DBSCAN

DBSCAN

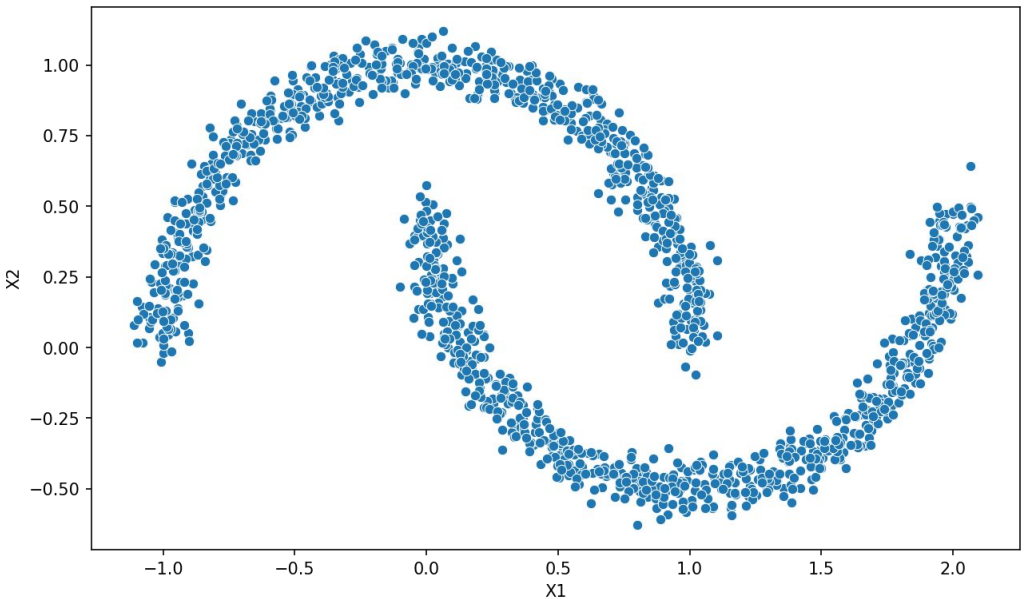
- DBSCAN stands for **D**ensity-**b**ased **s**patial **c**lustering of **a**pplications with **n**oise.
- Let’s review a brief history of the algorithm and then explore an intuition based approach to understanding how it works.

DBSCAN

- DBSCAN Key Ideas
  - DBSCAN focuses on using **density** of points as its main factor for assigning cluster labels.
  - This creates the ability to find cluster segmentations that other algorithms have difficulty with.

