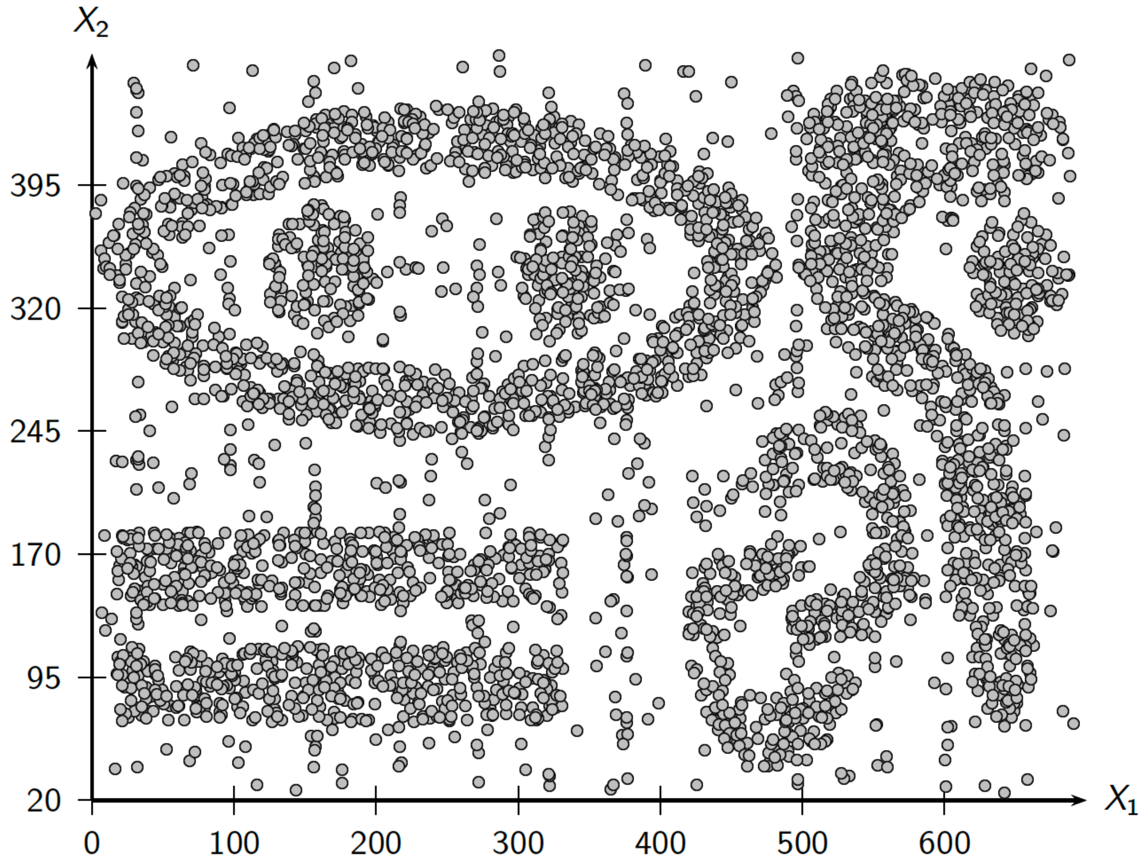


# Look at this example

- K-means results



# What about this synthetic dataset?



# K-means limitation

- It assumes the clusters are spherical shape (convex or ellipsoid-shaped)
- It is sensitive to outliers
- When the clusters are non-convex, two points in two neigborhood clusters might be closer than two points in same cluster.
- Density-based methods are able to mine nonconvex clusters, where distance-based methods may have difficulty.

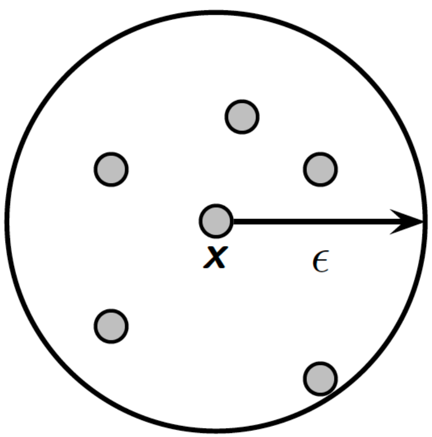
# The DBSCAN Approach

- Density-based Spatial Clustering of Applications with Noise (DBSCAN)
- Define a ball of radius ε around a point $x \in R^d$ , called that ε-neighborhood of x:

