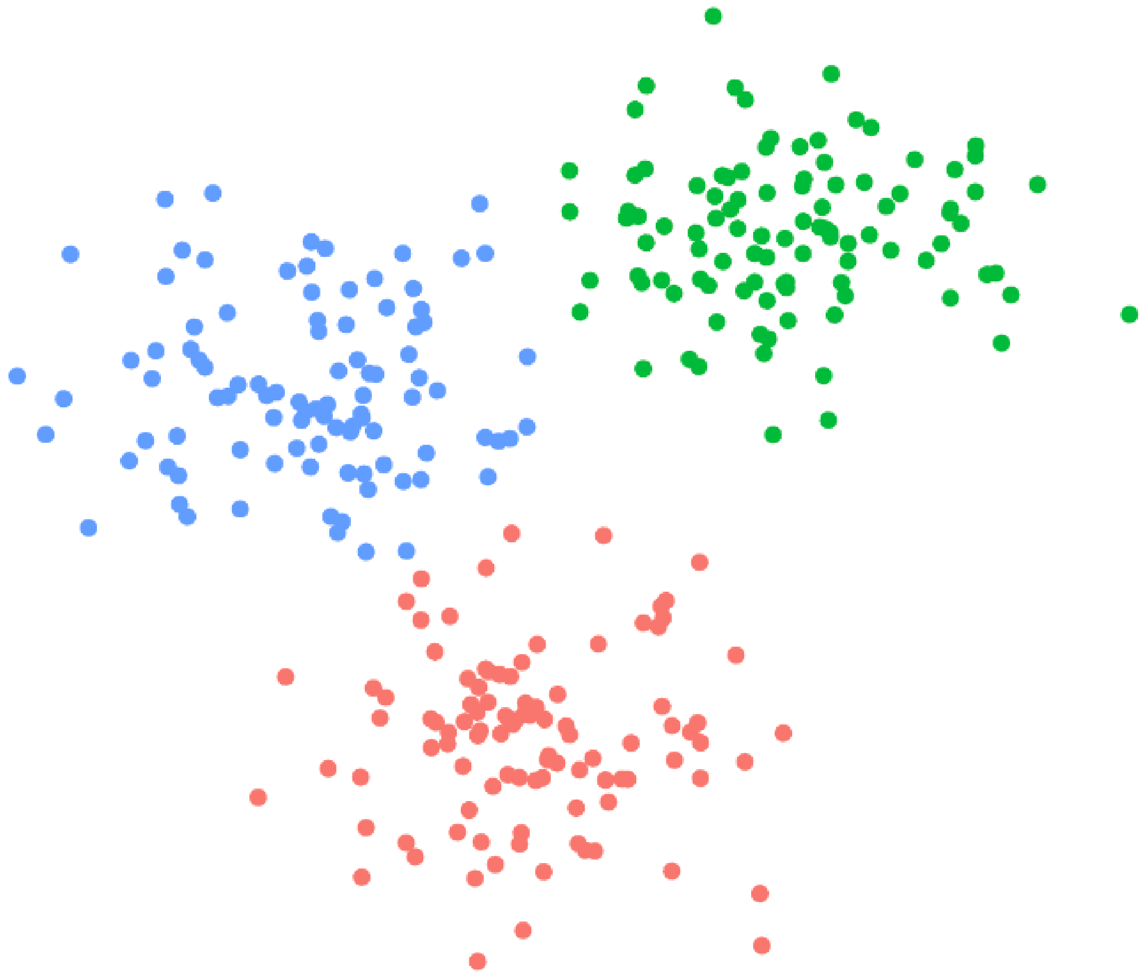
DBSCAN

Density based clustering

DBSCAN

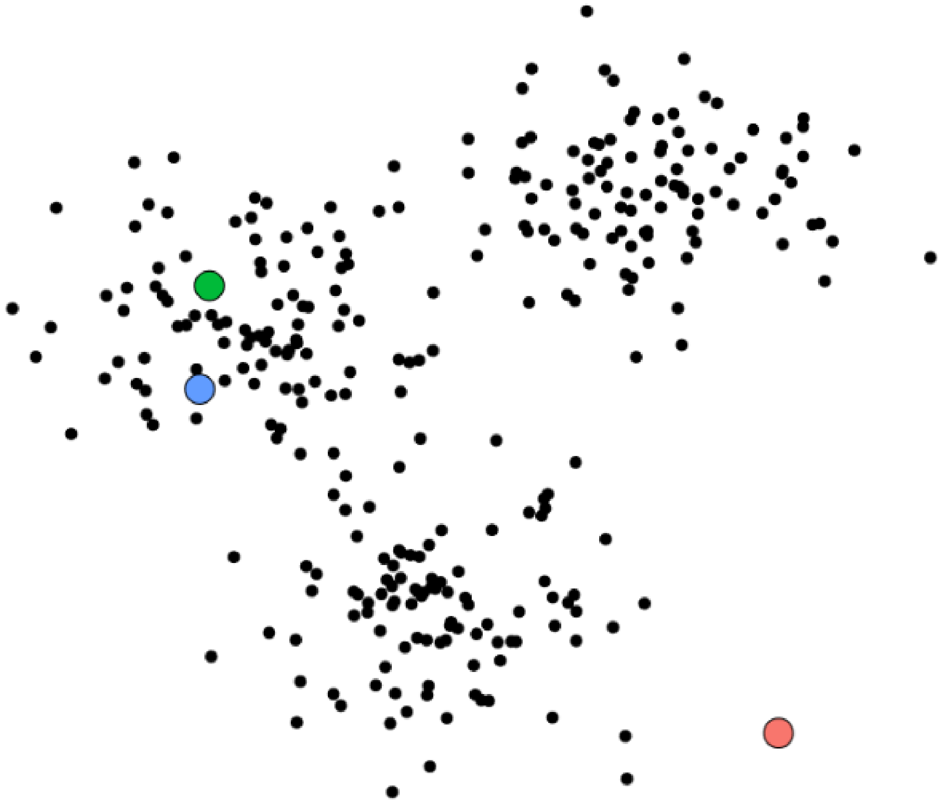
- Density-Based Spatial Clustering of Applications with Noise

