# Machine Learning – Density Based Clustering DBSCAN Clustering

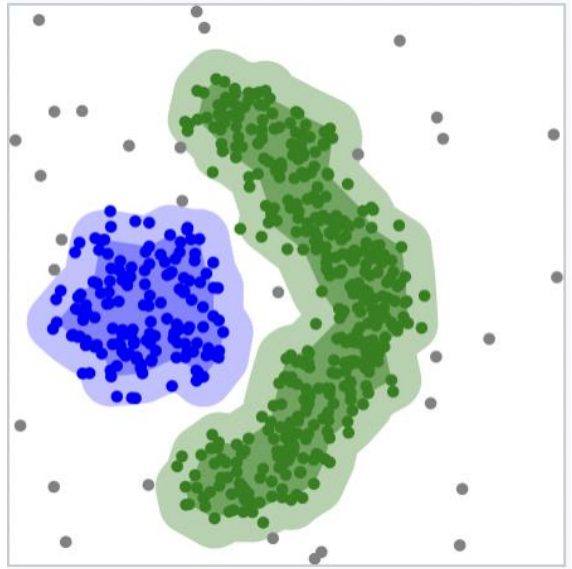
DR. BHARGAVI R

SCOPE

VIT CHENNAI

# Introduction

- Partitioning based and hierarchical clustering methods severely affected by the presence of noise and outliers in the data
- DBSCAN - Density-based spatial clustering of applications with noise.
- Clusters are dense regions in the data space, separated by regions of the lower density of points.
- DBSCAN requires minimum domain knowledge.
- It can discover clusters of arbitrary shape.
- Efficient for large databases.



# DBSCAN

- DBSCAN uses two parameters for measuring the density of regions
- Epsilon(eps): Defines the neighborhood around a data point
  - If the distance between two points is less or equal to ‘eps’ then they are considered as neighbors.
  - If the eps value is chosen too small then large part of the data will be considered as outliers.
  - If it is chosen very large then the clusters will merge and majority of the data points will be in the same clusters.

# Basics (cont…)

- MinPts: Minimum number of neighbors (data points) within eps radius.
- Larger the dataset, the larger value of MinPts must be chosen.
- As a general rule, the minimum MinPts can be derived from the number of dimensions D in the dataset as, MinPts >= D+1.
- The minimum value of MinPts must be chosen at least 3.

# Data Points - Different types

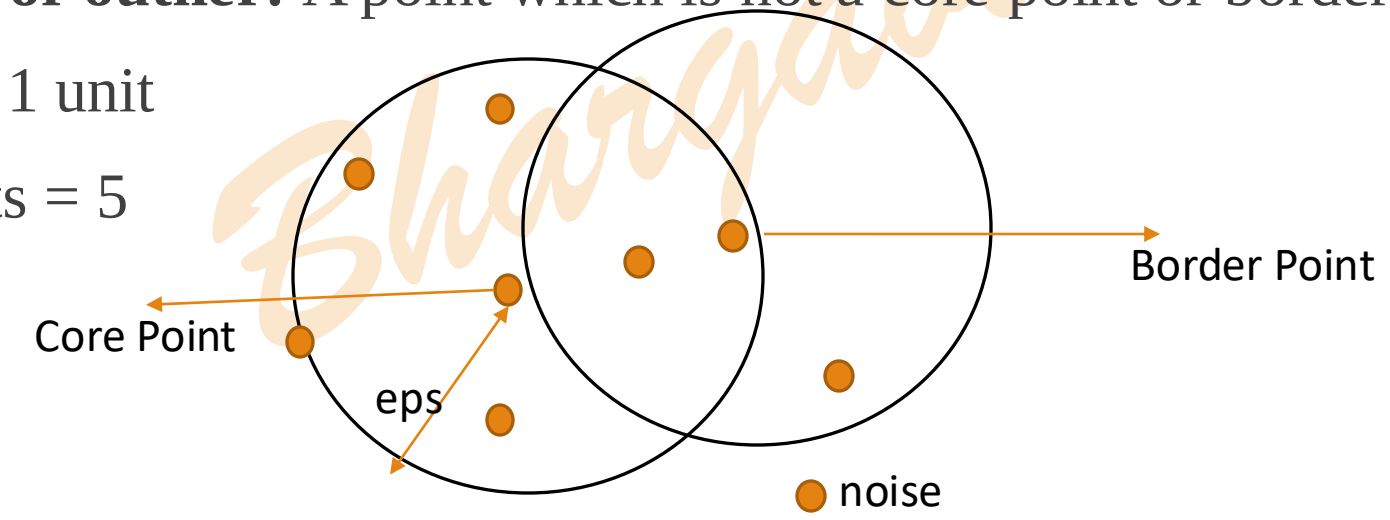
**Core Point:** A data point is a core point if it has atleast MinPts data points within eps neighborhood.

