Happy Wednesday!

- **Assignment 2 is out, due on Oct 5th 11:59pm (midnight)**
- **Fourth round of project seminars**, available Thursday, Sep 17th
- **Open office hours on Thursday, 7pm to 8pm**
  - https://primetime.bluejeans.com/a2m/live-event/qfsqxjec
- **Quiz 4, Friday, Sep 18th 6am until Sep 19th 11:59am (noon)**
  - Gaussian mixture models, hierarchical clustering, density based clustering

Coming up soon

- **Assignment 2 Early bird special** → 1 complete programming question by Wed, Sep 23rd
- **Touch-point 1**, survey for in-person version available tonight, deliverables due Sep 28th

CS4641B Machine Learning

Lecture 09: Density-based clustering

Rodrigo Borela ▸ rborelav@gatech.edu

These slides are based on slides from Jing Gao, Chao Zhang, Jiawei Han and Mahdi Roozbahani

Outline

- Overview
- Basic concepts
- The DBSCAN Algorithm
- Analysis of DBSCAN

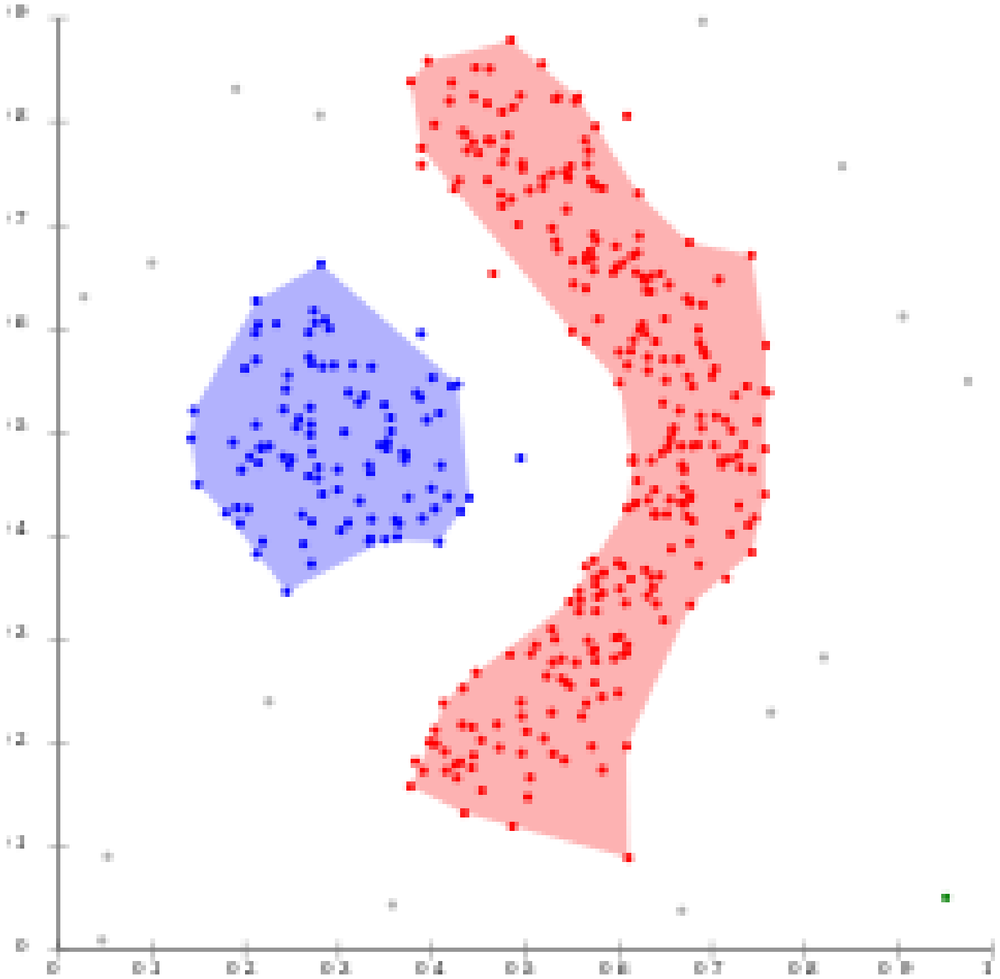
Outline

- Overview
- Basic concepts
- The DBSCAN Algorithm
- Analysis of DBSCAN

Density-based clustering

- **Basic Idea**
  - Clusters are dense regions in the data space, separated by regions of lower density
  - A cluster is defined as a maximal set of density-connected points
  - Detect arbitrarily shaped clusters

