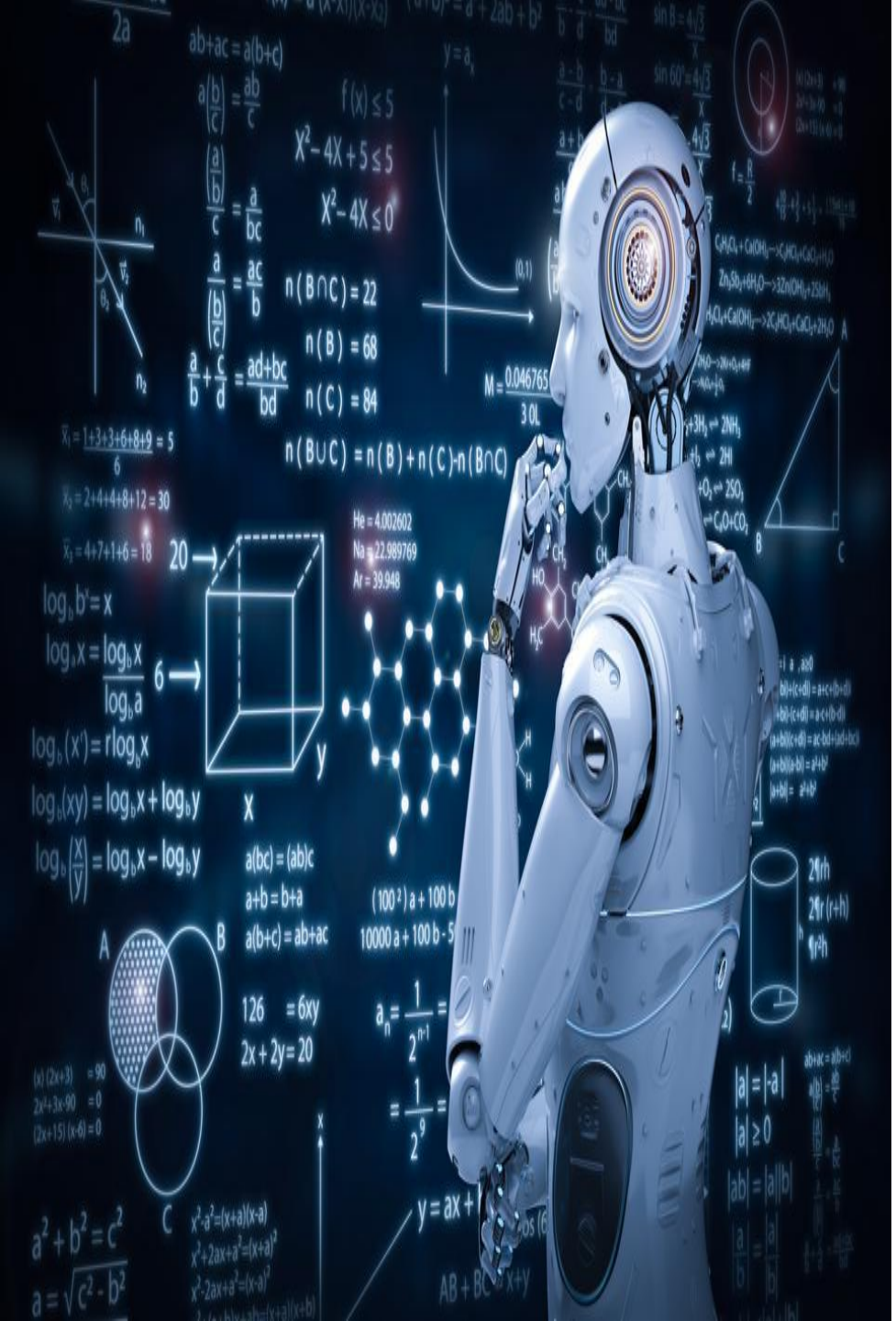


Practical Machine Learning

Day 14: Mar23 DBDA

Kiran Waghmare

Agenda

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

Concepts: Preliminary

- **DBSCAN is a density-based algorithm**

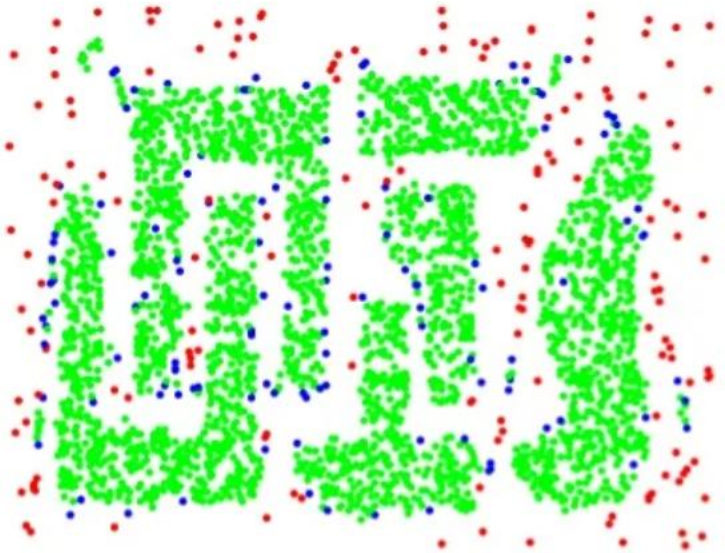
- DBScan stands for Density-Based Spatial Clustering of Applications with Noise

- Density-based Clustering locates regions of high density that are separated from one another by regions of low density

  Density = number of points within a specified radius (Eps)