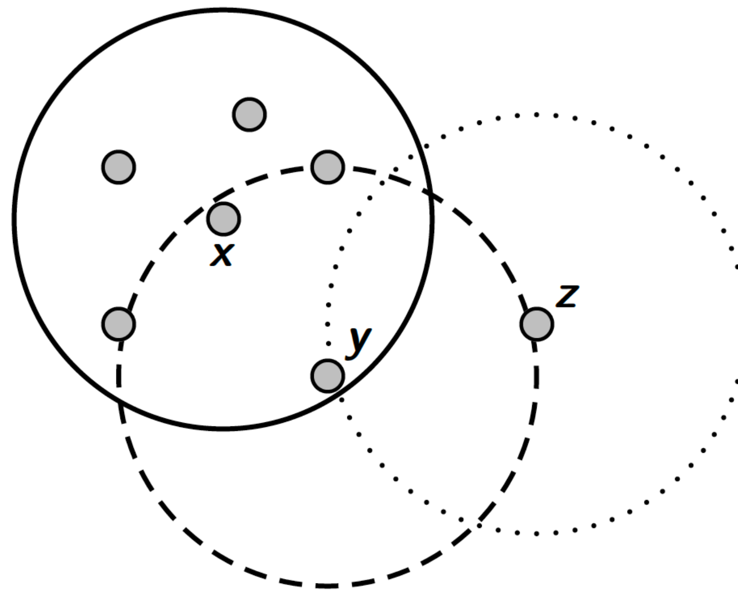
$$N_\epsilon(\mathbf{x}) = B_d(\mathbf{x}, \epsilon) = \{\mathbf{y} \mid \delta(\mathbf{x}, \mathbf{y}) \le \epsilon\}$$

- Here δ(x,y) represents the distance between points x and y. which is usually assumed to be the Euclidean
- We say that x is a *core point* if there are at least *minpts* points in its ε-neighborhood, i.e., if $|N_\varepsilon(x)| \geq$ minpts.
- *A border point* does not meet the minpts threshold, i.e., $|N_\varepsilon(x)| <$ minpts , but it belongs to the ε-neighborhood, of some core point z, that is, $x \in N_\varepsilon(z)$.
- If a point is neither a core nor a border point, then it is called a *noise point* or an outlier.

# Core, Border and Noise Points




(a) Neighborhood of a Point




(b) Core, Border, and Noise Points

# The DBSCAN Approach

- A point x is directly density *reachable* from another point y if $x \in N_\varepsilon(y)$ and y is a core point.
- A point x is density *reachable* from y if there exists a chain of points, $x_0, x_1, \ldots, x_l$ , such that $x = x_0$ and $y = x_l$ , and $x_i$ is directly density reachable from $x_{i-1}$ for all $i = 1, \ldots, l$ . In other words, there is set of core points leading from y to x.
- Two points x and y are *density connected* if there exists a core point z, such that both x and y are density reachable from z.
- A *density-based cluster* is defined as a maximal set of density connected points.

# The DBSCAN Approach

- DBSCAN computes the ε-neighborhood $N_\varepsilon(x_i)$ for each point $x_i$ in the dataset D, and checks if it is a core point. It also sets the cluster id, $id(x_i) = \emptyset$ for all points, indicating that they are not assigned to any cluster.
- Starting from each unassigned core point, the method recursively finds all its density connected points, which are assigned to the same cluster.
- Some border point may be reachable from core points in more than one cluster; they may either be arbitrarily assigned to one of the clusters or to all of them (if overlapping clusters are allowed).
- Those points that do not belong to any cluster are treated as outliers or noise.
- Each DBSCAN cluster is a maximal connected component over the core point graph.
- DBSCAN is sensitive to the choice of ε, in particular if clusters have different densities.