- **Method**
  - DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

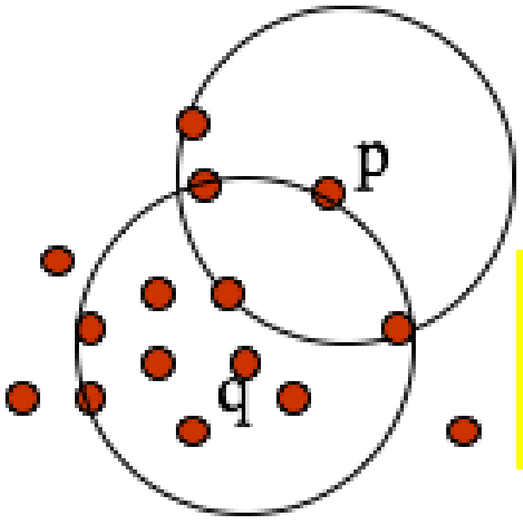


Outline

- Overview
- **Basic concepts**
- The DBSCAN Algorithm
- Analysis of DBSCAN

High Density vs. Low Density

- Two parameters
  - **Eps** (ε): Maximum radius of the neighborhood
  - **MinPts**: Minimum number of points in the **eps**-neighborhood of a point
- High density: ε-Neighborhood of an object contains at least MinPts of objects



