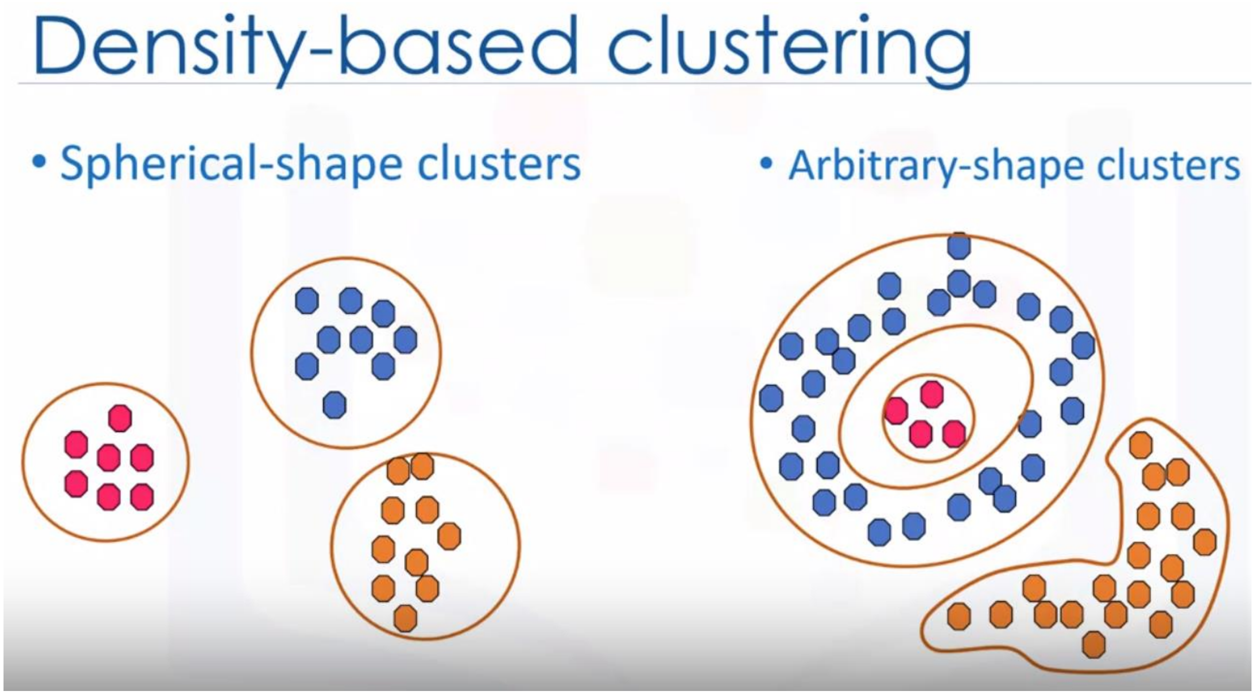
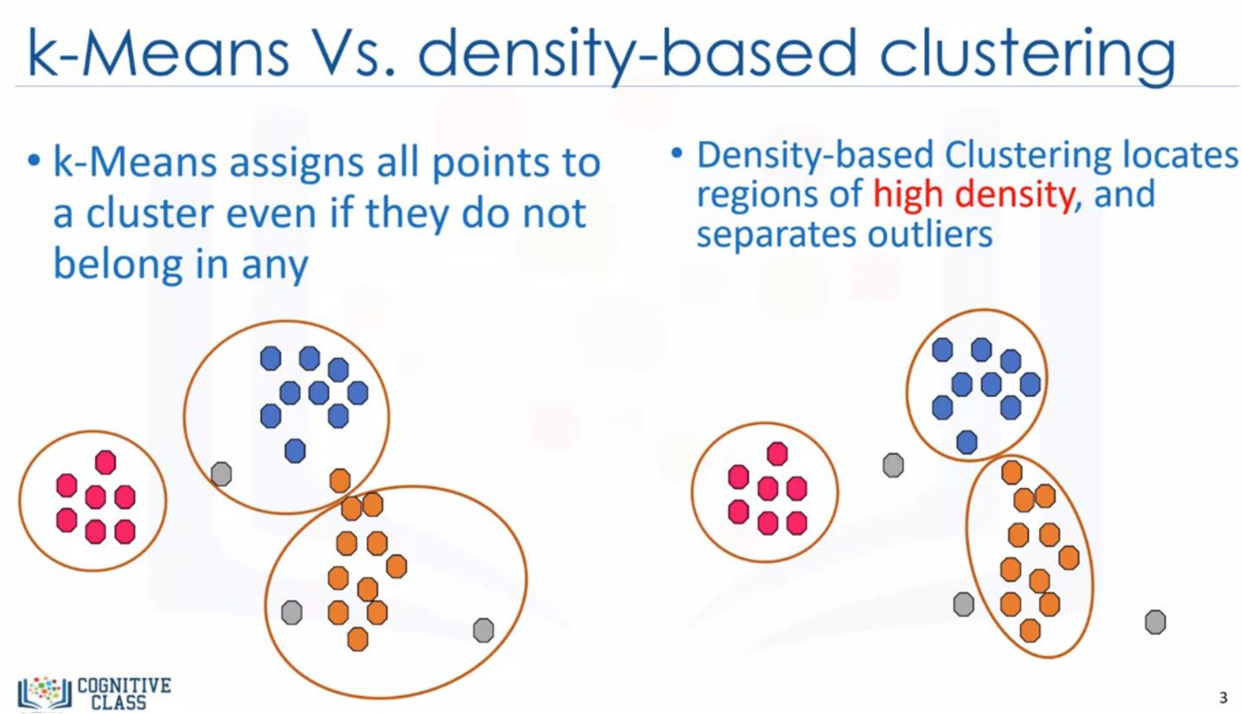