Look at this sample labeled dataset



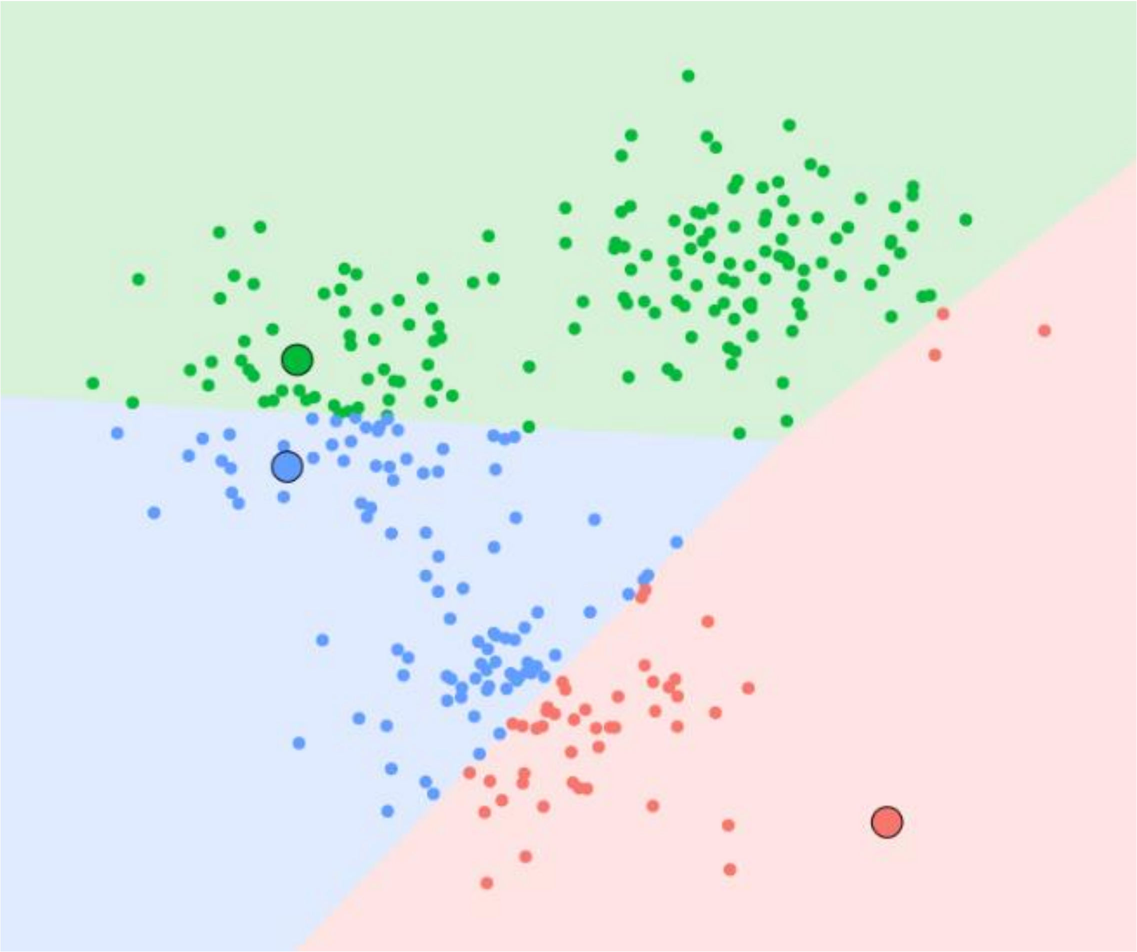
How k-means cluster this data with $k = 3$?