MinPts = 5

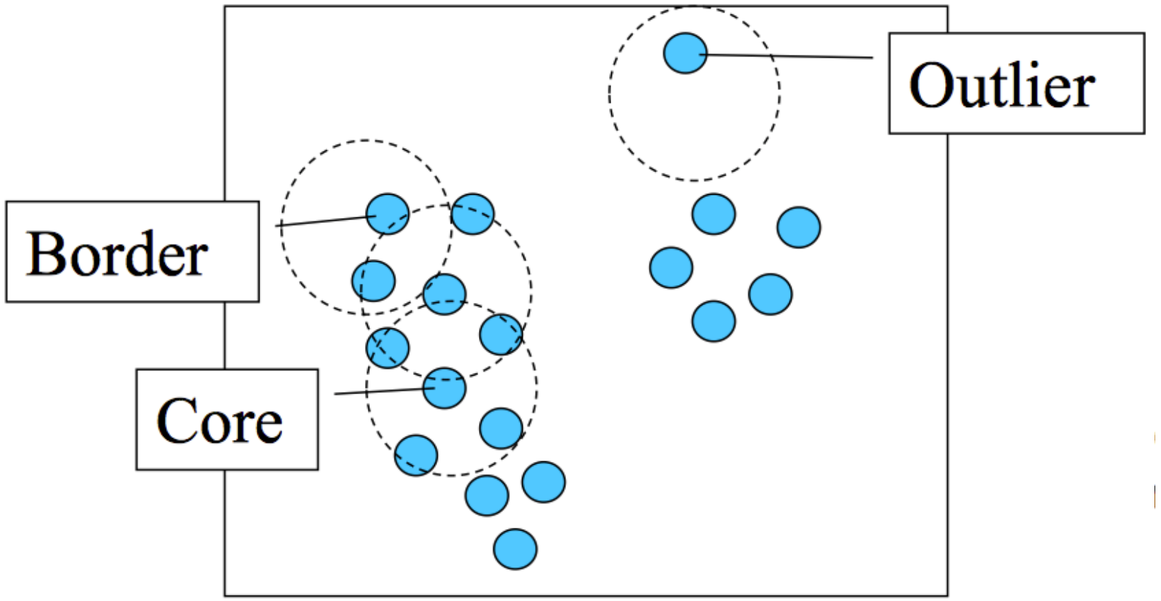
Eps = 1 cm

Density of ***p*** is low

Density of ***q*** is high

Core points, border points and outliers

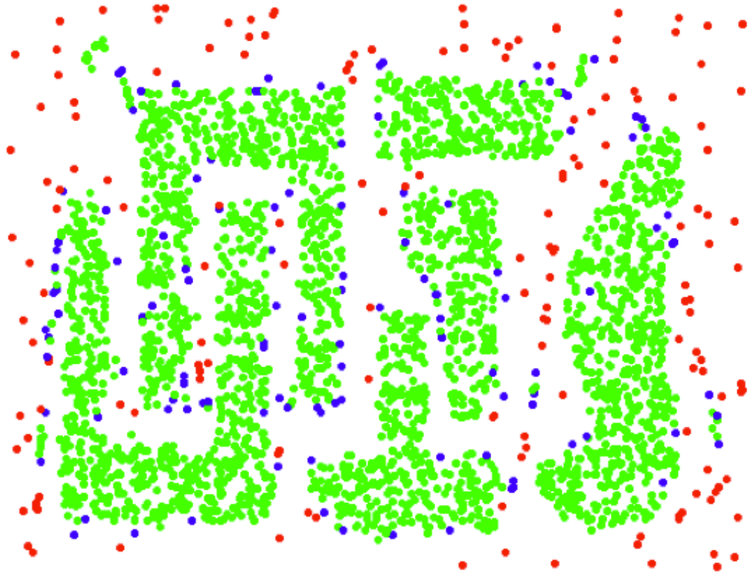
- Given *eps* and *minPts* categorize the objects into three exclusive groups:
- A point is a **core point** if it has more than a specified number of points (*minPts*) within *eps* – these are points that are at the interior of a cluster
- A **border point** has fewer than *minPts* within *eps*, but is in the neighborhood of a core point
- A **noise point** is any point that is not a core point nor a border point