# The DBSCAN Algorithm

DBSCAN in action

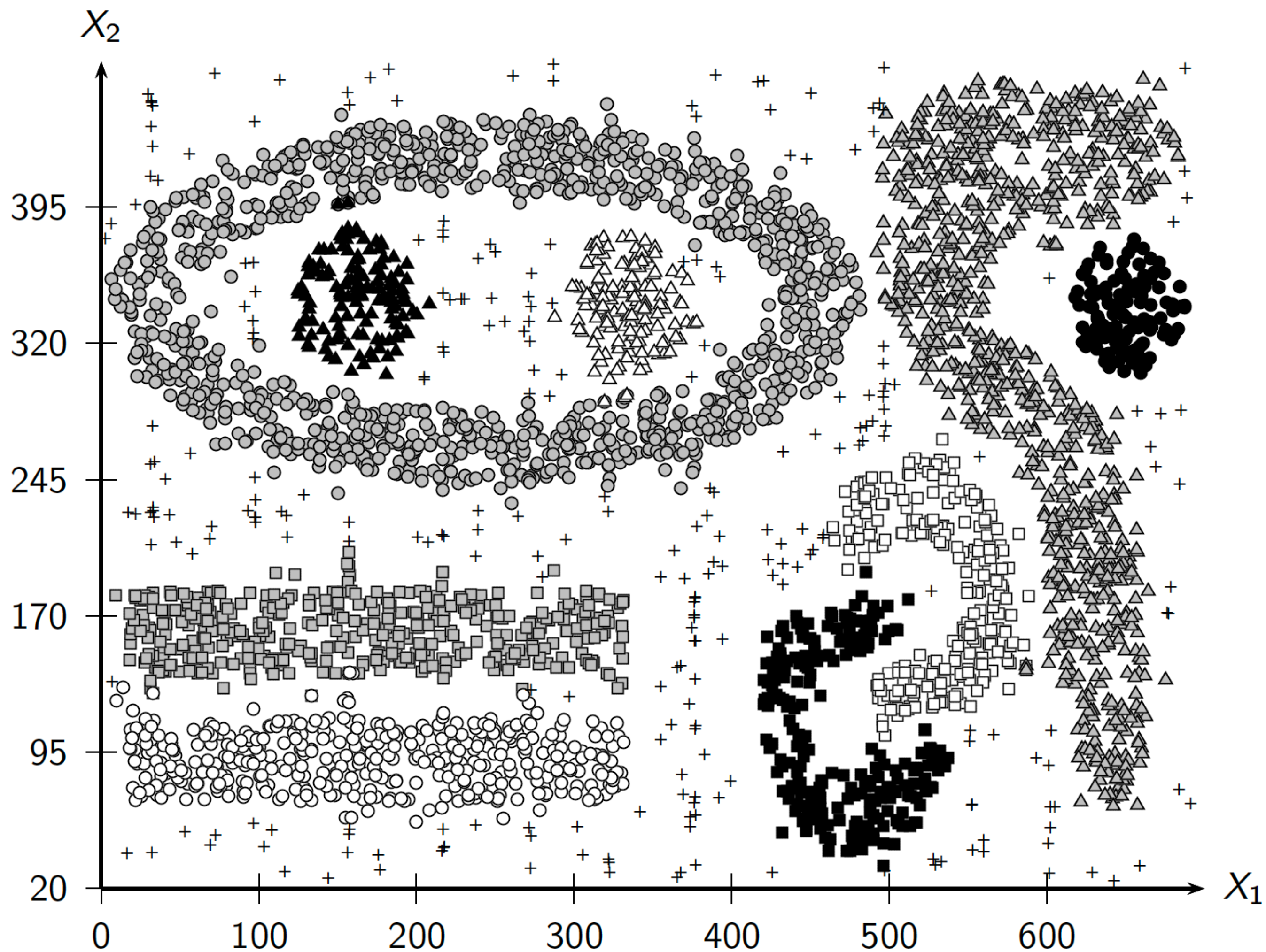
**dbscan** ($\boldsymbol{D}$, $\epsilon$, *minpts*)**:**

$Core \leftarrow \emptyset$<br>
**foreach** $\boldsymbol{x}_i \in \boldsymbol{D}$ **do** `// Find the core points`<br>
    Compute $N_\epsilon(\boldsymbol{x}_i)$<br>
    $id(\boldsymbol{x}_i) \leftarrow \emptyset$ `// cluster id for` $\boldsymbol{x}_i$<br>
    **if** $N_\epsilon(\boldsymbol{x}_i) \geq minpts$ **then** $Core \leftarrow Core \cup \{\boldsymbol{x}_i\}$<br>
$k \leftarrow 0$ `// cluster id`<br>
**foreach** $\boldsymbol{x}_i \in Core$, *such that* $id(\boldsymbol{x}_i) = \emptyset$ **do**<br>
    $k \leftarrow k + 1$<br>
    $id(\boldsymbol{x}_i) \leftarrow k$ `// assign` $\boldsymbol{x}_i$ `to cluster id` $k$<br>
    DensityConnected $(\boldsymbol{x}_i, k)$<br>
$\mathcal{C} \leftarrow \{C_i\}_{i=1}^{k}$, where $C_i \leftarrow \{\boldsymbol{x} \in \boldsymbol{D} \mid id(\boldsymbol{x}) = i\}$<br>
$Noise \leftarrow \{\boldsymbol{x} \in \boldsymbol{D} \mid id(\boldsymbol{x}) = \emptyset\}$<br>
$Border \leftarrow \boldsymbol{D} \setminus \{Core \cup Noise\}$<br>
**return** $\mathcal{C}, Core, Border, Noise$