Start with k randomly chosen centroids



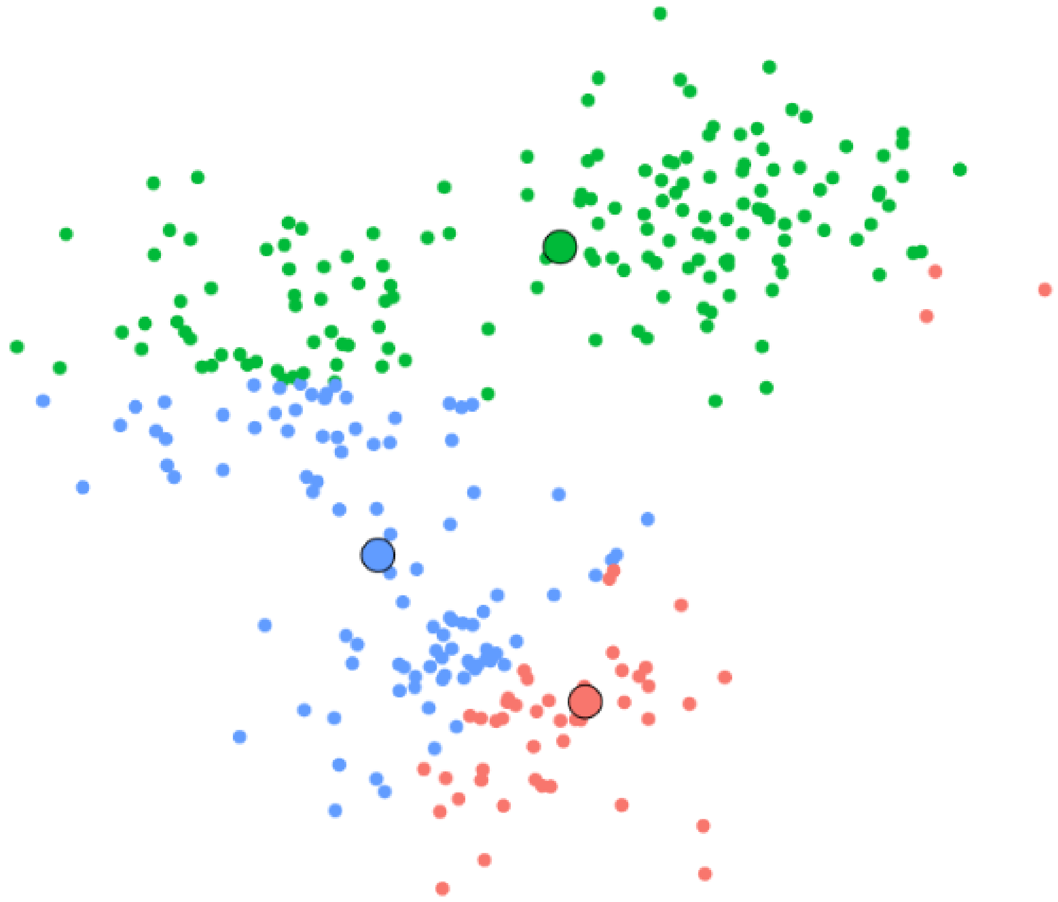
How k-means cluster this data with $k = 3$?

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



How k-means cluster this data with $k = 3$?