**Border Point:** A data point which has fewer than MinPts within eps but it is in the neighborhood of a core point.

**Noise or outlier:** A point which is not a core point or border point.

Eps = 1 unit

MinPts = 5



# Cont…

**Directly Density Reachable**: Data-point **a** is directly density reachable from a point **b** if

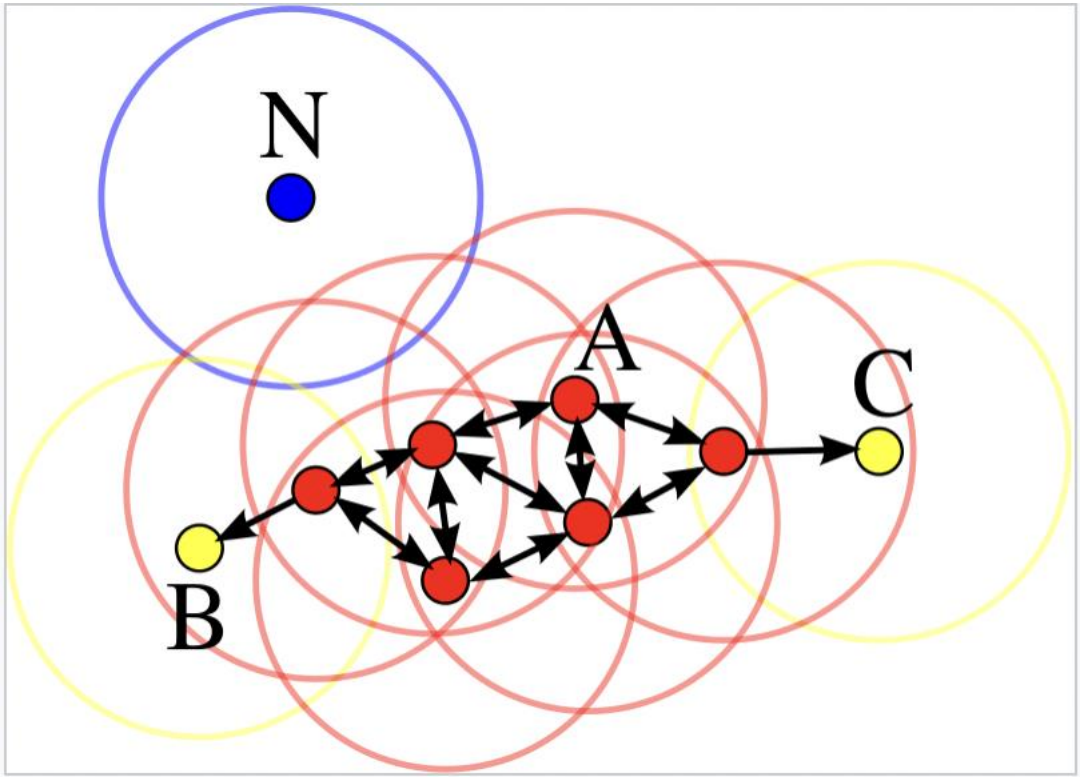
- Epslon neighborhood of b >= MinPts i.e **b** is a core point
- Data point **a** is in the epsilon neighborhood of **b**.