Concepts: Preliminary

Original Points



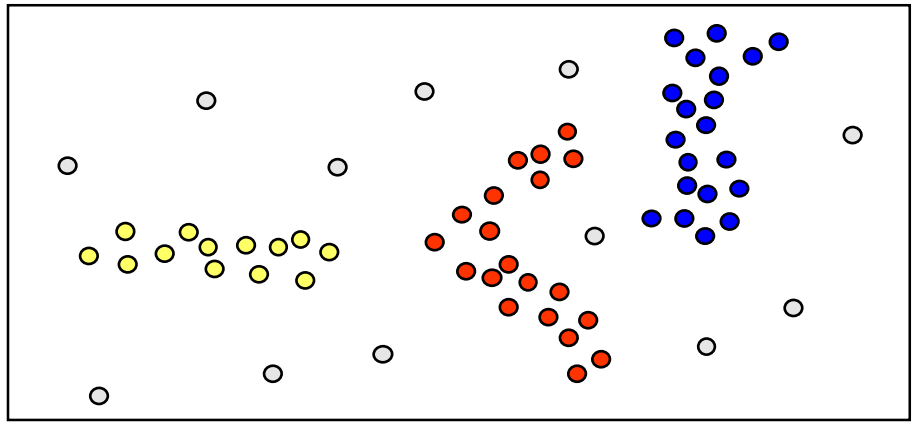
Point types: core, border and noise

Eps = 10, MinPts = 4

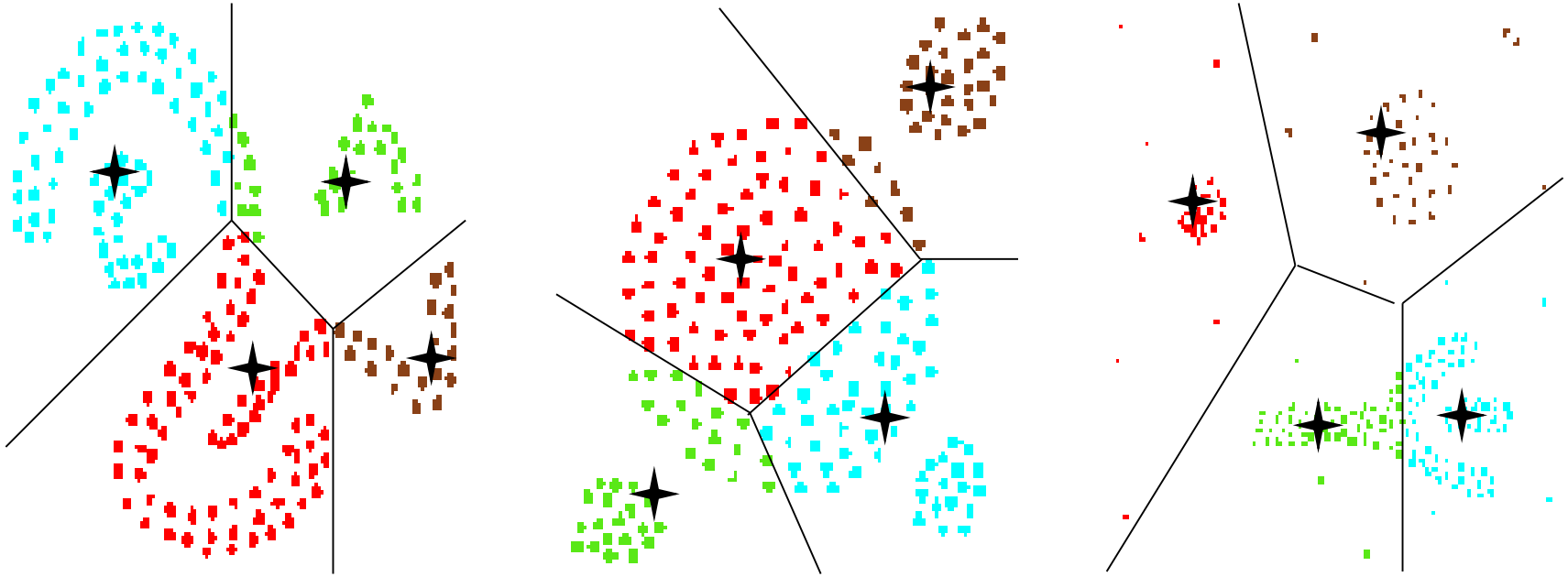
Density-Based Clustering

- *Basic Idea*:

  Clusters are dense regions in the data space, separated by regions of lower object density