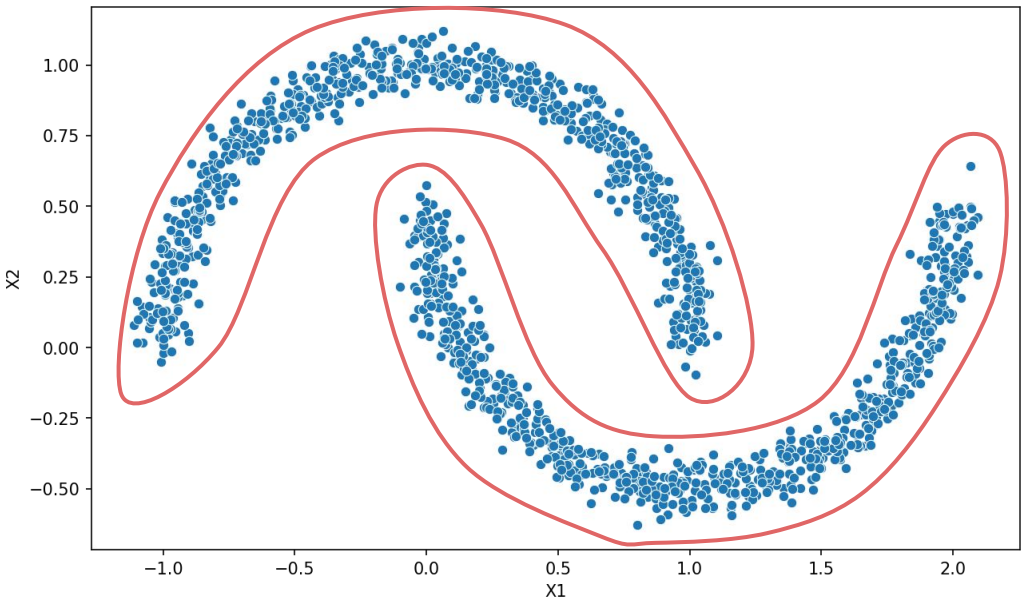
DBSCAN

- Consider the following data set:



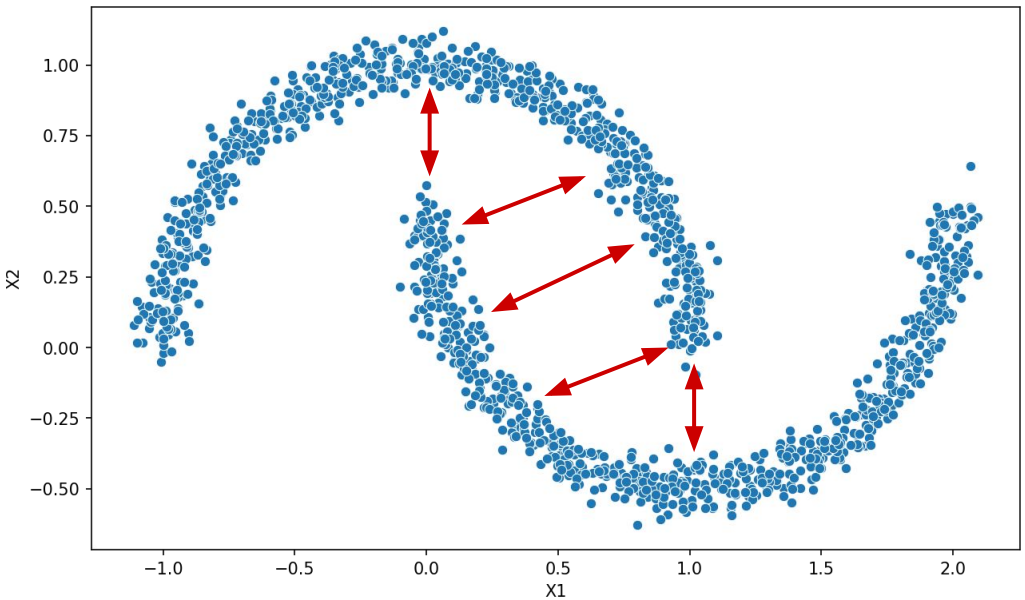
DBSCAN

- Cleary two “moon” shaped clusters:



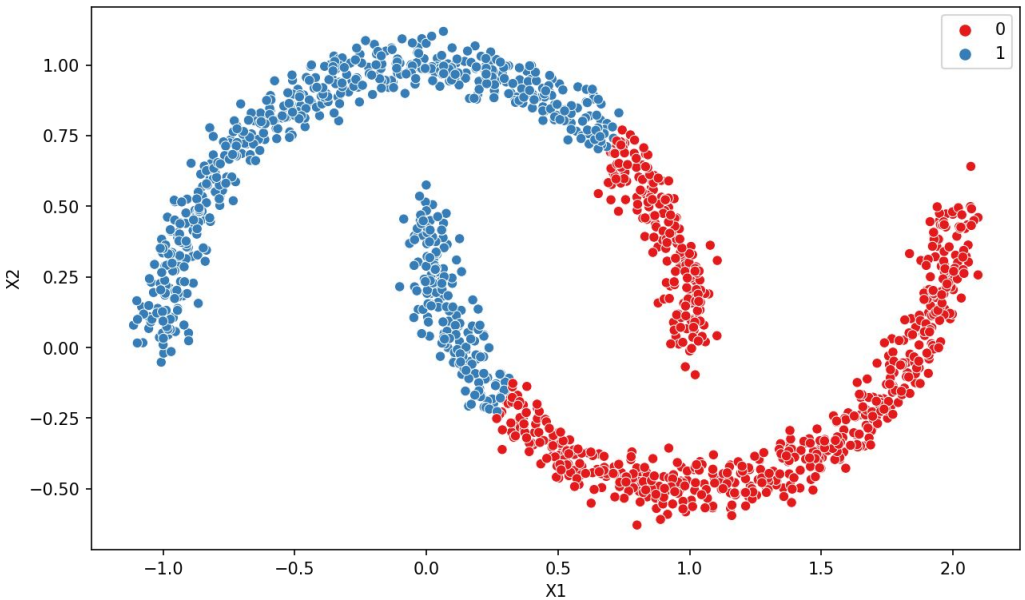
DBSCAN

- But distance based clustering has issues:



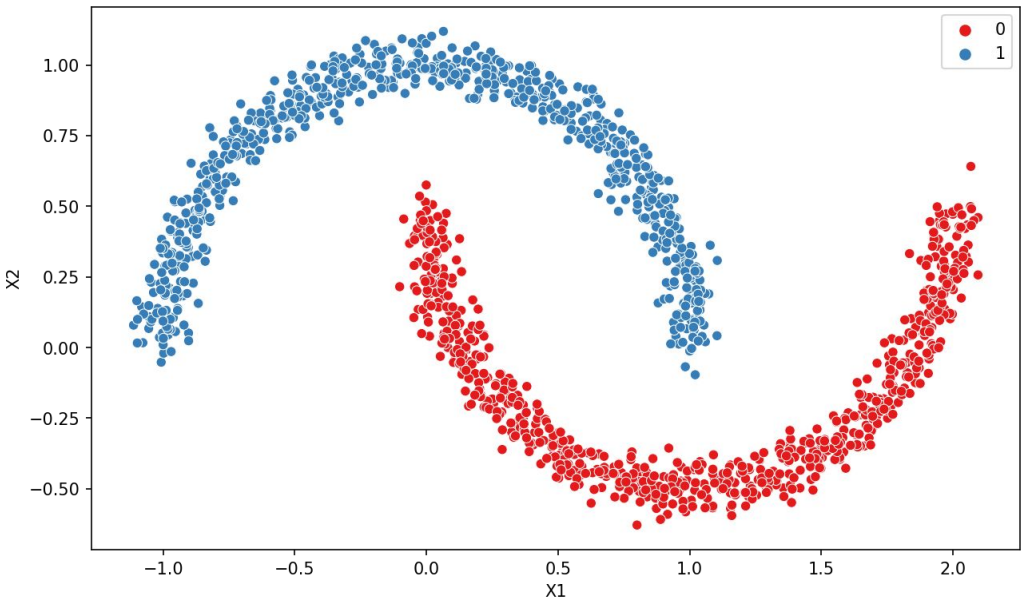
DBSCAN

- Results of K-Means:



DBSCAN

- Results of DBSCAN:



DBSCAN

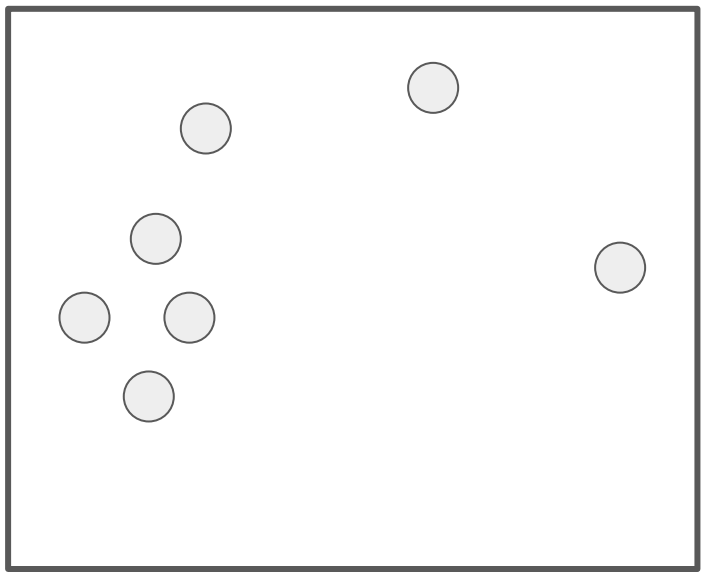
- DBSCAN iterates through points and uses two key hyperparameters (epsilon and minimum number of points) to assign cluster labels.
- Unlike K-Means, it focuses on density as the main factor for cluster assignment of points.