**Density Reachable:** Point **a** is density reachable from a point **b** with respect to eps and *MinPts*, if there exists a chain of poitns $b_1 = b$ , $b_2$, $b_3$, …., $b_n = a$, such that $b_{i+1}$ is directly density reachable from $b_i$.

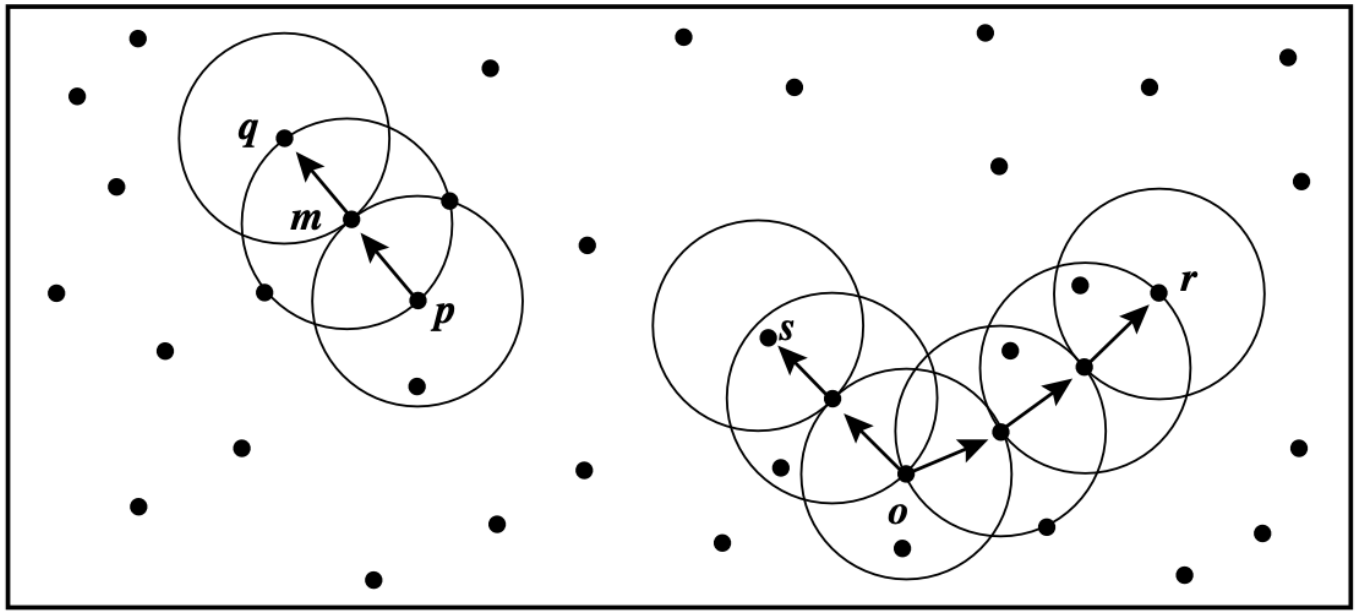
**Density Connected:** There can be cases when 2 border points will belong to the same cluster but they don't share a specific core point, then we say that they are density connected if, there exists a common core point, from which these borders points are density reachable.

# Cont…

MinPts = 4



# Cont….




Let MinPts = 3. Then,
Core Points : ***m, p, o,*** and ***r***
Directly density reachable : ***q*** is directly density reachable from ***m***
Density reachable : *q* is (indirectly) density reachable from ***p***. ***q*** is directly density-reachable from **m** and ***m*** is directly density-reachable from **p.** But p is not density reachable from q since q is not a core point.
Similarly, r and s are density-reachable from o, and o is density-reachable from r.

