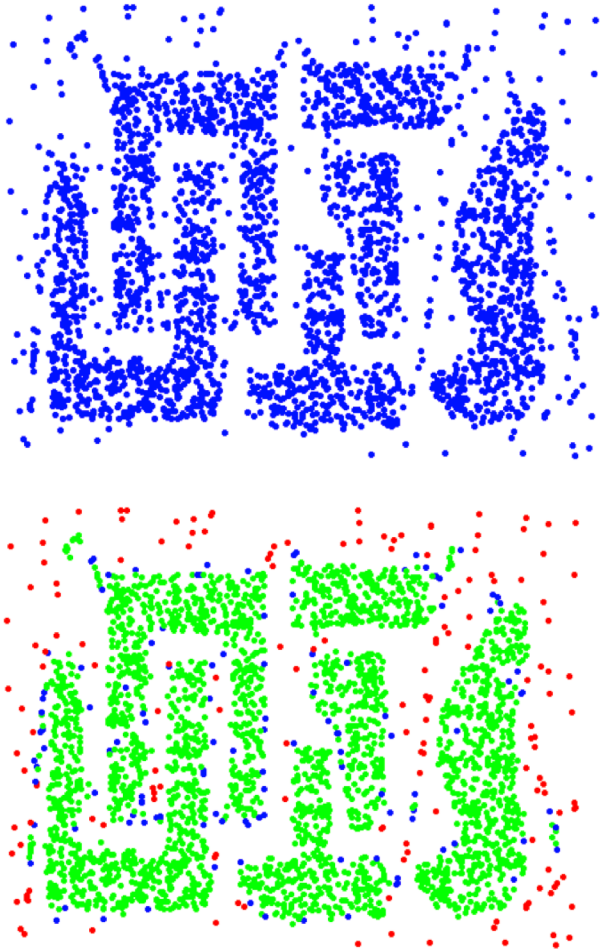
Density-based Clustering - DBSCAN

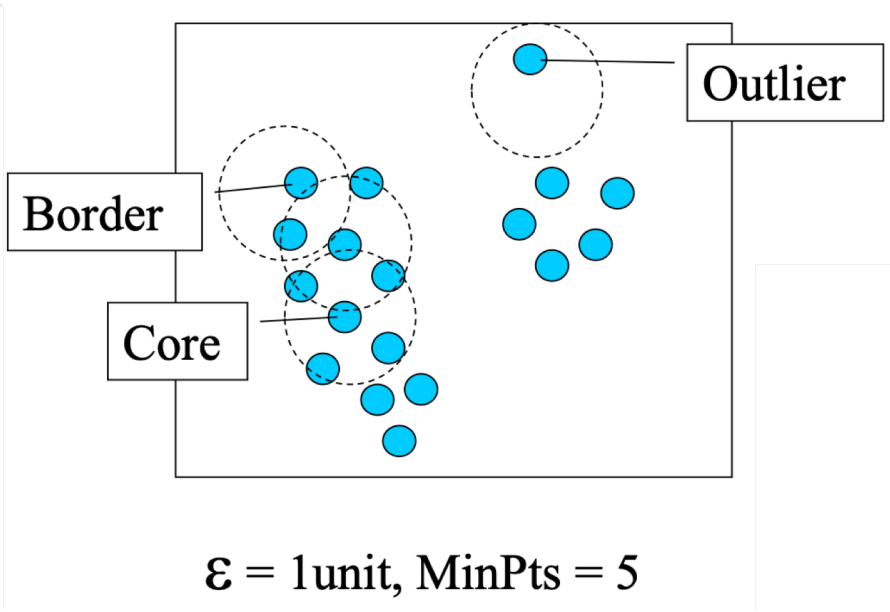
- Acronym for: Density-based spatial clustering of applications with noise
- Clusters are dense regions in the data space separated by regions of lower sample density.
- A cluster is defined as a maximal set of density connected points.
- Discover clusters of arbitrary shape.