- **Why Density-Based Clustering?**



Results of a *k*-medoid algorithm for *k*=4

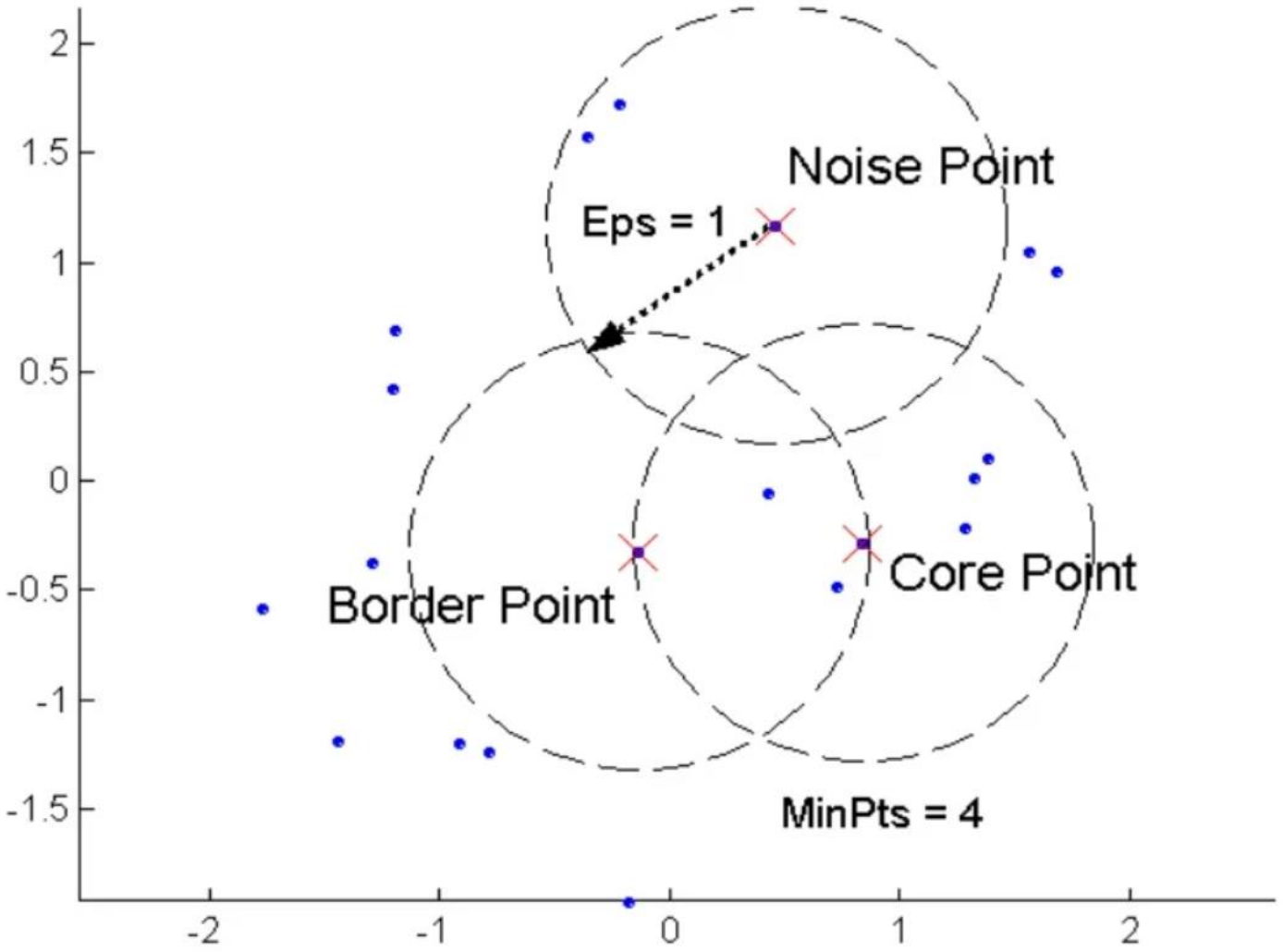
Concepts: Preliminary

- 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 or a border point

Concepts: Preliminary

- Any two core points are close enough– within a distance *Eps* of one another – are put in the same cluster
- Any border point that is close enough to a core point is put in the same cluster as the core point
- Noise points are discarded

Concepts: Core, Border, Noise



Parameter Estimation

parameters must be specified by the user.

ε = physical distance(radius),
minPts = desired minimum cluster size

minPts

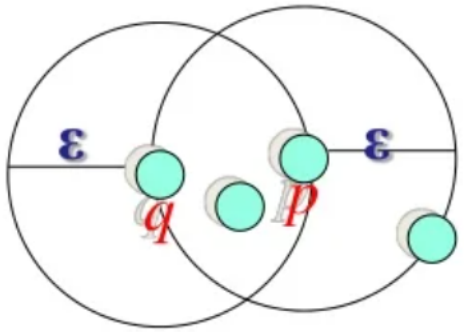
- derived from the number of dimensions *D* in the data set, as *minPts* ≥ *D* + *1*
- *minPts* = *1* does not make sense, as then every point on its own will already be a cluster
- *minPts* must be chosen at least 3. larger is better.