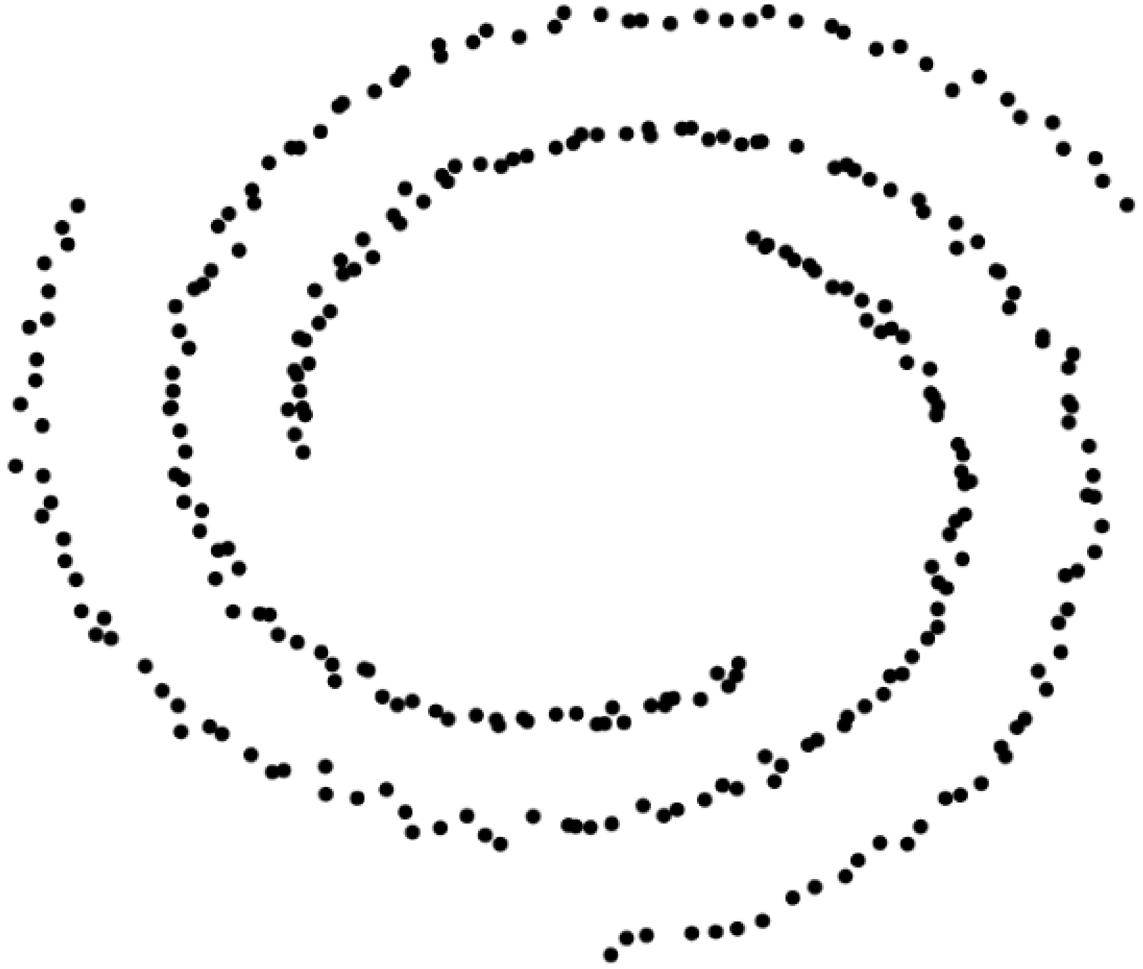
Update the centroids



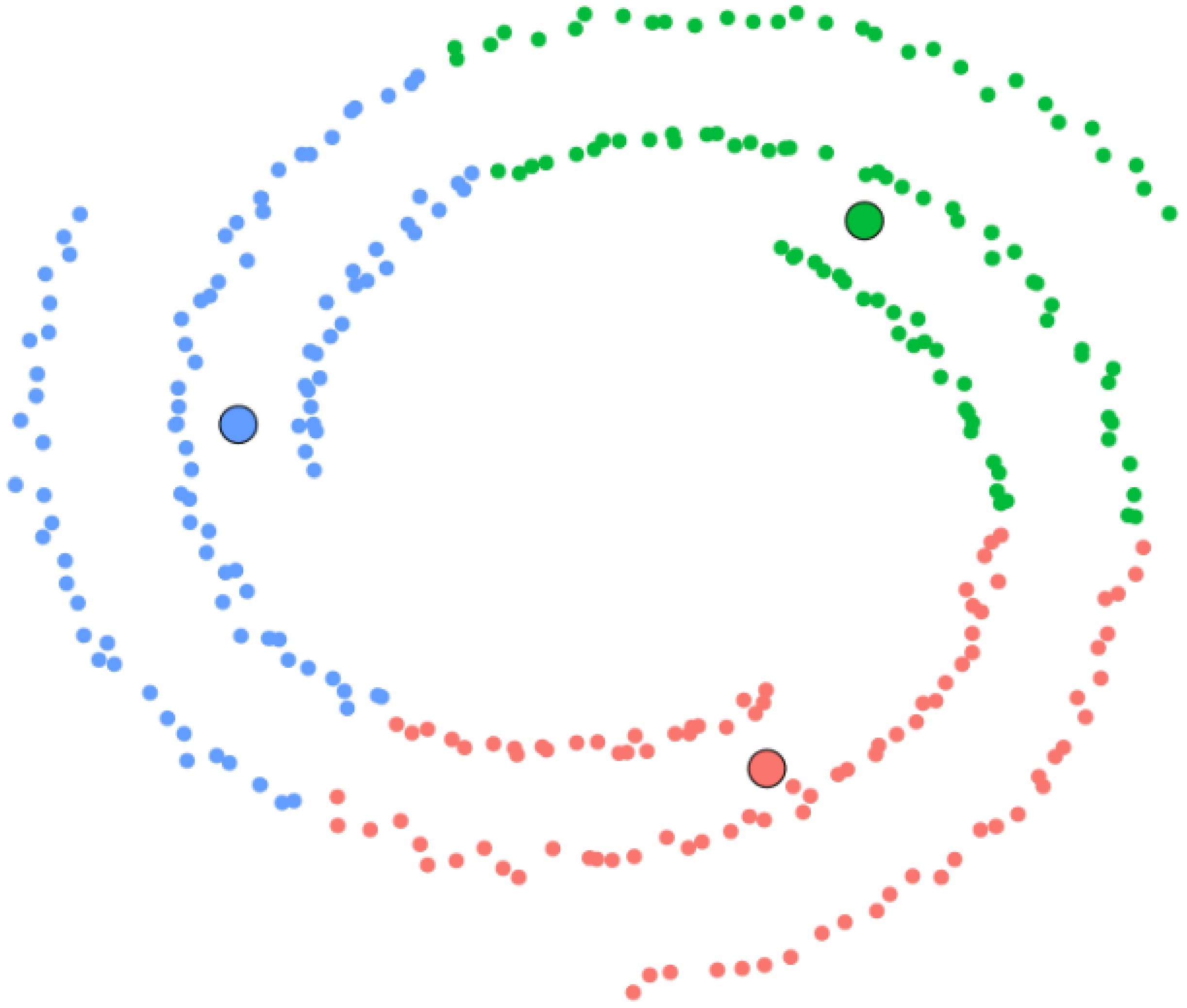
How k-means cluster this data with $k = 5$?

Repeat the steps until convergence

How about this dataset?

