Data Mining
Cluster Analysis: Basic Concepts and Algorithms

Lecture Notes for Chapter 7

Introduction to Data Mining, 2nd Edition

by

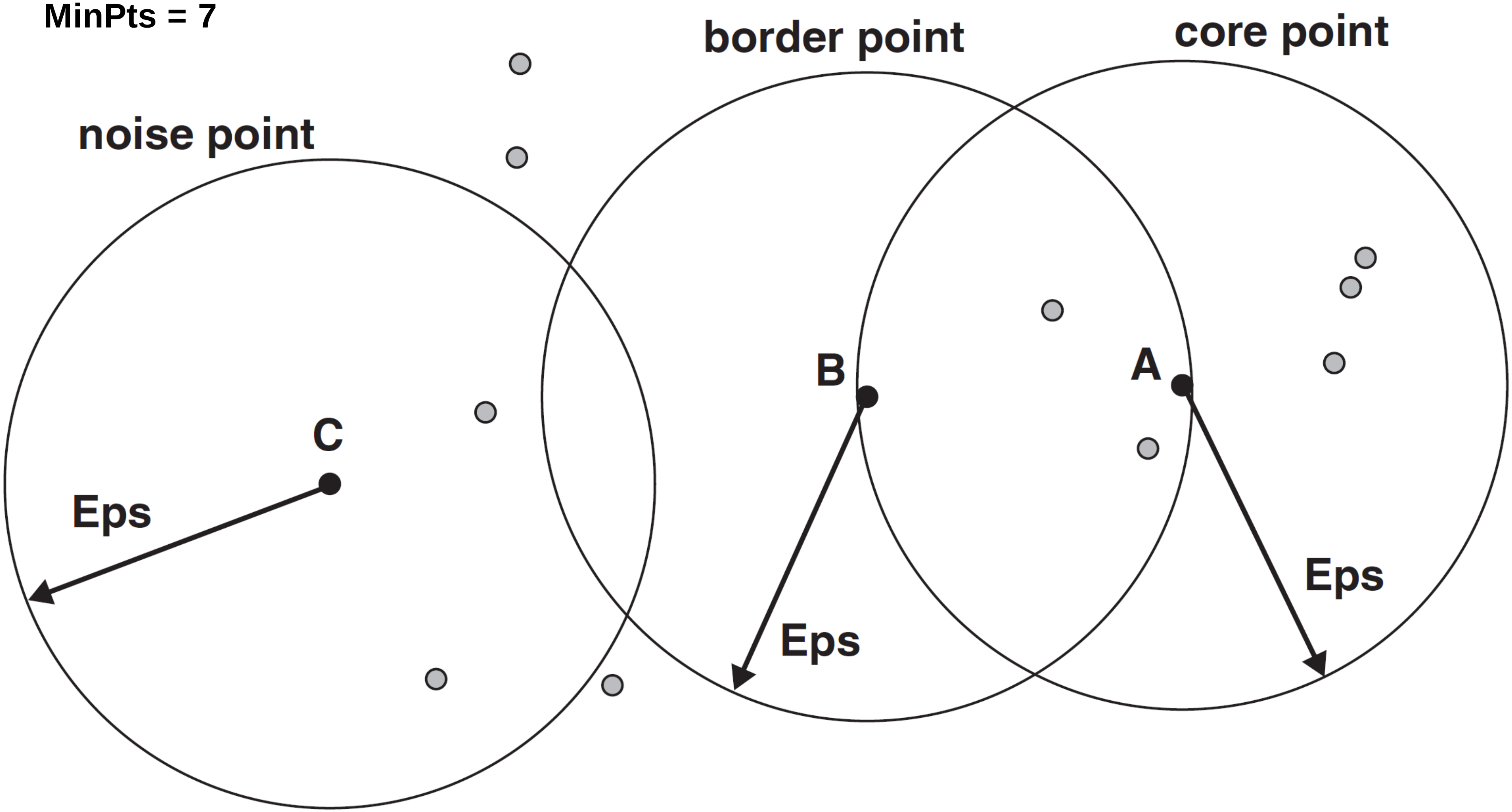
Tan, Steinbach, Karpatne, Kumar

DBSCAN

- DBSCAN is a density-based algorithm.
  - Density = number of points within a specified radius (Eps)
  - A point is a core point if it has at least a specified number of points (MinPts) within Eps
    - These are points that are at the interior of a cluster
    - Counts the point itself
  - A border point is not a core point, but is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point