DBSCAN

- DBSCAN Key Hyperparameters:
  - Epsilon:
    - Distance extended from a point.
  - Minimum Number of Points:
    - Minimum number of points in an epsilon distance.

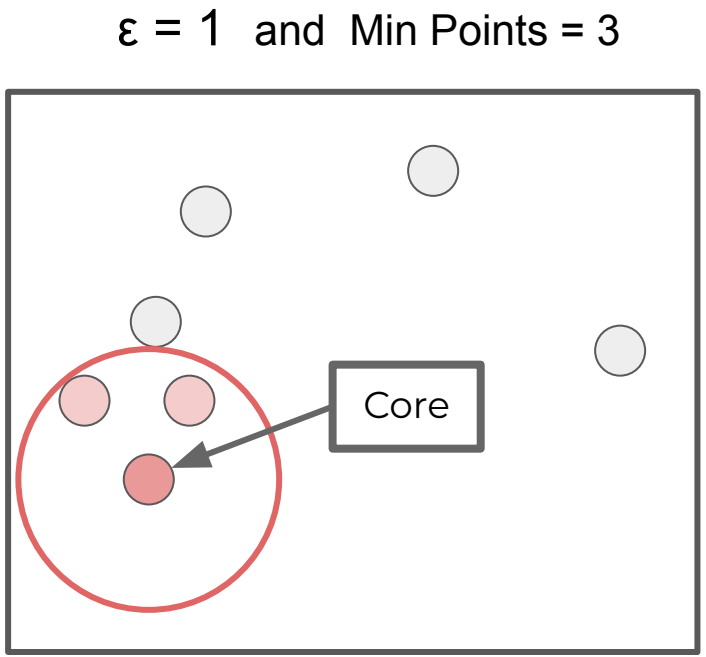
DBSCAN

- DBSCAN Point Types:
  - Core
  - Border
  - Outlier



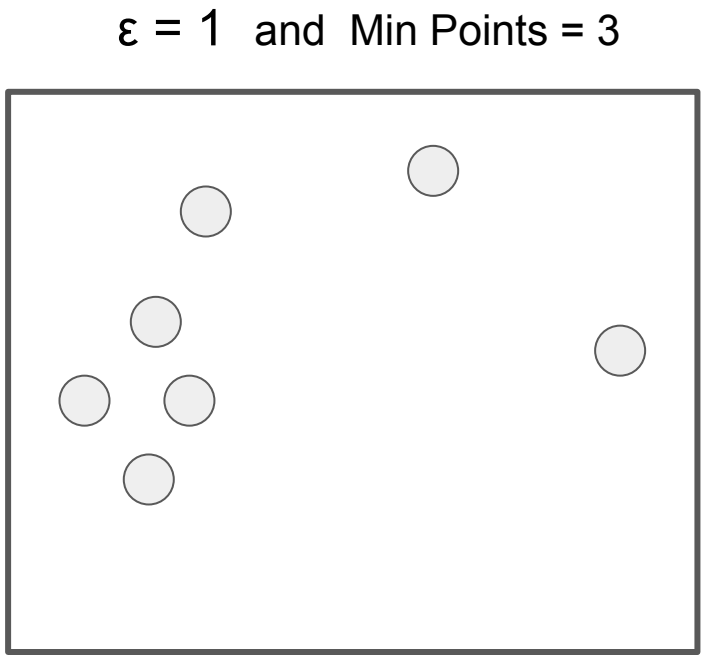
DBSCAN

- DBSCAN Point Types:
  - Core:
    - Point with min. points in epsilon range (including itself).