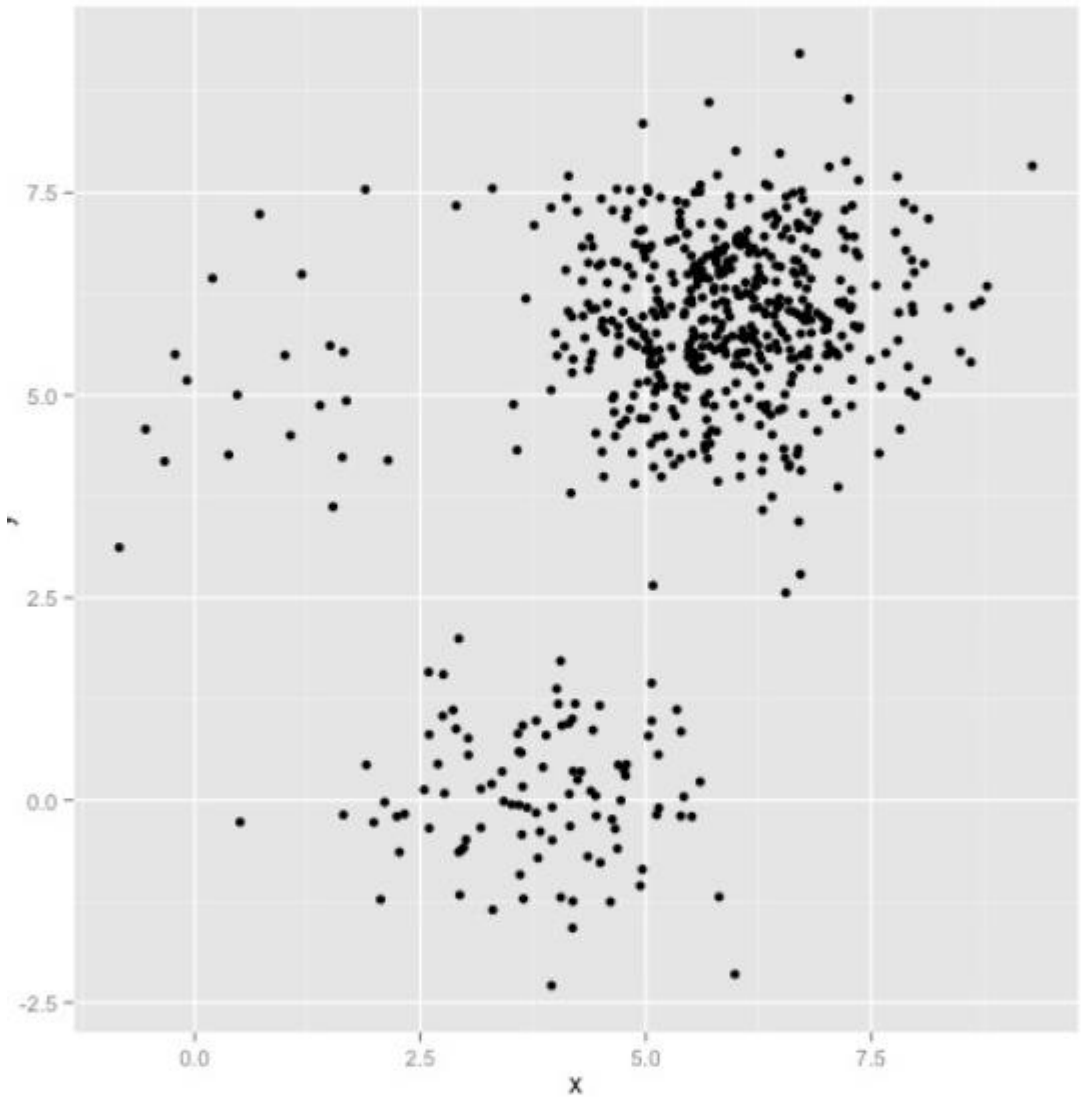


How about this dataset?