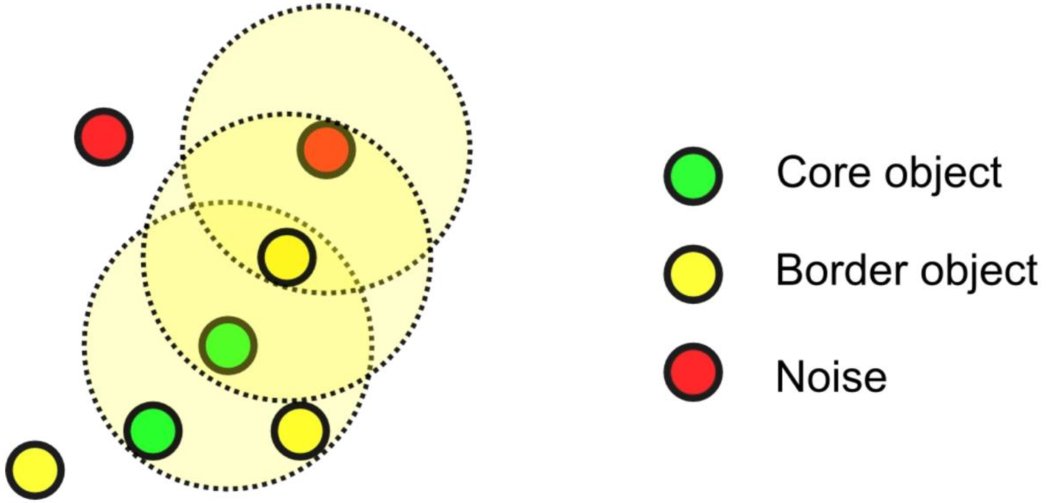
DBSCAN: Core, Border, and Noise Points

MinPts = 7



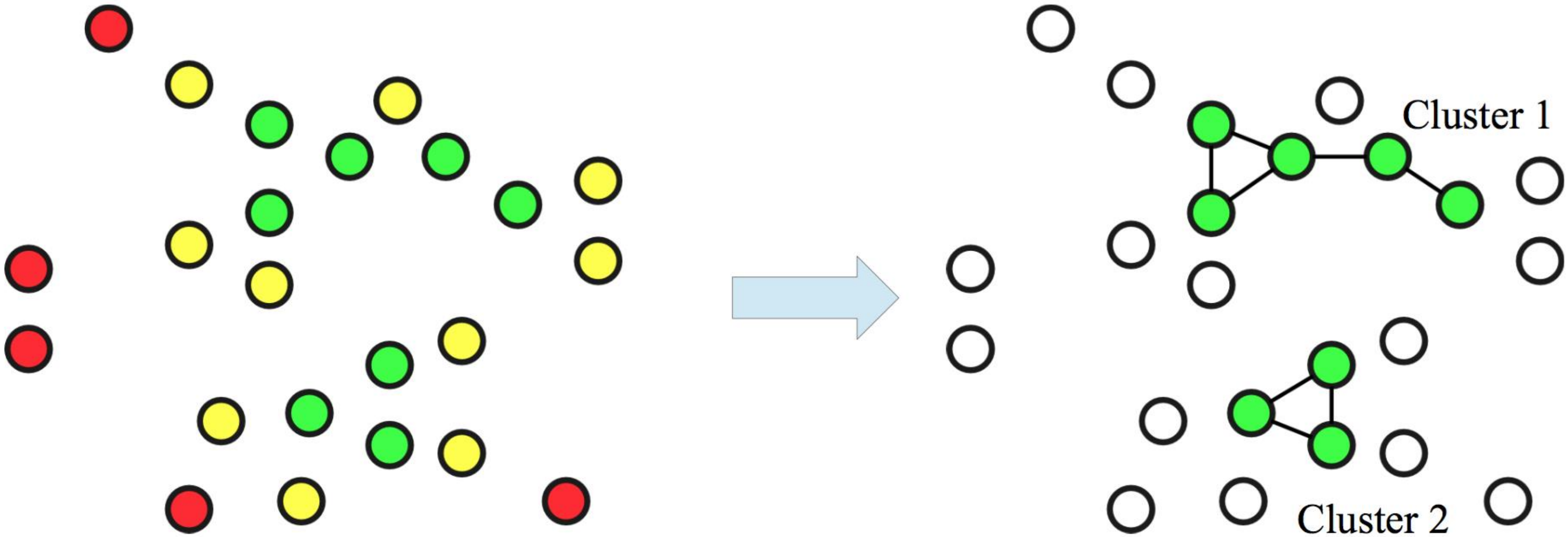
DBSCAN: Step 1

- Label points as **core** (dense), **border** and **noise**
  - Based on thresholds R (radius of neighborhood) and min_pts (min number of neighbors)