- larger the data set, the larger the value of *minPts* should be chosen.

ε

- value can be chosen by using a k-distance graph.
- if ε is chosen much too small, a large part of the data will not be clustered.
- if too high value ,majority of objects will be in the same cluster
- In general, small values of ε are preferable.

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



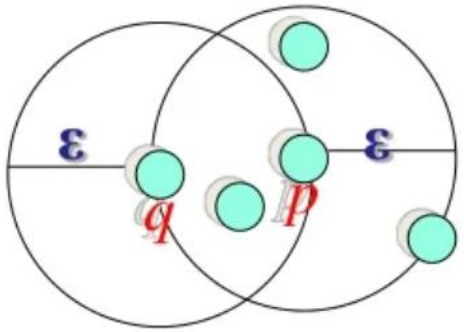
ε-Neighborhood of *p*
ε-Neighborhood of *q*

p is a core object (MinPts = 4)

q is not a core object

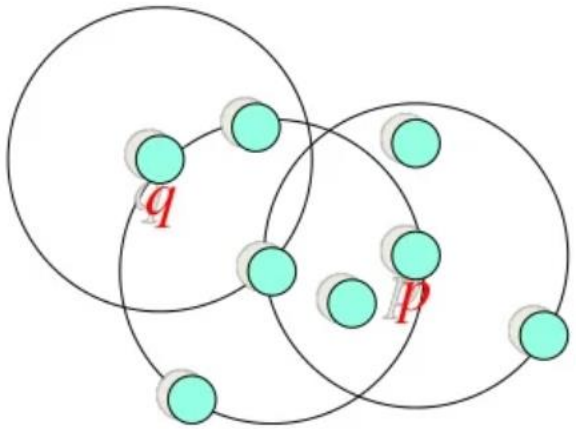
DBScan : Reachability

- **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

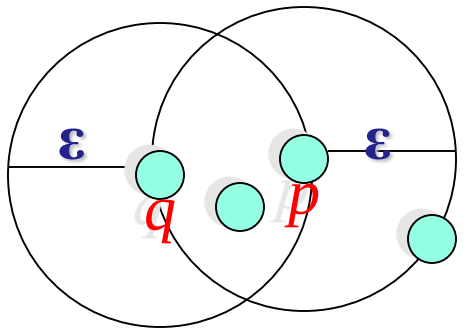


ε-Neighborhood

- ε-Neighborhood – Objects within a radius of ε from an object.

$$N_\varepsilon(p) : \{q \mid d(p,q) \leq \varepsilon\}$$

- "High density" - ε-Neighborhood of an object contains at least *MinPts* of objects.