Questions

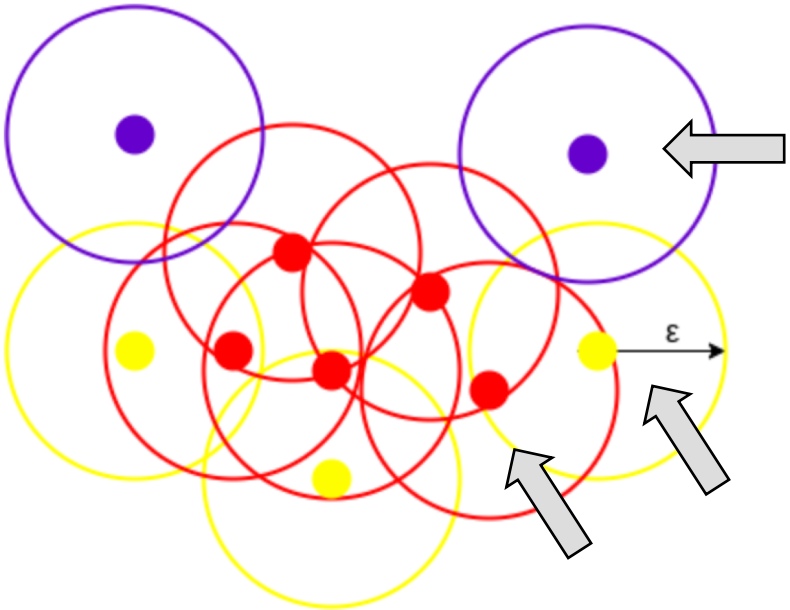
- What is a dense region?
- How do we measure density?

- Define three exclusive types of points
  Core, Border (or Edge) and Noise (or outlier)
  Core points -- dense region
  Noise -- sparse region
- Need two parameters
  1) a circle of *epsilon* radius
  2) a circle containing at least *minPts* number of points



Three types of points

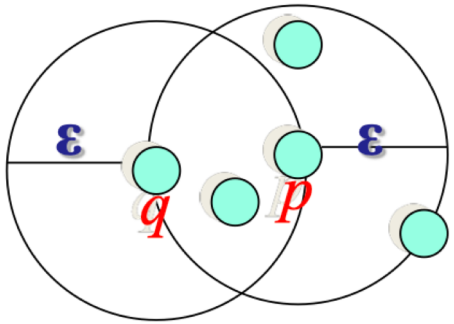
core	The point has at least minPts number of points within Eps
border	The point has fewer than minPts within Eps, but is in the neighbourhood (i.e. circle) of a core point.
noise	Any point that is not a core point or a border point.



minPts = 3
red: core
yellow: border
purple: noise

How to form core points into clusters? -- Density-reachability

- Directly density-reachable: a point *q* is directly density-reachable from point *p* if *p* is a core point and *q* is in *p*'s neighbourhood.

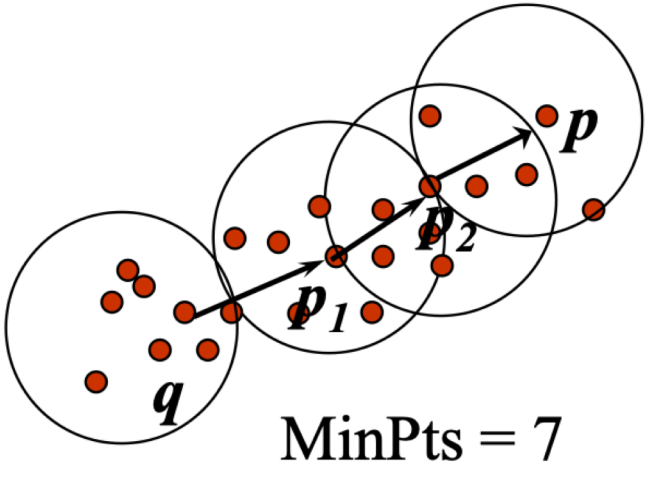


MinPts = 4

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

How to form core points into clusters? -- 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
  - q -> p1 -> p2 -> p form a chain
    (p is the border)



- p is indirectly density-reachable from q
- q is not density-reachable from p

The algorithm

- 1. Label all points as core, border or noise.
- 2. Eliminate noise points.
- 3. For every core point p that has not been assigned to a cluster:
  - Create a new cluster with the point p and all the points that are density-reachable from p
- 4. For border points belonging to more than 1 cluster, assign it to the cluster of the closest core point.

The distance measure could be Euclidean or others.

Some key points

- DBSCAN can find non-linearly separable clusters. (an advantage over K-means and GMM)
- Resistant to noise
- Not entirely deterministic: border points that are reachable from more than one cluster can be part of either cluster, depending on the implementation.