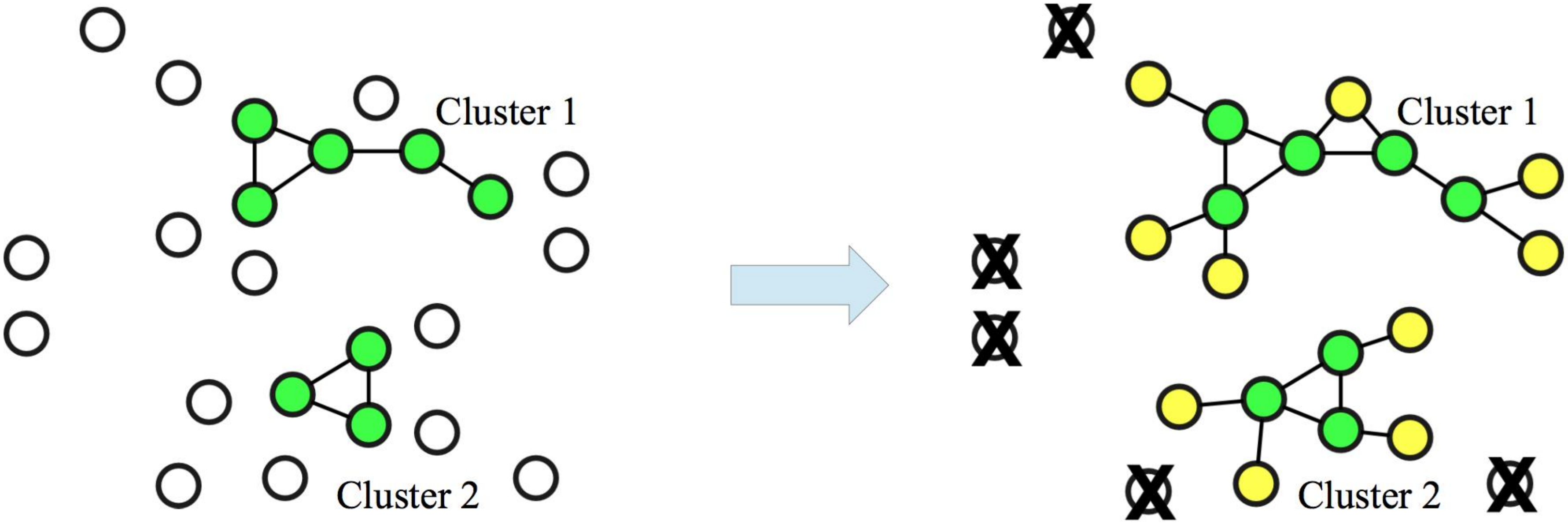
DBSCAN: Step 2

- Connect core objects that are neighbors, and put them in the same cluster



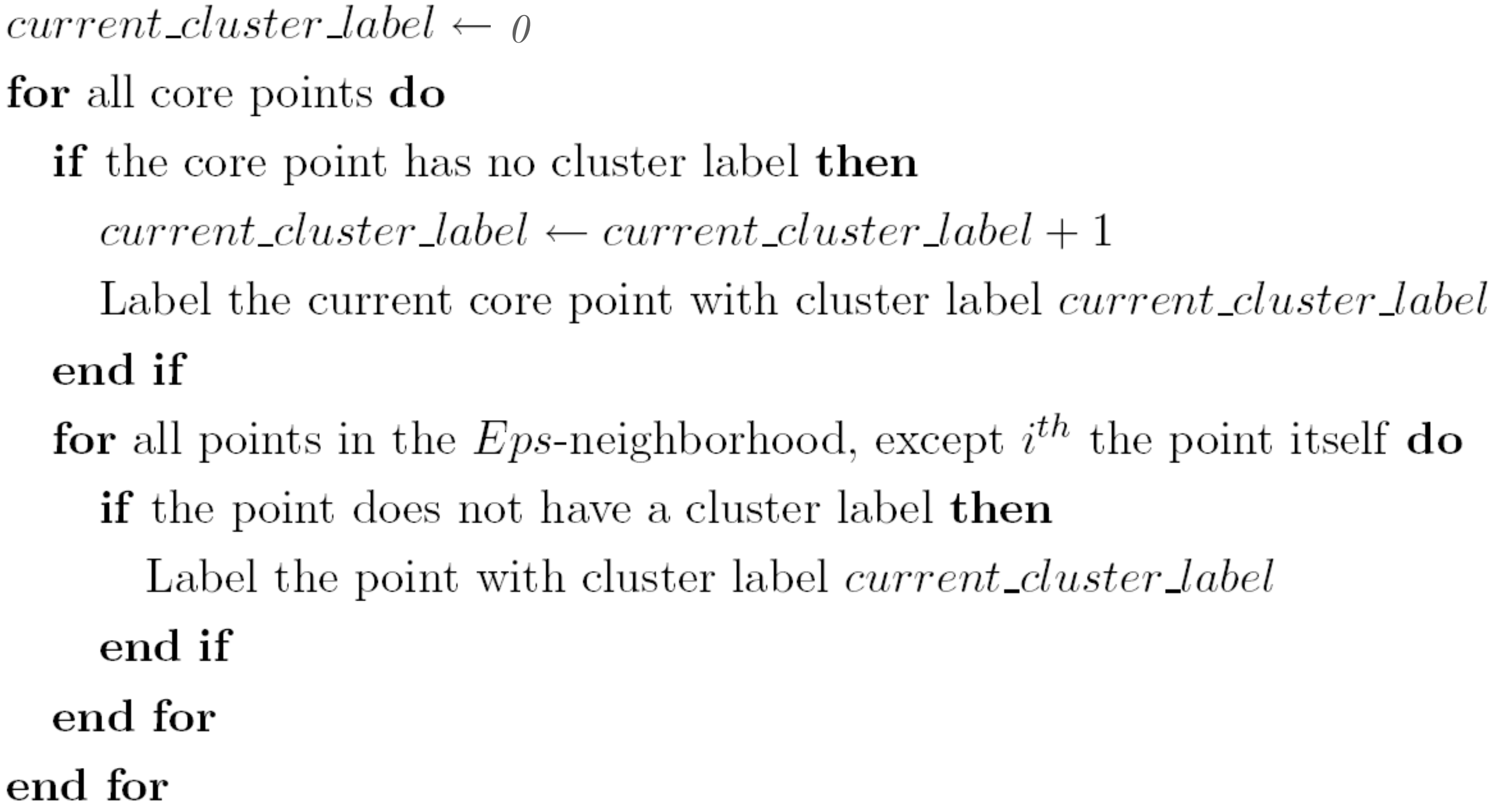
DBSCAN: Step 3

- Associate border objects to (one of) their core(s), and remove noise



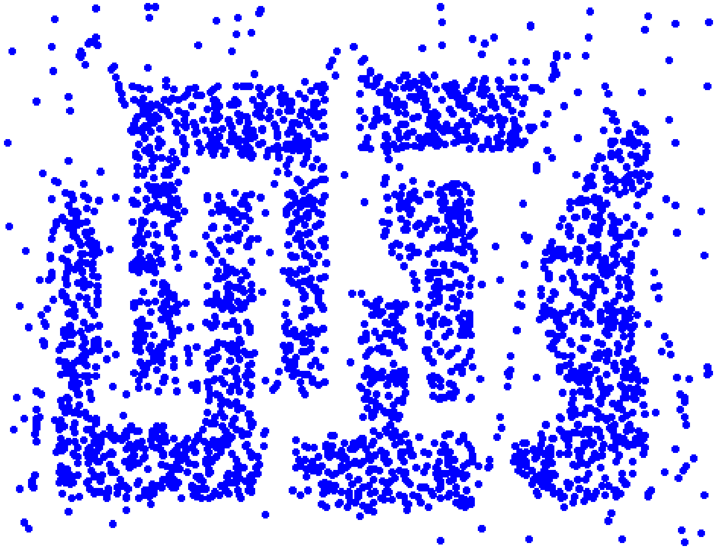