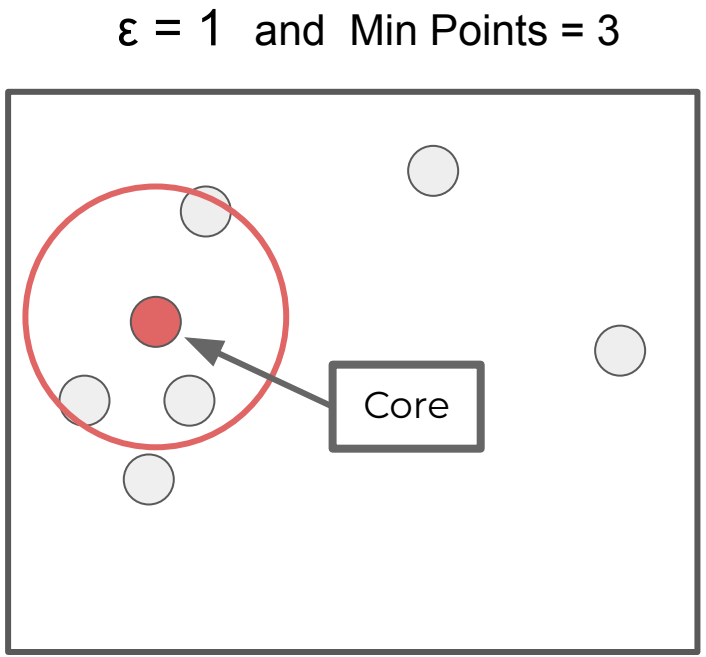
DBSCAN

- DBSCAN Point Types:
  - Border:
    - In epsilon range of core point, but does not contain min. number of points.



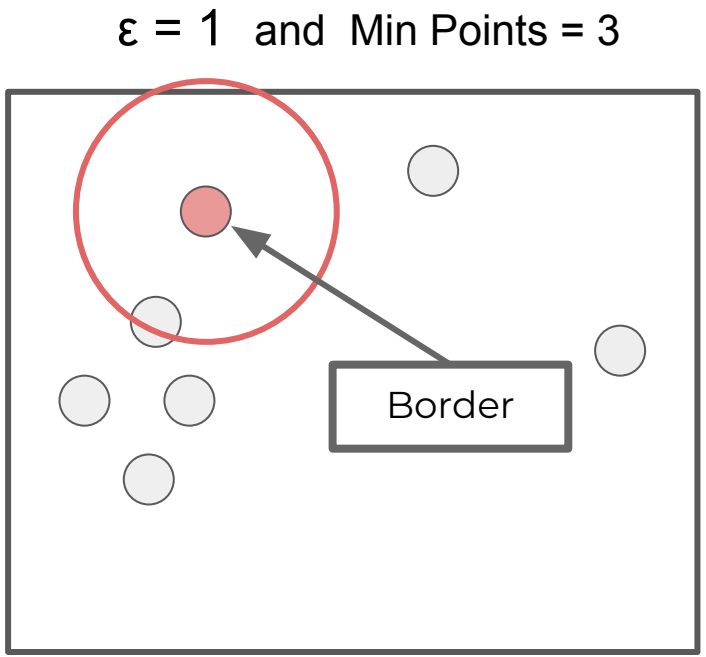
DBSCAN

- DBSCAN Point Types:
  - Border:
    - In epsilon range of core point, but does not contain min. number of points.