$\varepsilon = 1\text{unit}, \text{MinPts} = 5$

Examples

Original Points

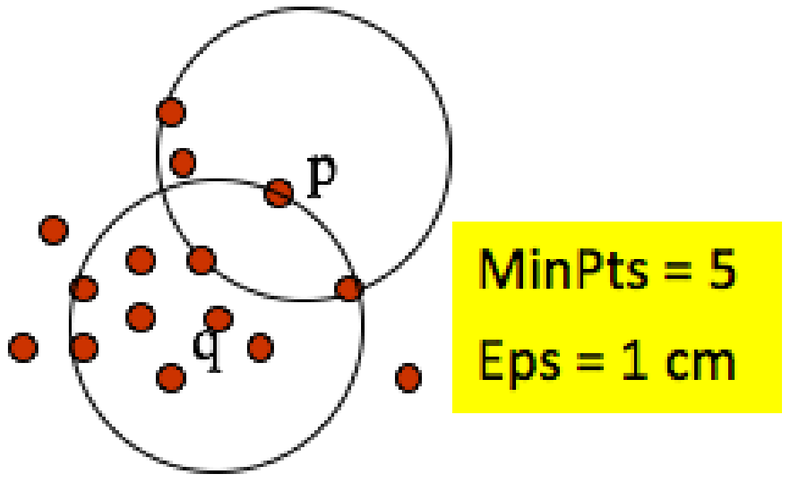


Point types: core, border and outliers

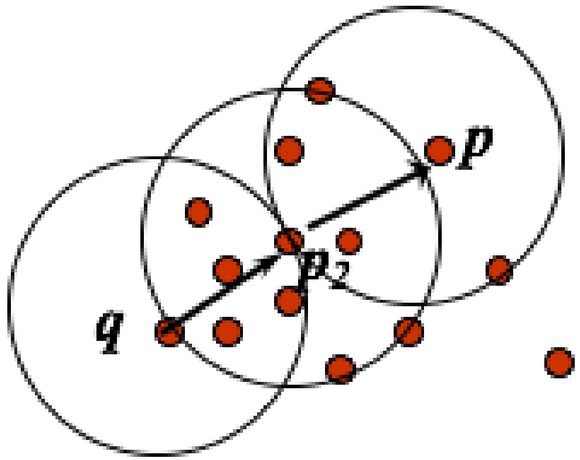
$\varepsilon = 10$, MinPts = 4

Density-based related points

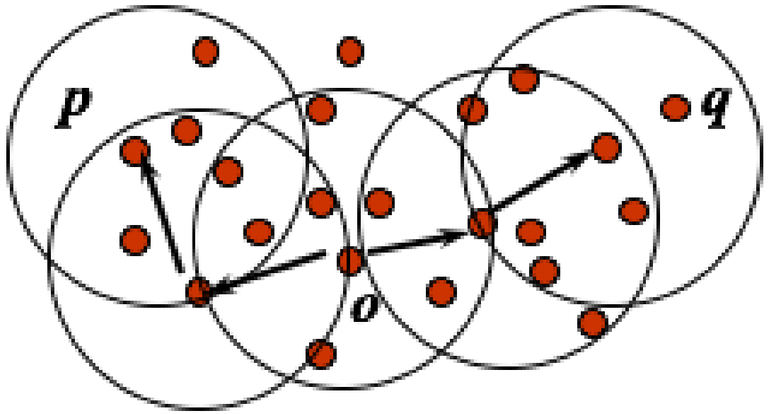
- **Direct density reachability**:
  - An object **p** is directly density-reachable from object **q** if:
    1. **q** is a core object
    2. **p** is in **q**'s ε-neighborhood



Directly density-reachable



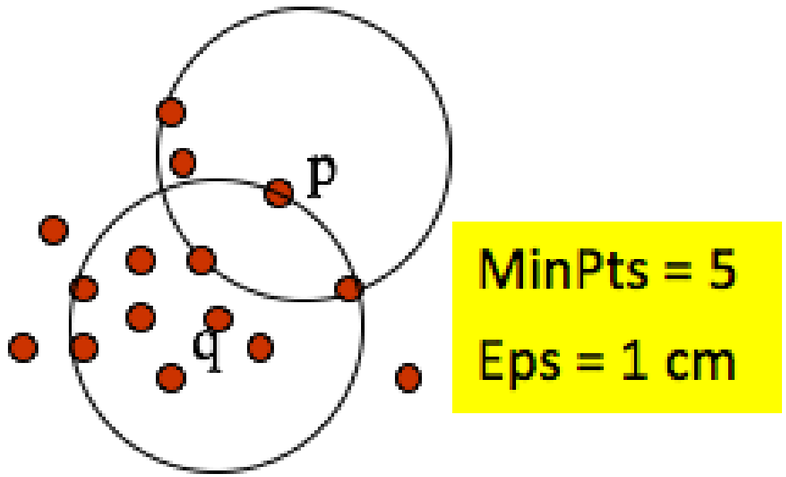
Density-Reachable



Density-Connected

Density-based related points

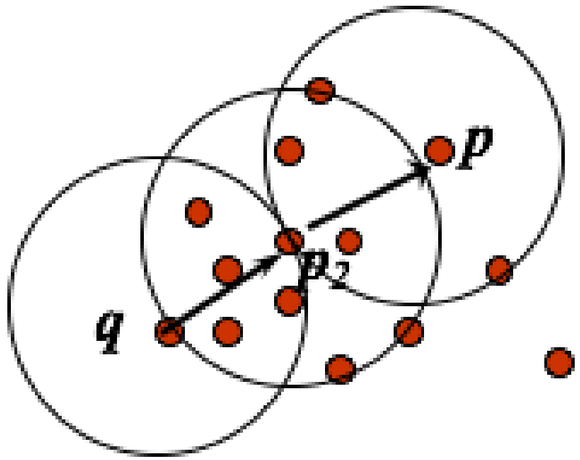
- Density reachability:
  - A point $\mathbf{p}$ is density-reachable from a point $\mathbf{q}$ if there is a chain of points $\mathbf{p_1}, \ldots, \mathbf{p_n}$ $\mathbf{p_1} = \mathbf{q}, \mathbf{p_n} = \mathbf{p}$ such that $\mathbf{p_{i+1}}$ is directly density-reachable from $\mathbf{p_i}$
  - $\mathbf{p_1} = \mathbf{q} \rightarrow \mathbf{p_2} \rightarrow \ldots \rightarrow \mathbf{p_n} = \mathbf{q}$