When applied to tasks with arbitrary shaped clusters or cluster within clusters, traditional techniques might not be able to achieve good results, that is elements in the same cluster might not share enough similarity or the performance may be poor.



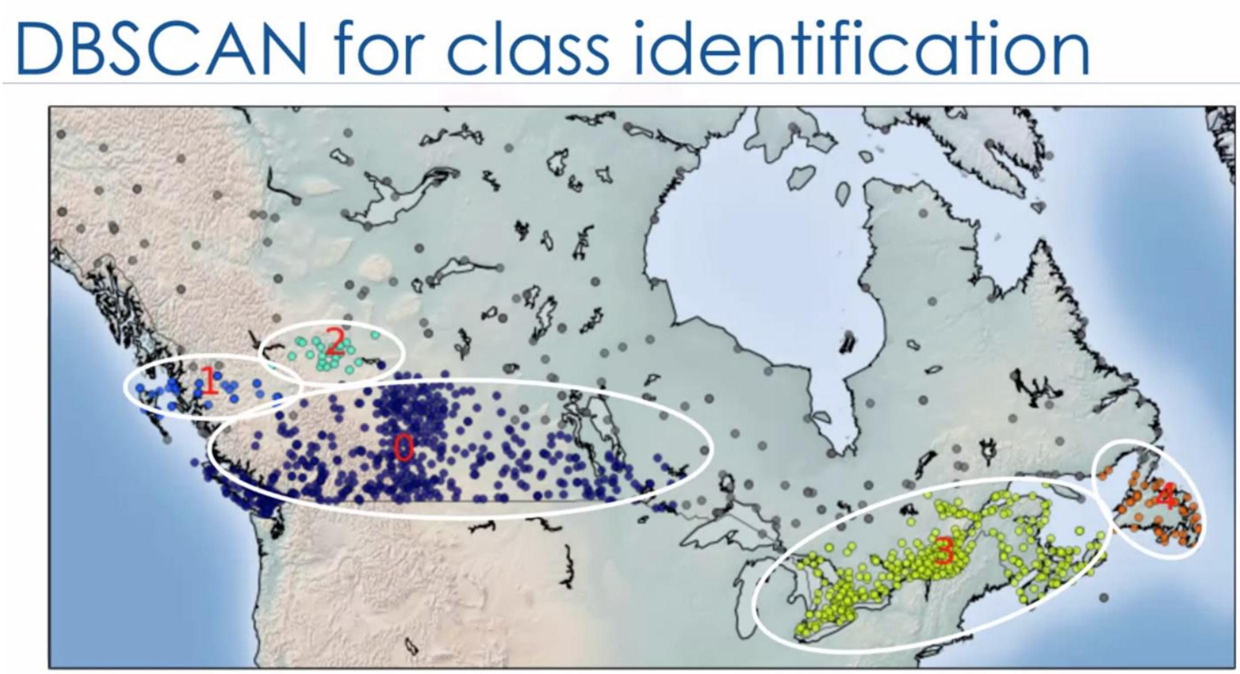
K-Means algorithm has no notion of outliers that is , all points are assigned to a cluster even if they do not belong in any.

In the domain of anomaly detection, this causes problems as anomalous points will be assigned to the same cluster as normal data points.

The anomalous points pull the cluster centroid towards them making it harder to classify them as anomalous points.

In contrast, density-based clustering locates regions of high density that are separated from one another by regions of low density.

Density in this context is defined as the number of points within a specified radius.

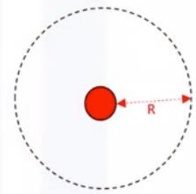


A specific and very popular type of density-based clustering is DBSCAN. DBSCAN is particularly effective for tasks like class identification on a spatial context. It can find out any arbitrary shaped cluster without getting effected by noise.

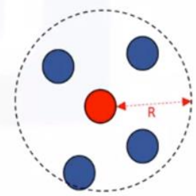
What is DBSCAN?