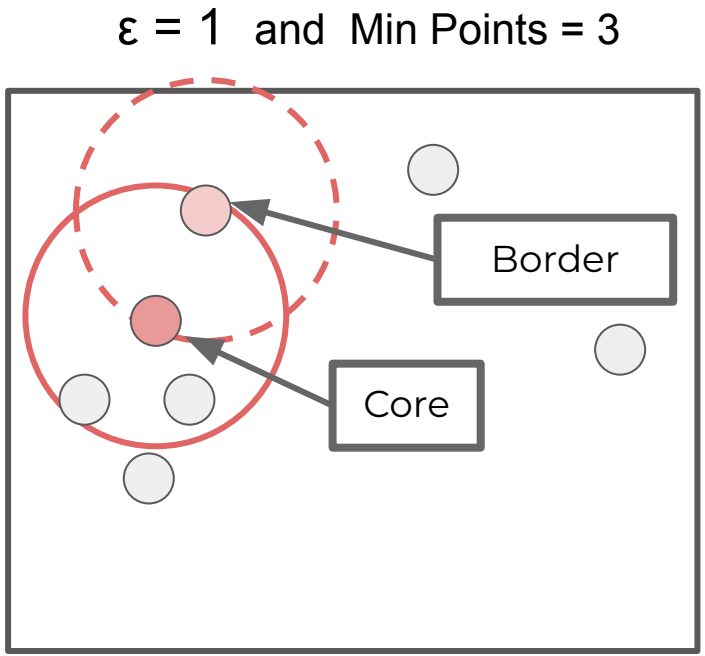
DBSCAN

- DBSCAN Point Types:
  - Border:
    - In epsilon range of core point, but does not contain min. number of points.



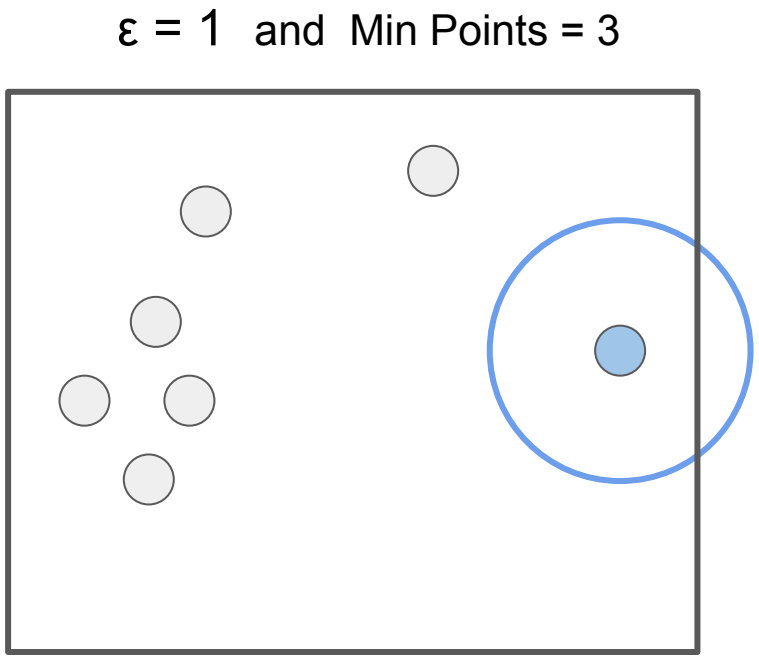
DBSCAN

- DBSCAN Point Types:
  - Border:
    - In epsilon range of core point, but does not contain min. number of points.



