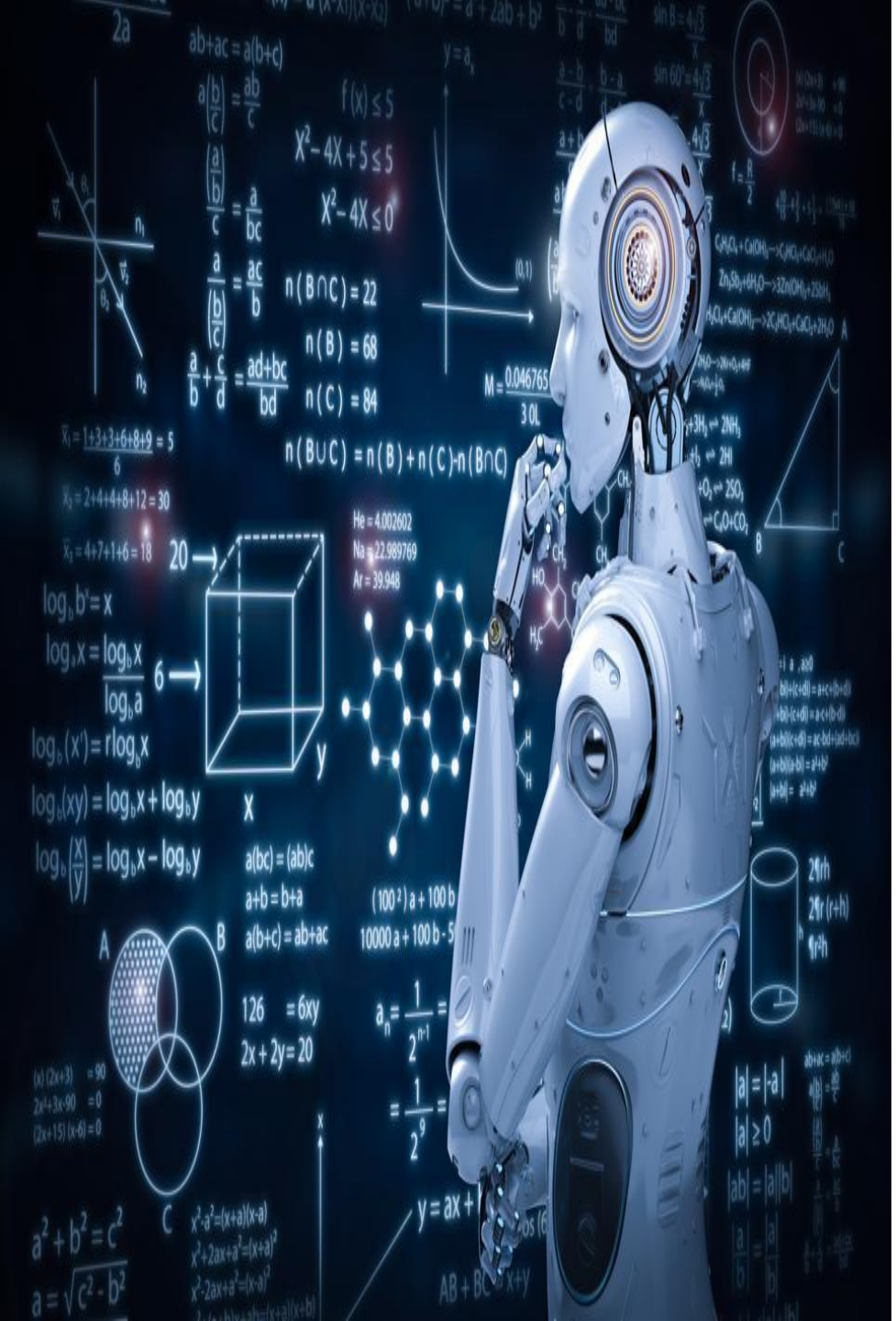


Practical Machine Learning

Day 15: Sep22 DBDA

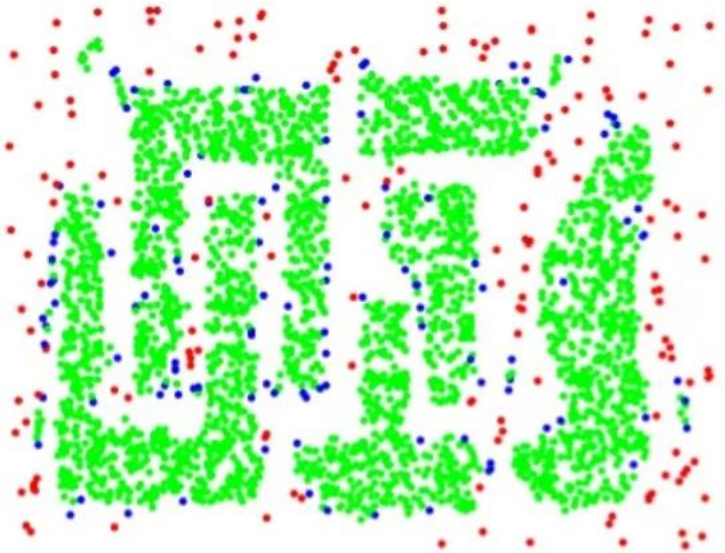
Kiran Waghmare

Agenda

- Clustering
- K-Means
- Hierarchical
- DB-SCAN

Concepts: Preliminary

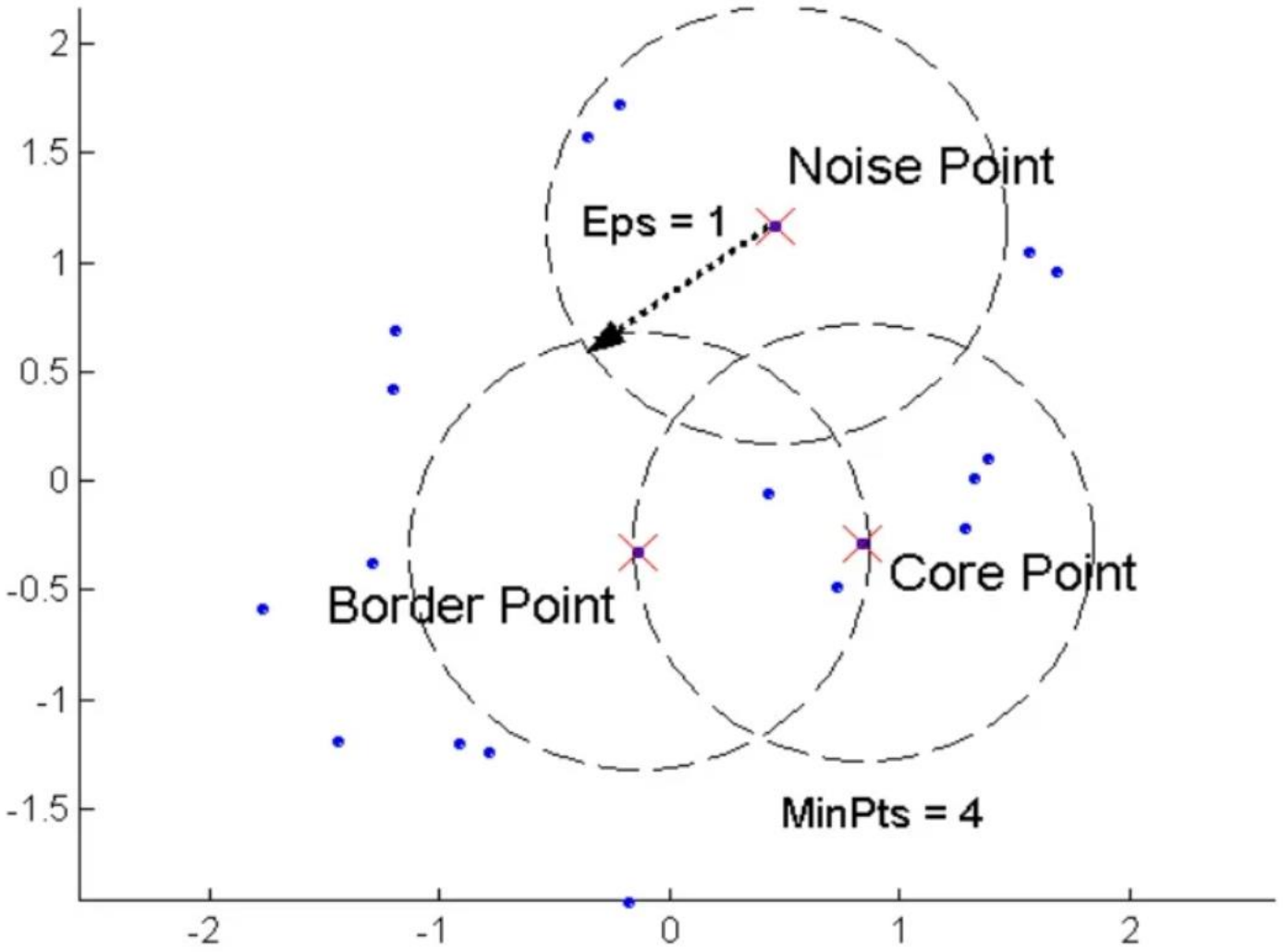
Original Points



Point types: core, border and noise