DBSCAN Algorithm

- Eliminate noise points
- Perform clustering on the remaining points

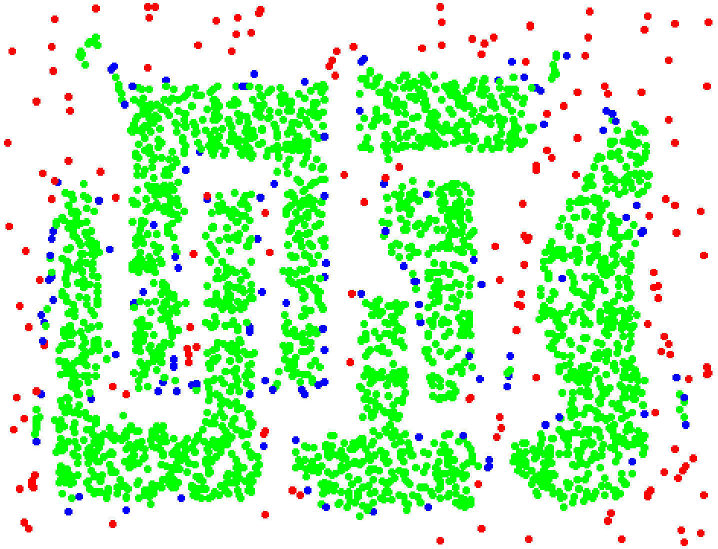


```
$current\_cluster\_label \leftarrow 0$
for all core points do
  if the core point has no cluster label then
    $current\_cluster\_label \leftarrow current\_cluster\_label + 1$
    Label the current core point with cluster label $current\_cluster\_label$
  end if
  for all points in the $Eps$-neighborhood, except $i^{th}$ the point itself do
    if the point does not have a cluster label then
      Label the point with cluster label $current\_cluster\_label$
    end if
  end for
end for
```