- DBSCAN (**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise)
  - Is one of the most common clustering algorithms
  - Works based on density of objects

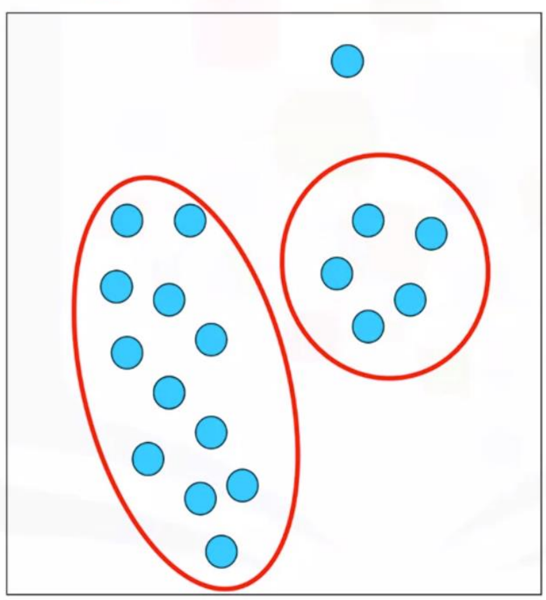
- R (**R**adius of neighborhood)
  - Radius (R) that if includes enough number of points within, we call it a dense area



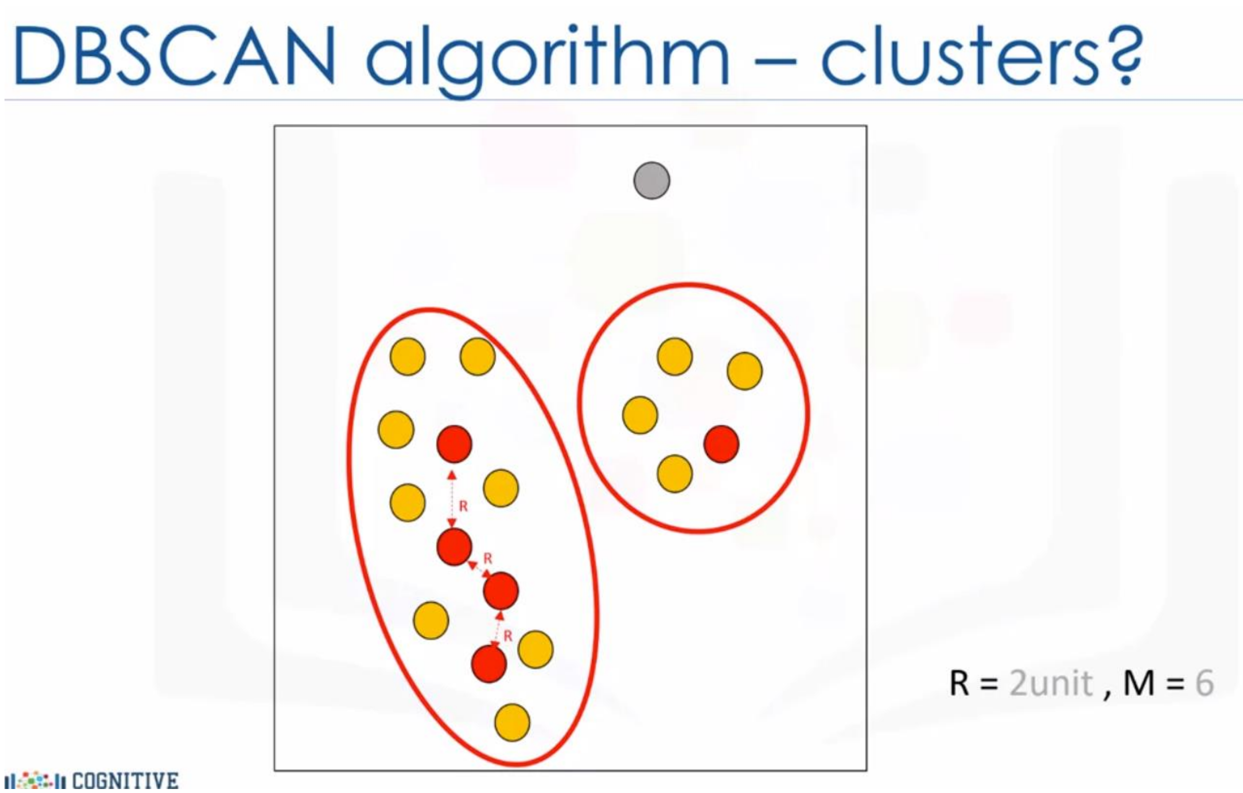
- M (**M**in number of neighbors)
  - The minimum number of data points we want in a neighborhood to define a cluster



Advantages of DBSCAN



1. Arbitrarily shaped clusters
2. Robust to outliers
3. Does not require specification of the number of clusters



Core point: there are 6 points in the two centimeter neighbor of the red point, we mark this point as a core point

Border point: its neighbourhood contains less than M data points, or it is reachable from some core points. Here reachability means it is within our distance from a core point. It means that even though the yellow point is within the 2 centimeter neighborhood of the red point, it is not by itself a core point because it does not have at least 6 points in its neighborhood.

Outlier: a point that is not a core point and also is not close enough to be reachable from a core point.

We continue and visit all the points in the dataset and label them as either core, border, or outlier.

The next step is to connect core points that are neighbors and put them in the same cluster.