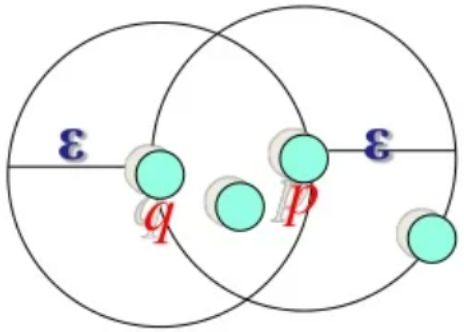
Eps = 10, MinPts = 4

Concepts: Core, Border, Noise



Concepts: ε-Neighborhood

- ε-Neighborhood - Objects within a radius of ε from an object. (epsilon-neighborhood)
- Core objects - ε-Neighborhood of an object contains at least MinPts of objects



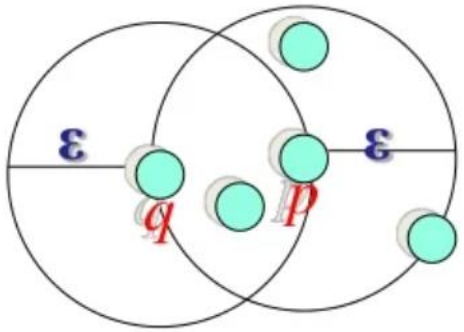
ε-Neighborhood of *p*
ε-Neighborhood of *q*

p is a core object (MinPts = 4)

q is not a core object

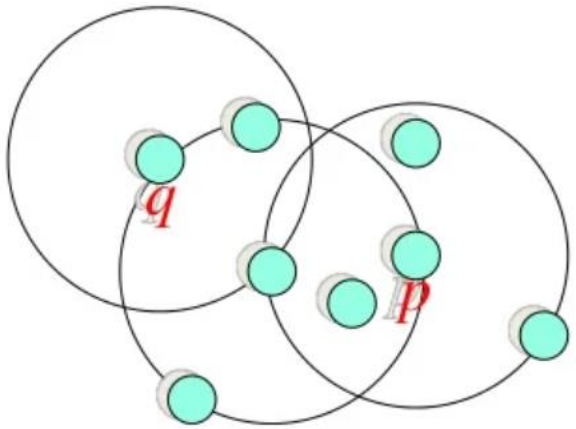
DBScan : Reachability

- **Directly density-reachable**
  - An object q is directly density-reachable from object p if q is within the ε-Neighborhood of p and p is a core object.