DBSCAN: Core, Border and Noise Points



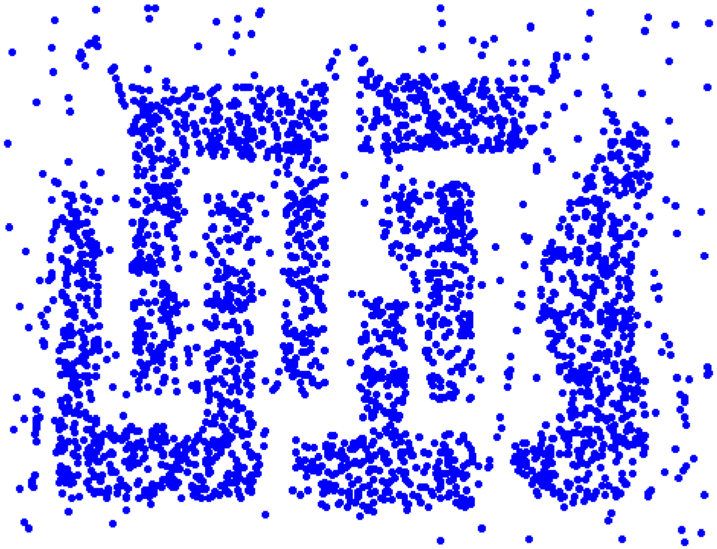
Original Points



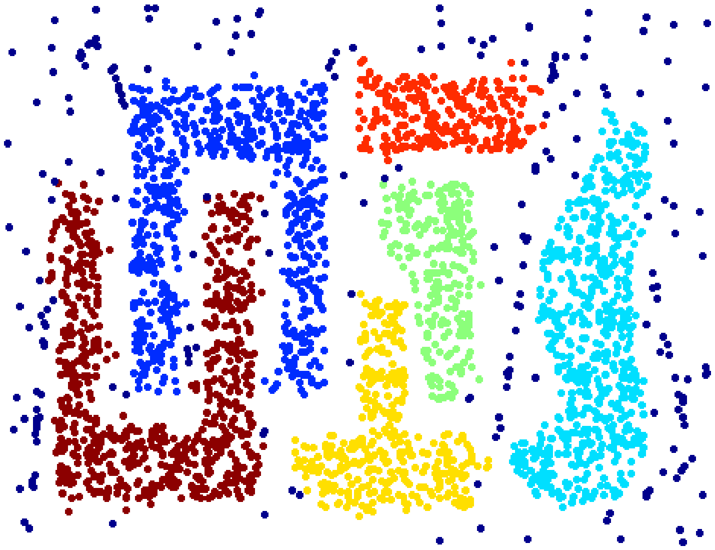
Point types: core, border and noise

Eps = 10, MinPts = 4

When DBSCAN Works Well



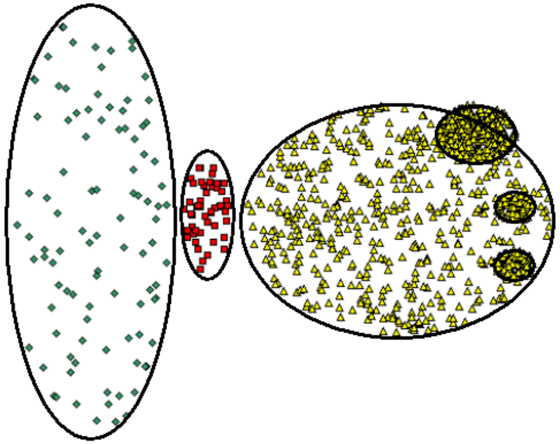
Original Points



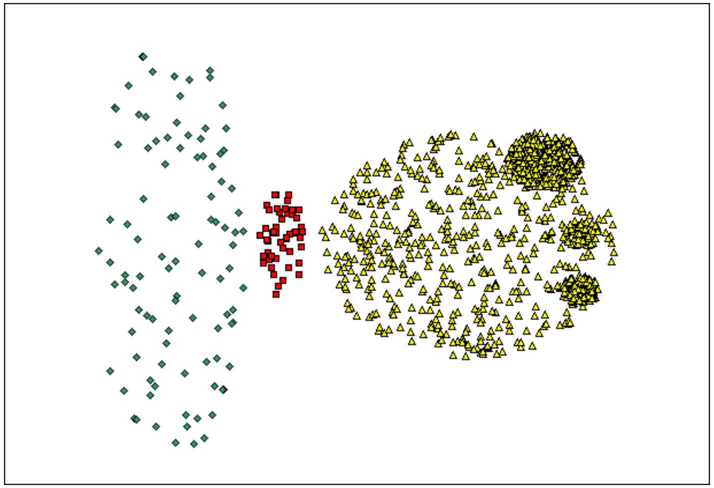
Clusters

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

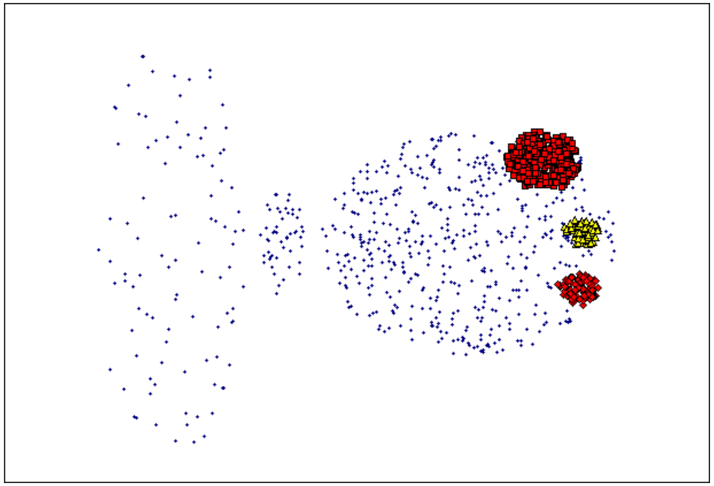
When DBSCAN Does NOT Work Well



Original Points



(MinPts=4, Eps=9.92)