# Algorithm

- Find all the neighbor points within eps and identify the core points.
- For each core point if it is not already assigned to a cluster, create a new cluster.
- Find recursively all its density connected points and assign them to the same cluster as the core point.
  - A point **a** is density connected to a point **b** with respect to eps and MinPts, if there is a point **c** such that, both **a** and **b** are density reachable from **c** w.r.t. to eps and MinPts
- Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.

# Example

| ID | X1 | X2 |
|---|---|---|
| 1 | 3 | 7 |
| 2 | 4 | 6 |
| 3 | 5 | 5 |
| 4 | 6 | 4 |
| 5 | 7 | 3 |
| 6 | 6 | 2 |
| 7 | 7 | 2 |
| 8 | 8 | 4 |
| 9 | 3 | 3 |
| 10 | 2 | 6 |
| 11 | 3 | 5 |
| 12 | 2 | 4 |

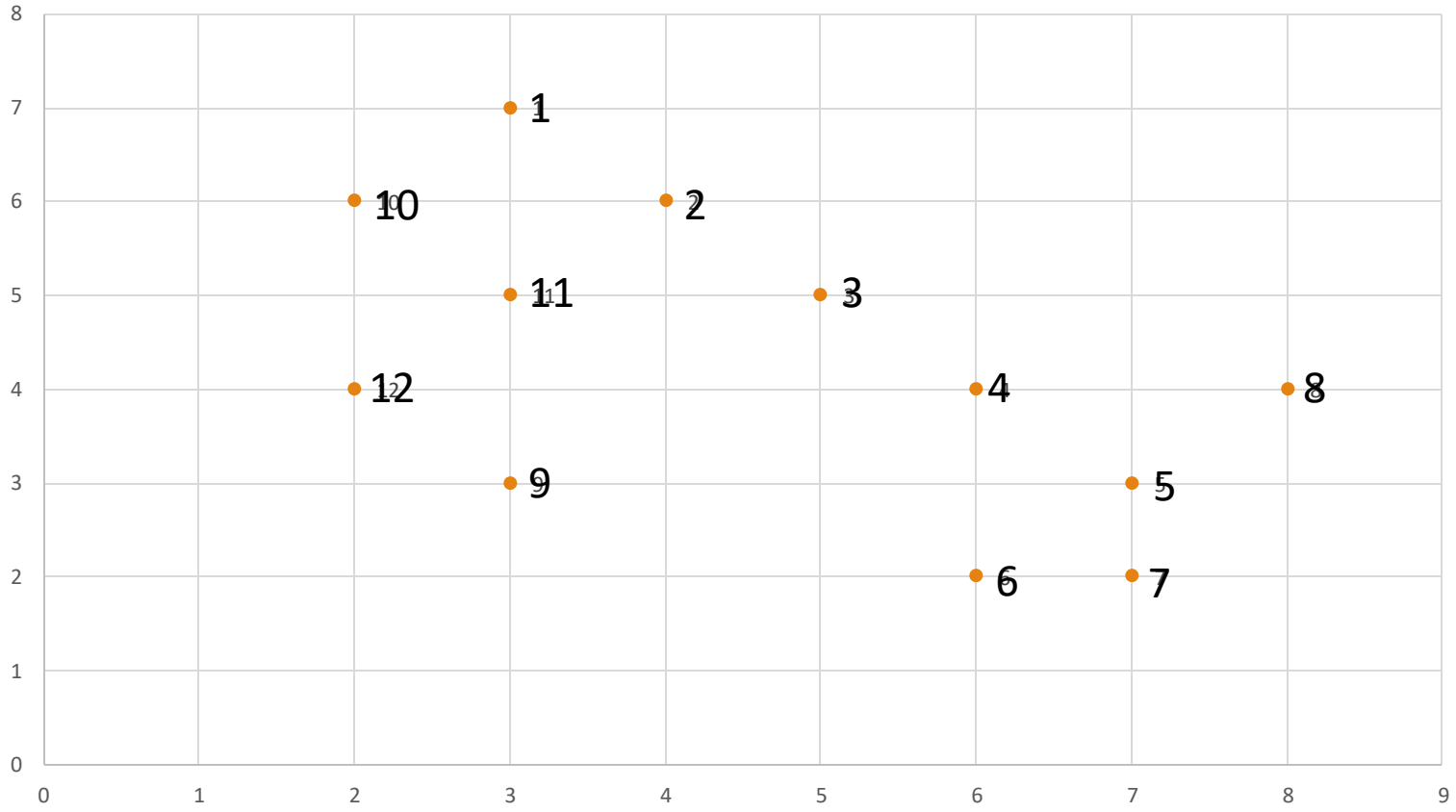
Eps = 1.9
MinPts =4

# Distance Computations

| ID | X1 | X2 | 1 (3,7) | 2(4,6) | 3(5,5) | 4(6,4) | 5(7,3) | 6(6,2) | 7(7,2) | 8(8,4) | 9(3,3) | 10(2,6) | 11(3,5) | 12(2,4) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 7 | 0.00 | | | | | | | | | | | |
| 2 | 4 | 6 | 1.41 | 0.00 | | | | | | | | | | |
| 3 | 5 | 5 | 2.83 | 1.41 | 0.00 | | | | | | | | | |
| 4 | 6 | 4 | 4.24 | 2.83 | 1.41 | 0.00 | | | | | | | | |