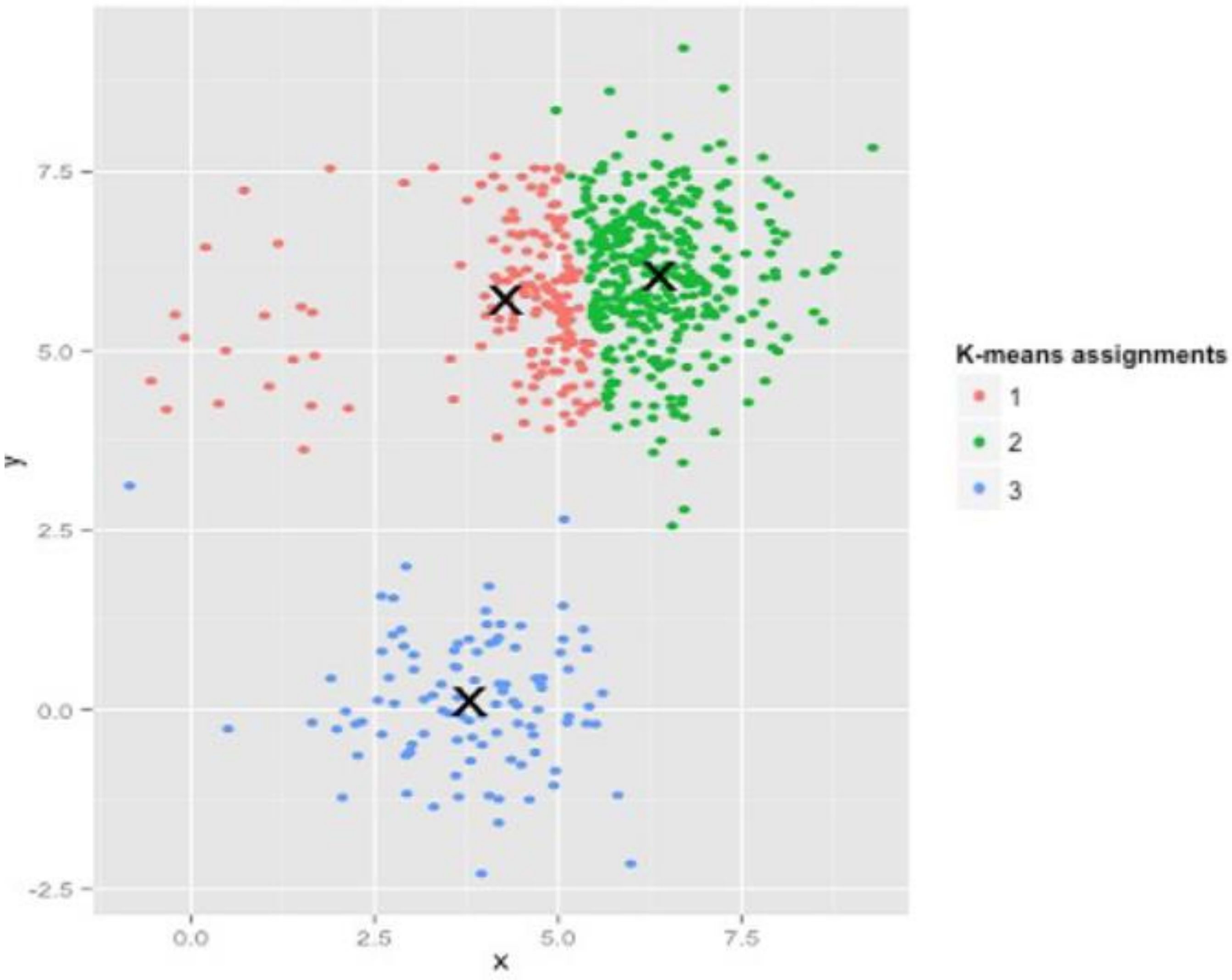
Let’s look at this example

How many clusters can you see in this plot?



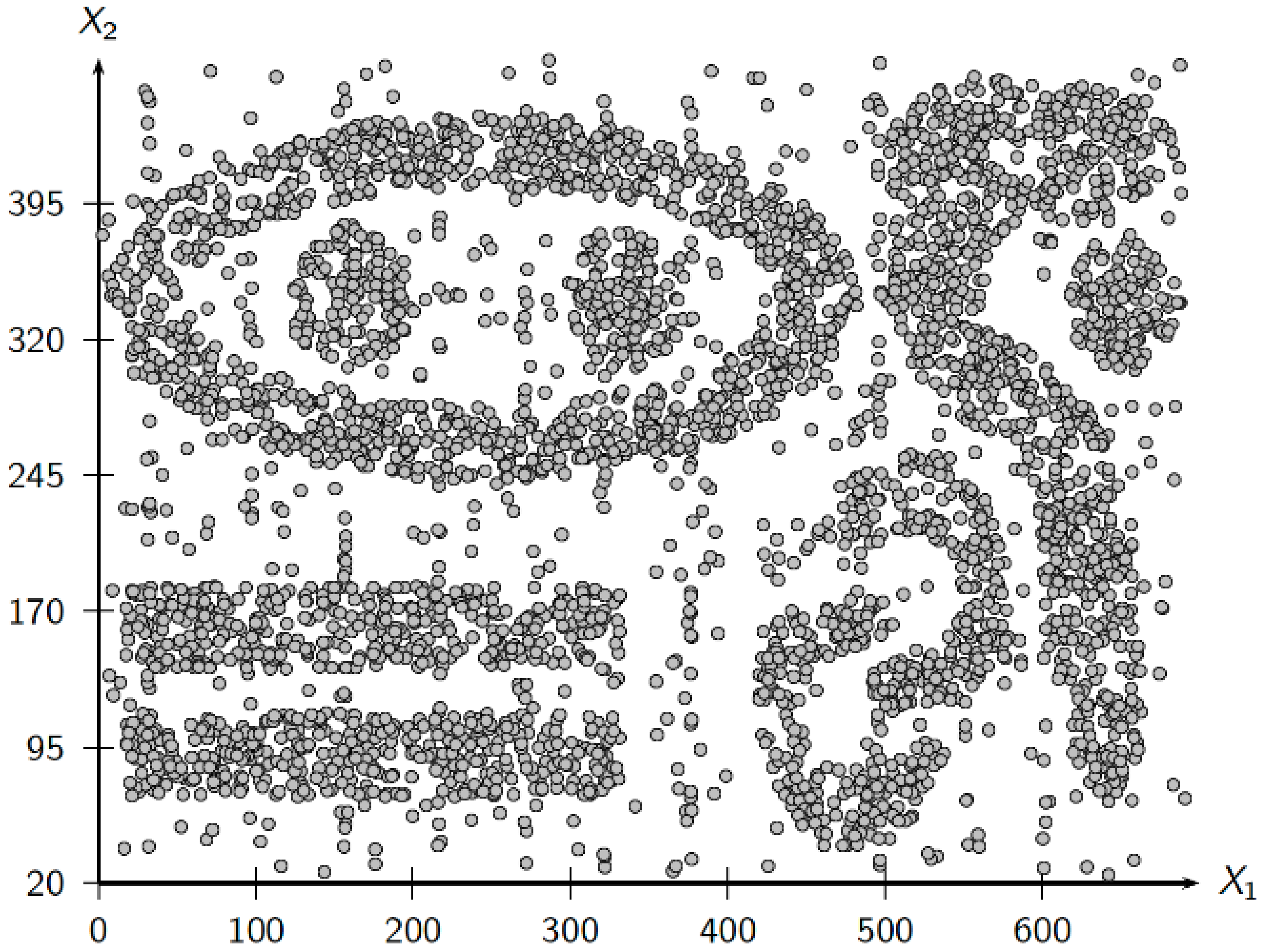
Let's look at this example

K-means results



MONTANA STATE UNIVERSITY

How about this synthetic dataset?



