# Density-based Spatial Clustering of Application with Noise

DBSCAN is a density-based clustering algorithm. In other words, it groups data considering that a cluster has high density, while the region that separates them does not. The main idea is that each point in the cluster must have a minimum number of neighbors or nearby points to determine density.

The parameters are:

1. **Epsilon (eps):** The maximum distance between two points to be considered neighbors (radius). If epsilon is very small, the algorithm may take a long time to converge. If it is very large, it may not converge.
2. **minPts:** The minimum number of neighbors a point must have to be considered a core point with high density.
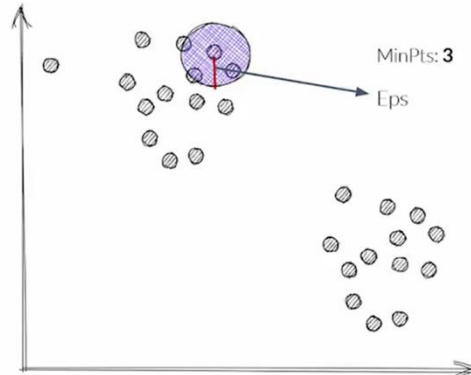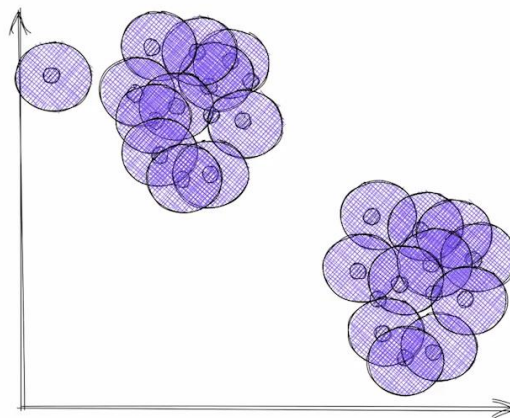

The operation of DBSCAN is based on two main concepts:

- **Core point:** A point that has at least **minPts** observations within a distance less than or equal to eps.
- **Reachable point:** A point that is within a distance less than or equal to eps from a core point.

Mathematical fundamentals

1. For each observation, it is checked whether it is a core point.

2. If the observation is a core point, a new cluster is created, and all reachable observations are added to it.

3. If the observation is a border point, it is added to the cluster of the core point to which it belongs.
4. If the observation is not a reachable point, it is classified as noise.
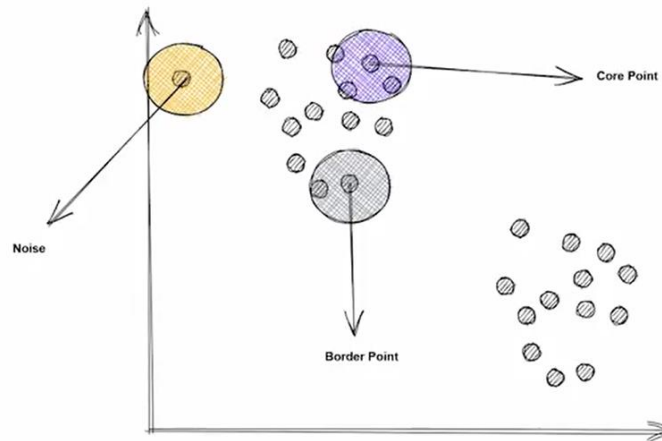
When a core point connects with another core point, they merge, and the cluster grows larger. However, a border point does not connect with others.

Useful cases

DBSCAN is an efficient and robust algorithm, particularly useful for problems where clusters have irregular shapes or are not clearly defined. For example:

- Image segmentation
- Clustering of geographical data
- Clustering of social network data (grouping users)

Advantages of DBSCAN

- No need to define the number of clusters beforehand.
- Detects outliers.
- Clusters non-spherical shapes

Disadvantages of DBSCAN

- Choosing the hyperparameters eps and minPts can be difficult and greatly influence performance.
- Border points may be arbitrarily assigned to any cluster.
- Not suitable for problems where clusters have regular shapes.

*See bibliography in the project references.