**DensityConnected** ($\boldsymbol{x}$, $k$)**:**

**foreach** $\boldsymbol{y} \in N_\epsilon(\boldsymbol{x})$ **do**<br>
    $id(\boldsymbol{y}) \leftarrow k$ `// assign` $\boldsymbol{y}$ `to cluster id` $k$<br>
    **if** $\boldsymbol{y} \in Core$ **then** DensityConnected $(\boldsymbol{y}, k)$

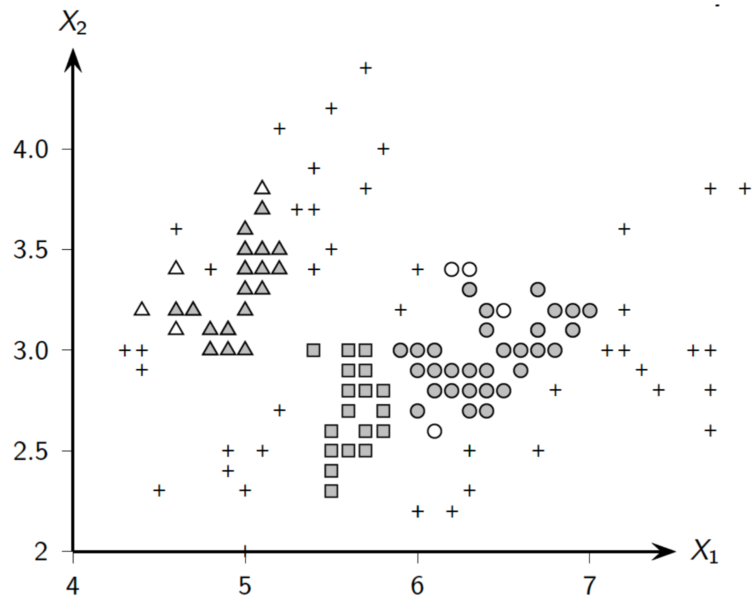
# Density-based Clusters
## ε = 15 and minpts = 10



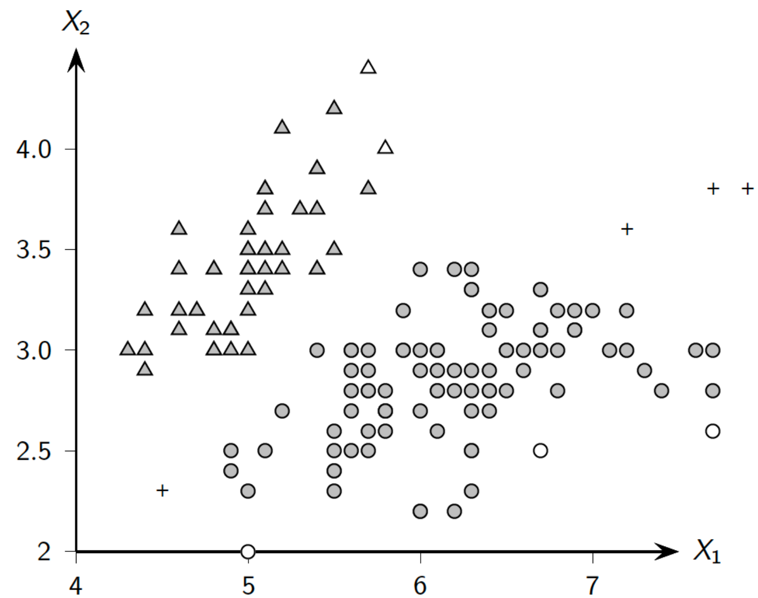
# The Disadvantages of DBSCAN

- Suffers from curse of dimensionality means in high dimensional space the ε-neighborhood is meaningless and all the point are fall close to each other
- Approximate appropriate values for ε and minpt could be a challenging

# DBSCAN Clustering
## Iris dataset




(a) $\epsilon = 0.2,\ minpts = 5$




(b) $\epsilon = 0.36,\ minpts = 3$

# The Disadvantages of DBSCAN

- Suffers from curse of dimensionality means in high dimensional space the ε-neighborhood is meaningless and all the point are fall close to each other
- Approximate appropriate values for ε and minpt could be a challenging
- Finding clusters with different densities could be difficult