K-means limitations

- It assumes the clusters are of convex shape.
- Sensitive to outliers.
- When clusters are non-convex, two points in two neighborhood clusters might be closer than two points in the same cluster.
- Density based methods are able to mine non-convex clusters, where distance-based methods may have difficulty.
- K-means time complexity $O(tnkd)$

K-means Demo

- https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

DBSCAN approach

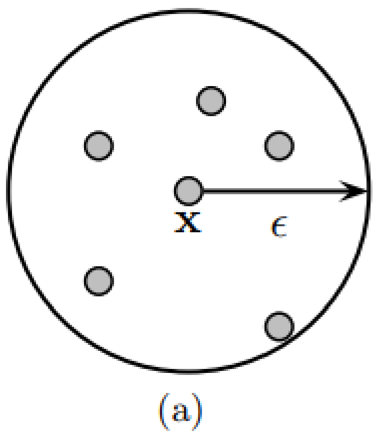
- Density-based Spatial Clustering of Applications with Noise (DBSCAN)
- Define a ball of radius ε around a point $x \in \mathbb{R}^d$, called the ϵ-neighborhood of x:

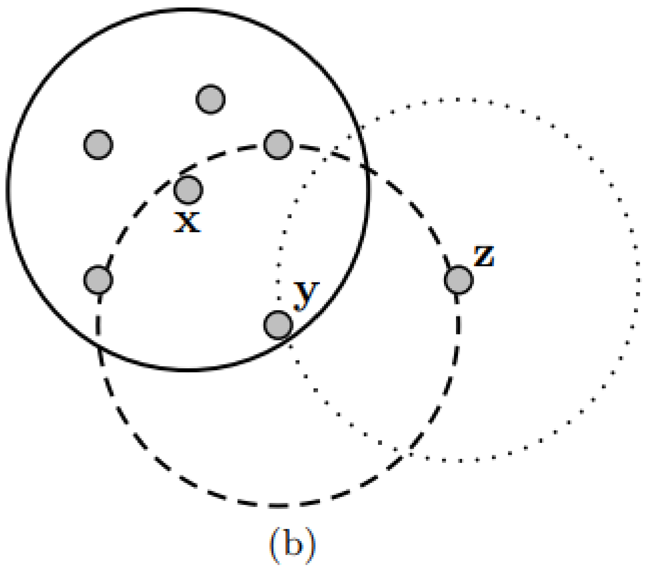
$$\boxed{N_\epsilon(x) = B_d(x, \epsilon) = \{y \mid \delta(x, y) \leq \epsilon\}}$$

- Here $\delta(x, y)$ represents the distance between points x and y, which is usually the Euclidean distance.
    - Other distance metrics can be used as well.
- We say that x is a core point if there are at least minpts points in its ϵ-neighborhood, i.e., if $|N_\epsilon(x)| \geq minpts$.
    - minpts is a user defined local density or frequency threshold .
- A border point does not meet the minpts threshold, i.e., $|N_\epsilon(x)| < minpts$, but it belongs to the ϵ-neighborhood, or core points z, that is, $x \in N_\epsilon(z)$.
- If point is neither core nor border point, then it is called a noise point or an outlier.

Core, border and noise points

Suppose minpts = 6





(a) Neighborhood of a Point

(b) Core, Border, and Noise Point

The DBSCAN approach

- A point x is directly density reachable from another point y if $x \in N_\epsilon(y)$ and y is a core point.
- A point x is density reachable from y if there exists a chain of points $x_0, x_1, x_2, \ldots, x_L$, such that $x = x_0$ and $y = x_L$, and x_i is directly density reachable from x_{i-1} $\forall i \in [1 \ldots L]$. In other words, 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 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 $N_\epsilon(x_i)$ 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 density connected points, which are assigned to the same cluster.
- Some border points 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).
- 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.

MONTANA STATE UNIVERSITY

The DBSCAN algorithm

Algorithm 15.1: Density-based Clustering Algorithm

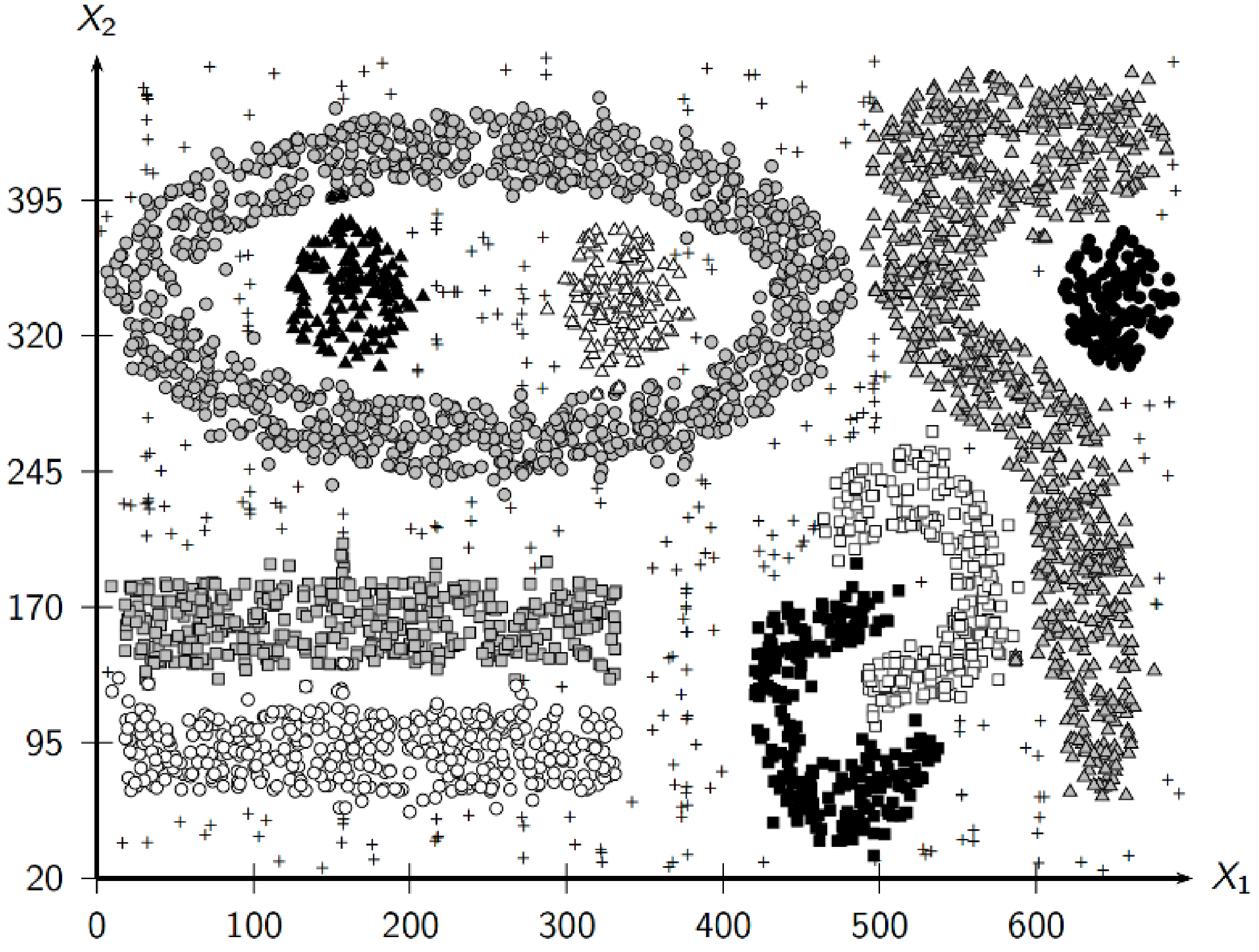
DBSCAN $(\mathbf{D}, \epsilon, minpts)$:

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

DensityConnected $(\mathbf{x}, k)$:

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

Density based clusters $\epsilon = 15$ and $minpts = 10$



MONTANA STATE UNIVERSITY

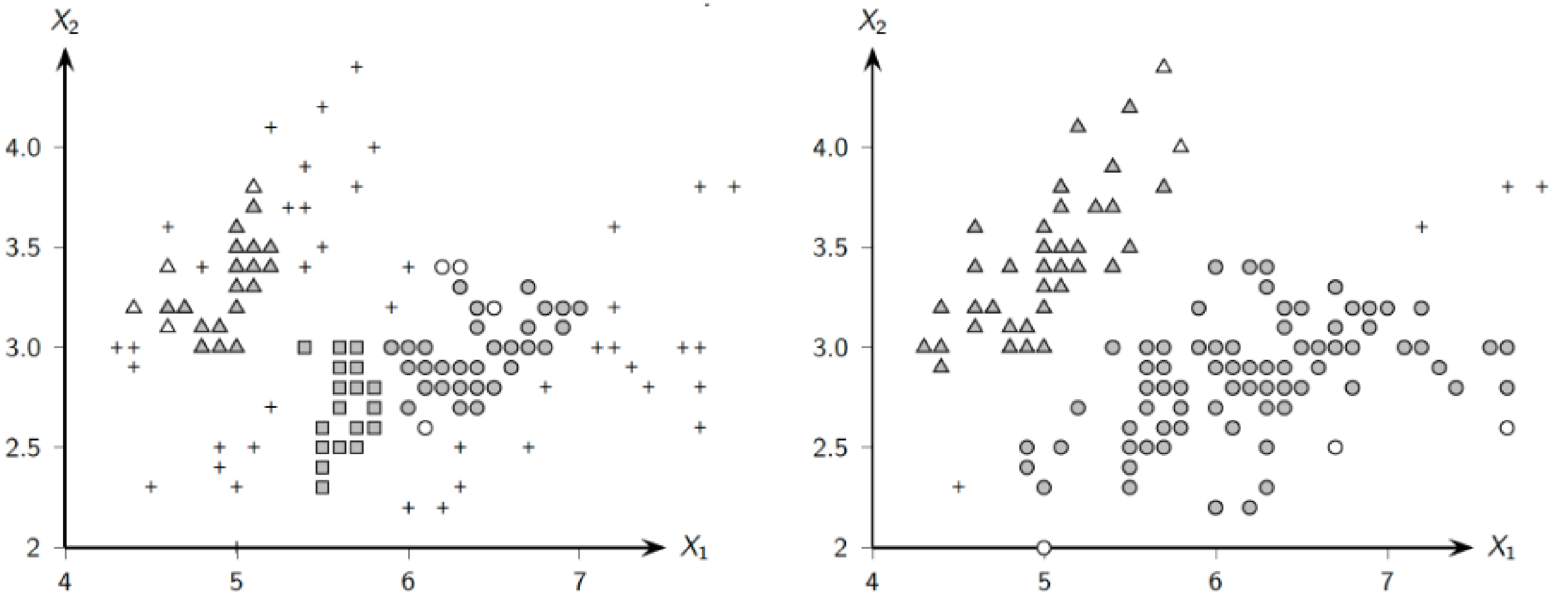
DBSCAN visualization

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

Disadvantages of DBSCAN

- Suffers from curse of dimensionality.
  - In high dimensions ϵ-neighborhood is meaningless
  - All the points fall closer to each other.
- Approximate appropriate values for ϵ and $minpts$ could be challenging.
- Finding clusters with different densities is difficult.

DBSCAN clustering IRIS dataset



(a) $\epsilon = 0.2$, $minpts = 5$ (b) $\epsilon = 0.36$, $minpts = 3$