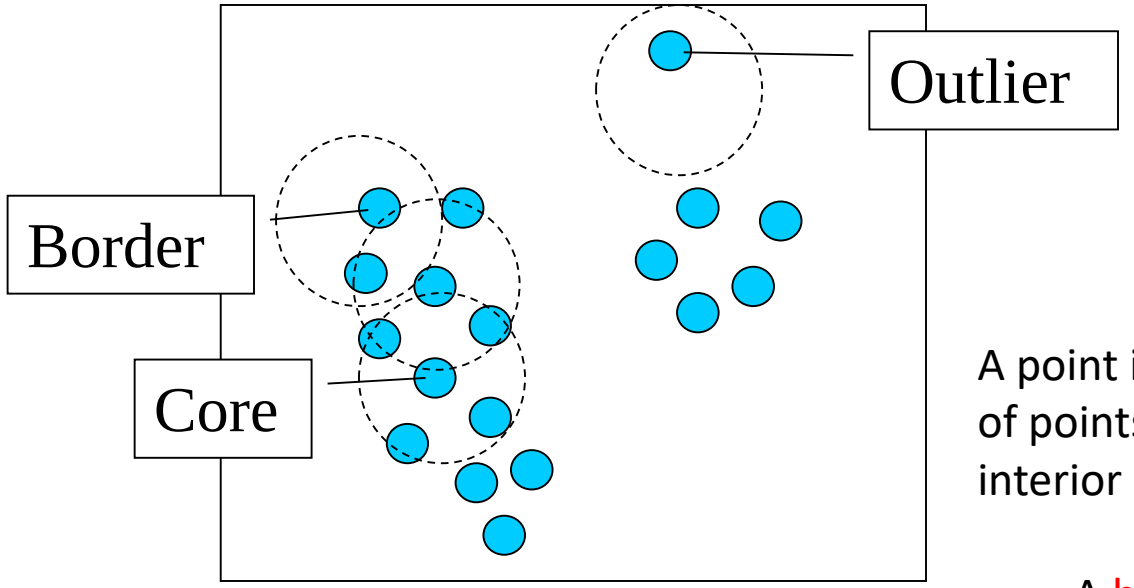
ε-Neighborhood of p

ε-Neighborhood of q

Density of p is "high" (MinPts = 4)

Density of q is "low" (MinPts = 4)

Core, Border & Outlier



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

ε = 1unit, MinPts = 5

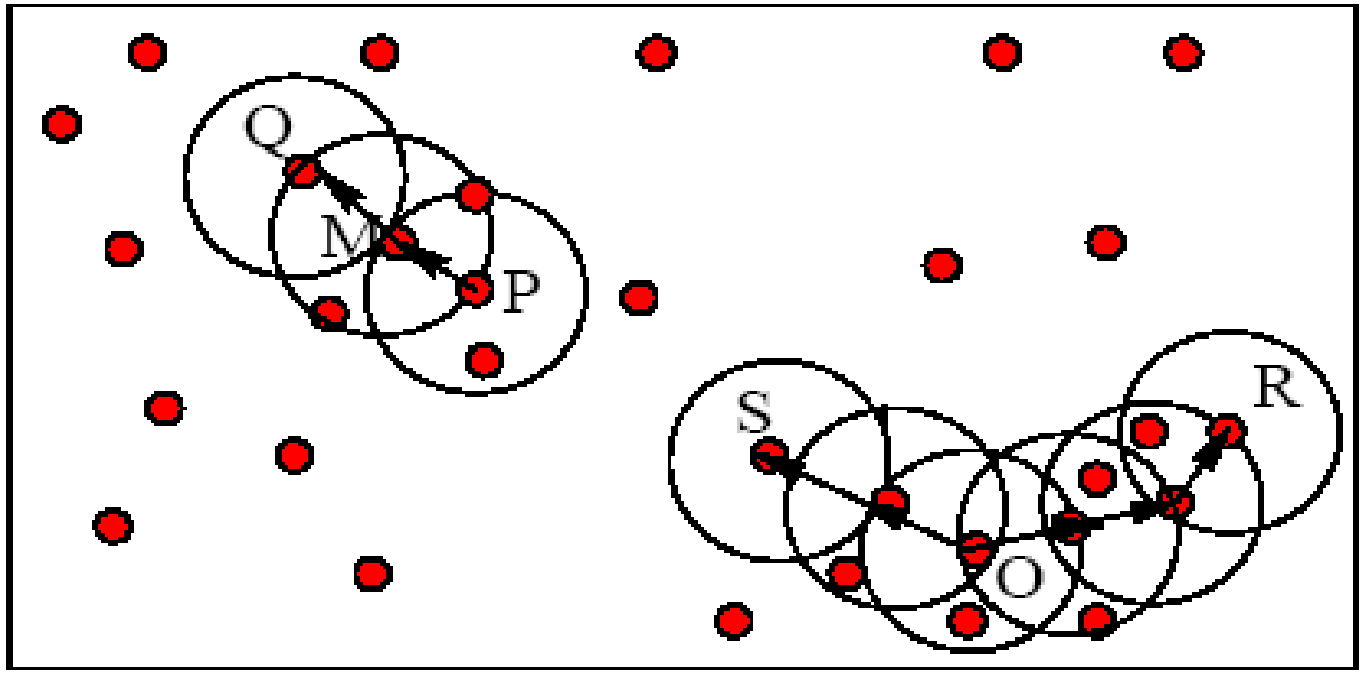
A point is a core point if it has more than a specified number of points (MinPts) within Eps These are points that are at the interior of a cluster.