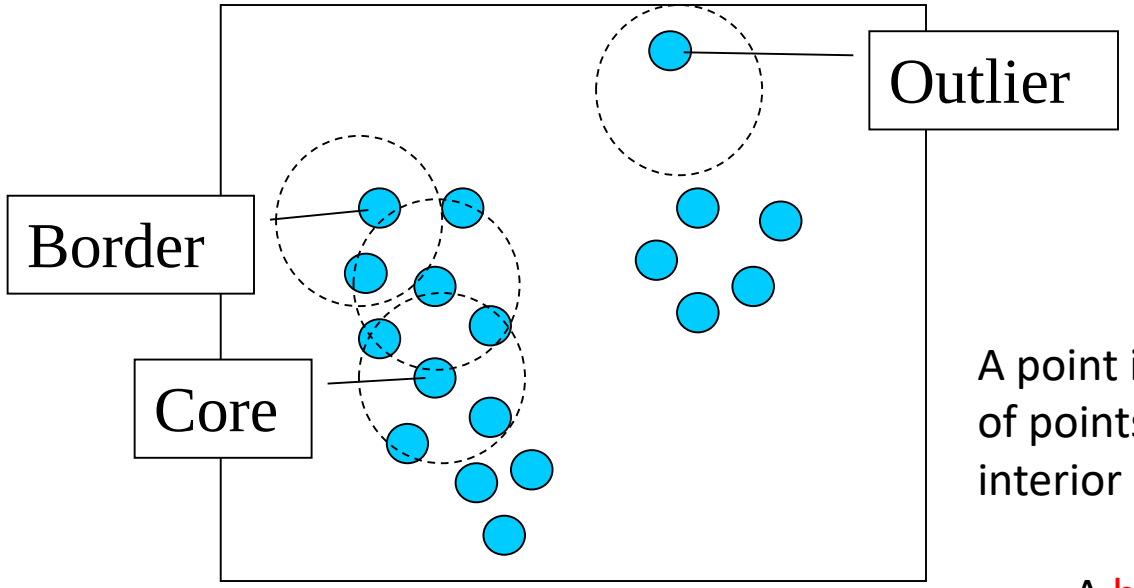


- q is directly density-reachable from p
- p is not directly density-reachable from q.

DBScan : Reachability



Core, Border & Outlier



Given ε and *MinPts*, categorize the objects into three exclusive groups.

ε = 1unit, MinPts = 5

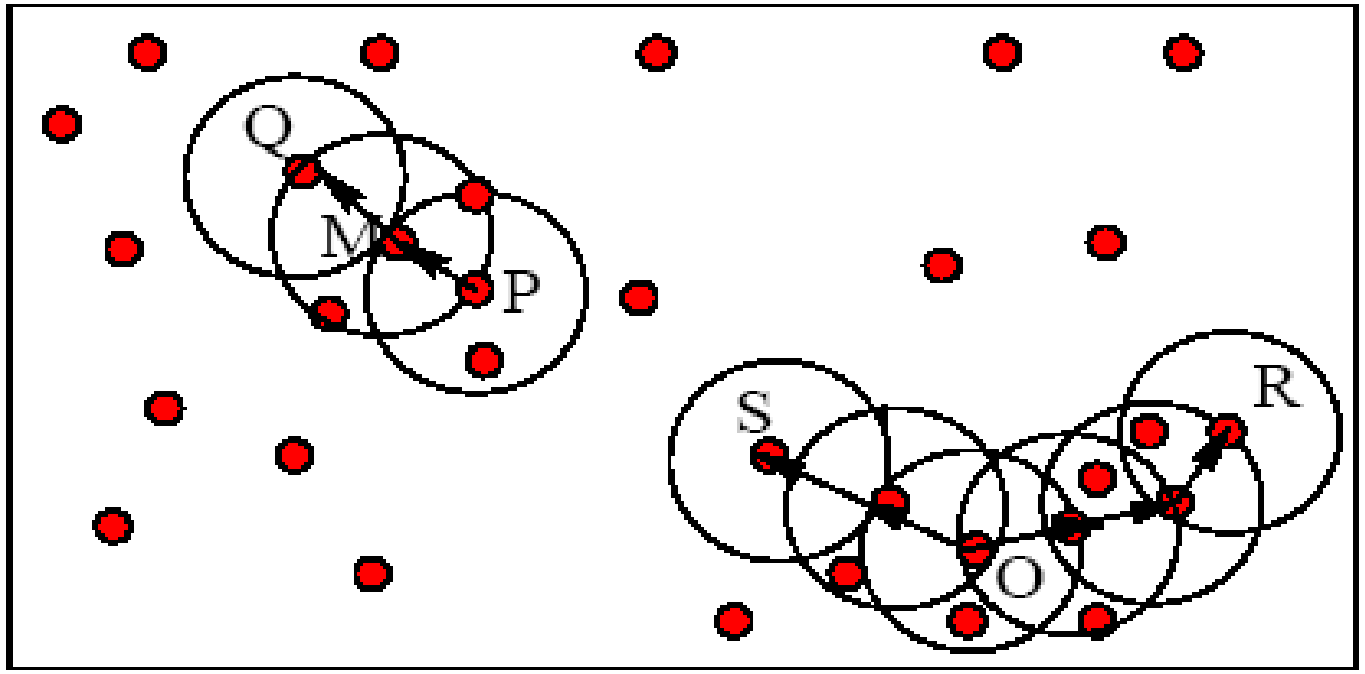
A point is a core point if it has more than a specified number of points (MinPts) within Eps These are points that are at the interior of a cluster.