DBSCAN

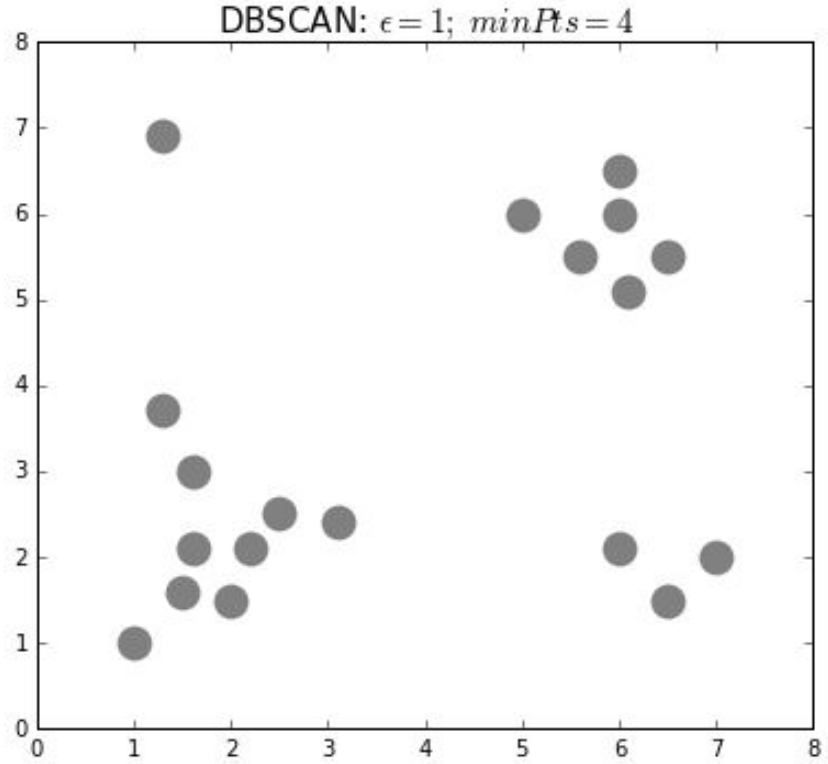
- DBSCAN Point Types:
  - Outlier:
    - Can not be “reached” by points in a cluster assignment.



DBSCAN

- DBSCAN Procedure:
  - Pick a random point not yet assigned.
  - Determine the point type.
  - Once a **core** point has been found, add all directly reachable points to the same cluster as core.
  - Repeat until all points have been assigned to a cluster or as an outlier.

DBSCAN



DBSCAN

- Epsilon Intuition:
  - Increasing epsilon allows more points to be **core** points which also results in more **border** points and less outlier points.
  - Imagine a huge epsilon, all points would be within the neighborhood and classified as the same cluster!

DBSCAN

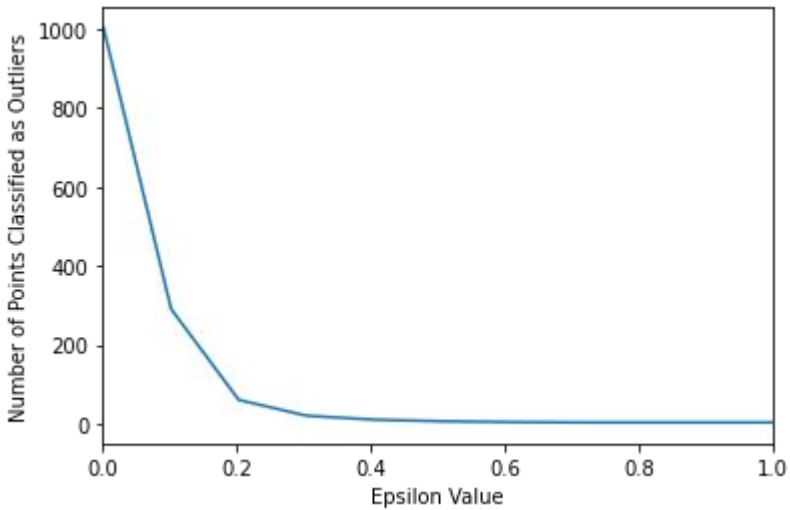
- Epsilon Intuition:
  - Decreasing epsilon causes more points not to be in range of each other, creating more unique clusters.
  - Imagine a tiny epsilon, the range would not extend far out enough to come into contact with any other points!

DBSCAN

- Methods for finding an epsilon value:
  - Run multiple DBSCAN models varying epsilon and measure:
    - Number of Clusters
    - Number of Outliers
    - Percentage of Outliers

DBSCAN

- Plot “elbow/knee” diagram comparing epsilon values:



DBSCAN

- Min. Number of Samples Intuition:
  - Larger number causes more points to be considered unique outliers.
  - Imagine if min. number of samples was close to total number of points available, then very likely all points would become outliers.

DBSCAN

- Choosing Min. Number of Samples:
  - Test multiple potential values and chart against number of outliers labeled.