| 5 | 7 | 3 | 5.66 | 4.24 | 2.83 | 1.41 | 0.00 | | | | | | | |
| 6 | 6 | 2 | 5.83 | 4.47 | 3.16 | 2.00 | 1.41 | 0.00 | | | | | | |
| 7 | 7 | 2 | 6.40 | 5.00 | 3.61 | 2.24 | 1.00 | 1.00 | 0.00 | | | | | |
| 8 | 8 | 4 | 5.83 | 4.47 | 3.16 | 2.00 | 1.41 | 2.83 | 2.24 | 0.00 | | | | |
| 9 | 3 | 3 | 4.00 | 3.16 | 2.83 | 3.16 | 4.00 | 3.16 | 4.12 | 5.10 | 0.00 | | | |
| 10 | 2 | 6 | 1.41 | 2.00 | 3.16 | 4.47 | 5.83 | 5.66 | 6.40 | 6.32 | 3.16 | 0.00 | | |
| 11 | 3 | 5 | 2.00 | 1.41 | 2.00 | 3.16 | 4.47 | 4.24 | 5.00 | 5.10 | 2.00 | 1.41 | 0.00 | |
| 12 | 2 | 4 | 3.16 | 2.83 | 3.16 | 4.00 | 5.10 | 4.47 | 5.39 | 6.00 | 1.41 | 2.00 | 1.41 | 0.00 |

| Data Point | Data points with in eps distance |
|---|---|
| 1 | 2, 10 |
| 2 | 1, 3, 11 |
| 3 | 2, 4 |
| 4 | 3, 5 |
| 5 | 4, 6, 7, 8 |
| 6 | 5, 7 |
| 7 | 5, 6 |
| 8 | 5 |
| 9 | 12 |
| 10 | 1, 11 |
| 11 | 2, 10, 12 |
| 12 | 9, 11 |

| Data Point | Status | |
|---|---|---|
| 1 | Noise | Border |
| 2 | Core | |
| 3 | Noise | Border |
| 4 | Noise | Border |
| 5 | Core | |
| 6 | Noise | Border |
| 7 | Noise | Border |
| 8 | Noise | Border |
| 9 | Noise | |
| 10 | Noise | Border |
| 11 | Core | |
| 12 | Noise | Border |


8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
8
9
1
10
2
11
3
12
4
8
9
5
6
7