A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point.

A noise point is any point that is not a core point nor a border point.

Example

- M, P, O, and R are core objects since each is in an Eps neighborhood containing at least 3 points

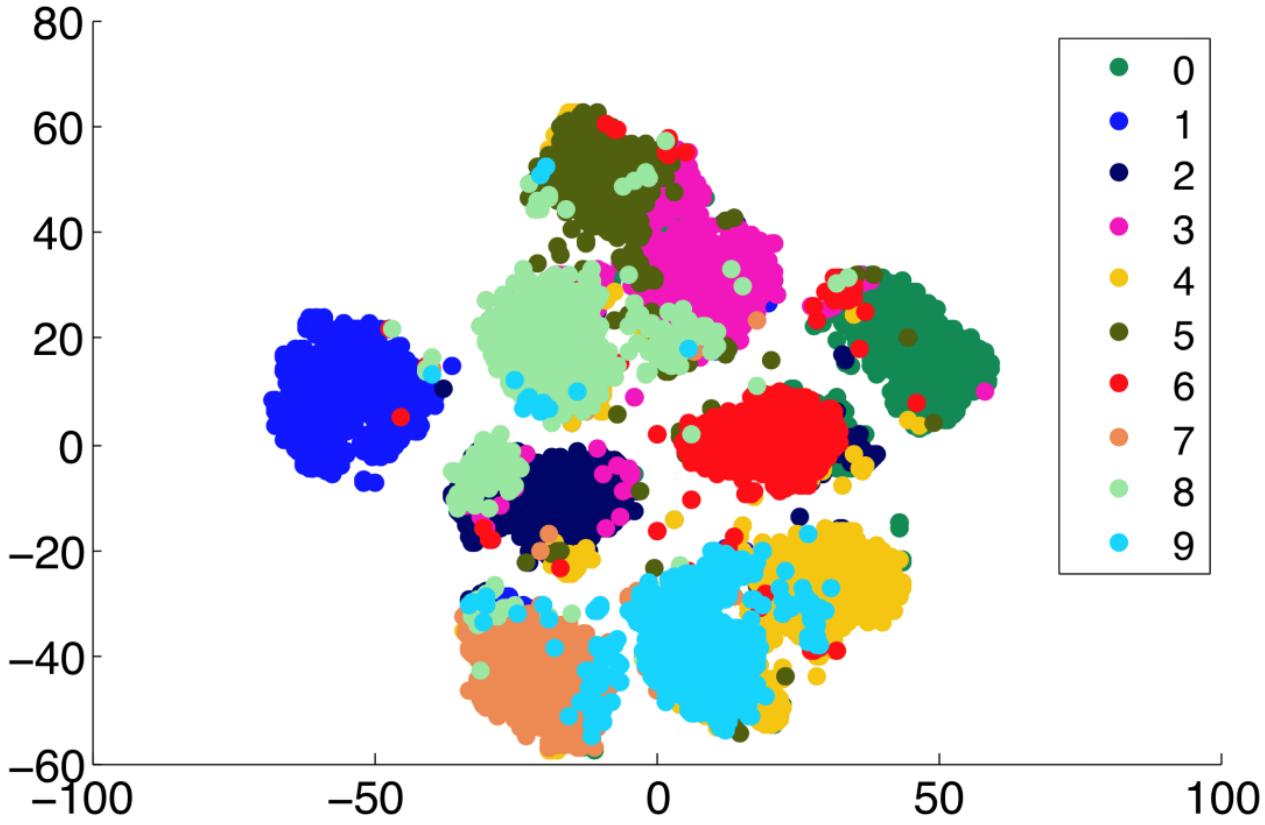


Minpts = 3

Eps=radius
of the circles

Reminder: Unsupervised Learning

- There are no labels for the training phase
- Our goal is to discover structure in data



DBScan Algorithm

Input: N objects to be clustered and global parameters *Eps*, *MinPts*.

Output: Clusters of objects.

Algorithm:

1) Arbitrary select a point *P*.
2) Retrieve all points density-reachable from *P* wrt ***Eps*** and ***MinPts***.
3) If *P* is a core point, a cluster is formed.
4) If *P* is a border point, no points are density-reachable from *P* and **DBSCAN** visits the next point of the database.
5) Continue the process until all of the points have been processed.

DBScan :Flowchart