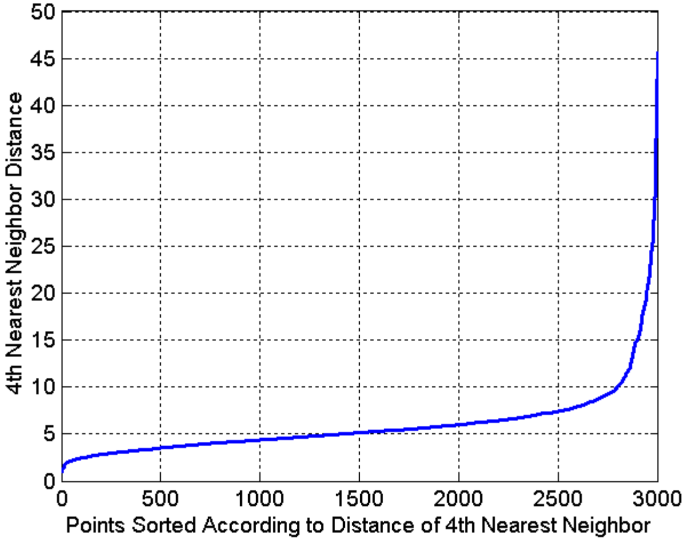


(MinPts=4, Eps=9.75).

- Varying densities
- High-dimensional data where density it is harder to define

DBSCAN: Determining EPS and MinPts

- Idea is that for points in a cluster, their k^th nearest neighbors are at roughly the same distance
- **Noise points have the k^th nearest neighbor at farther distance**
- So, plot sorted distance of every point to its k^th nearest neighbor


4th Nearest Neighbor Distance
Points Sorted According to Distance of 4th Nearest Neighbor