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

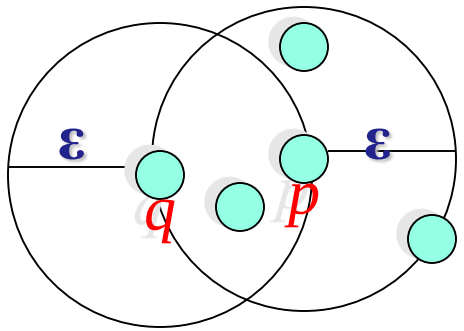


Minpts = 3

Eps=radius
of the circles

Density-Reachability

- **Directly density-reachable**
  - **An object q is directly density-reachable from object p if p is a core object and q is in p's ε-neighborhood.**

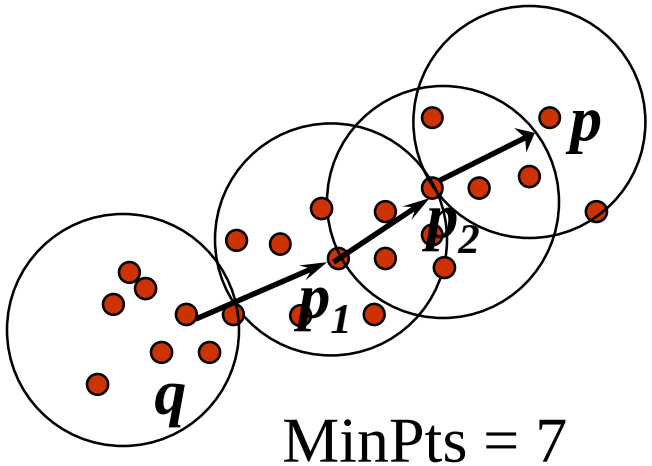


MinPts = 4

- **q is directly density-reachable from p**
- **p is not directly density- reachable from q?**
- **Density-reachability is asymmetric.**

Density-reachability

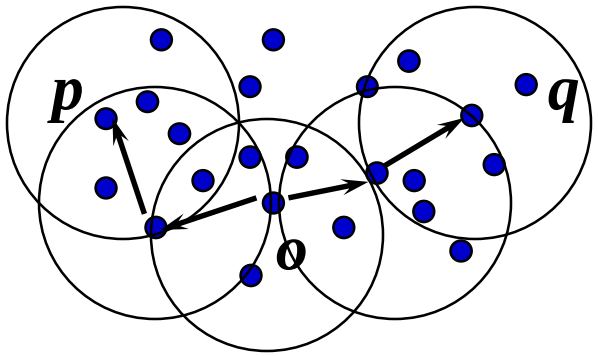
- Density-Reachable (directly and indirectly):
  - A point p is directly density-reachable from p2;
  - p2 is directly density-reachable from p1;
  - p1 is directly density-reachable from q;
  - p←p2←p1←q form a chain.



- p is (indirectly) density-reachable from q
- q is not density- reachable from p?

Density-Connectivity

- **Density-reachable is not symmetric**
  - not good enough to describe clusters
- **Density-Connected**
  - A pair of points p and q are density-connected if they are commonly density-reachable from a point o.



- Density-connectivity is symmetric