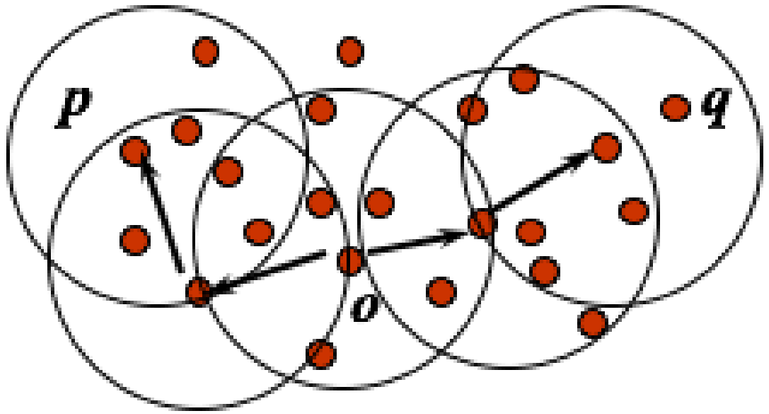
MinPts = 5

Eps = 1 cm

Directly density-reachable



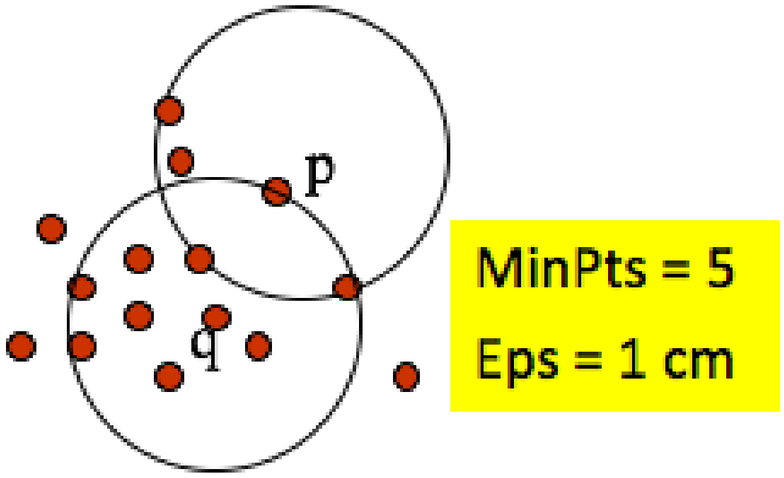
Density-Reachable



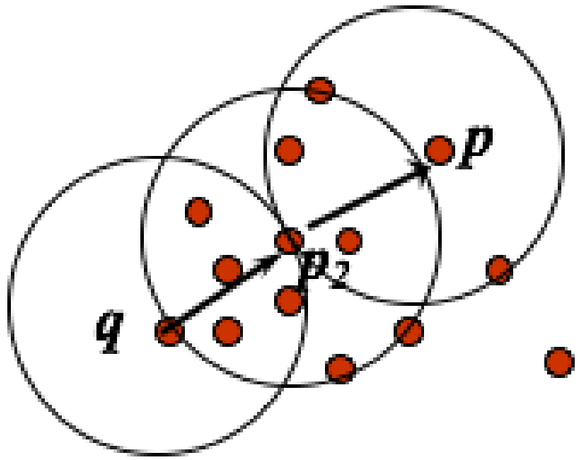
Density-Connected

Density-based related points

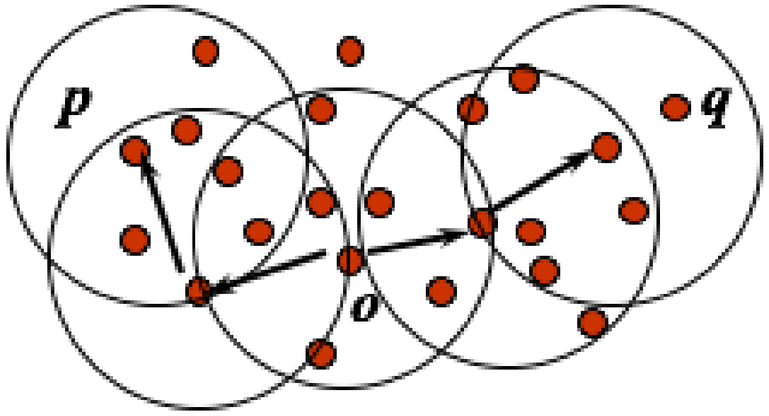
- Density connectivity:
  - A point **p** is density-connected to a point **q** if there is a point **o** such that both **p** and **q** are density-reachable from **o**



Directly density-reachable



Density-Reachable



Density-Connected

Outline

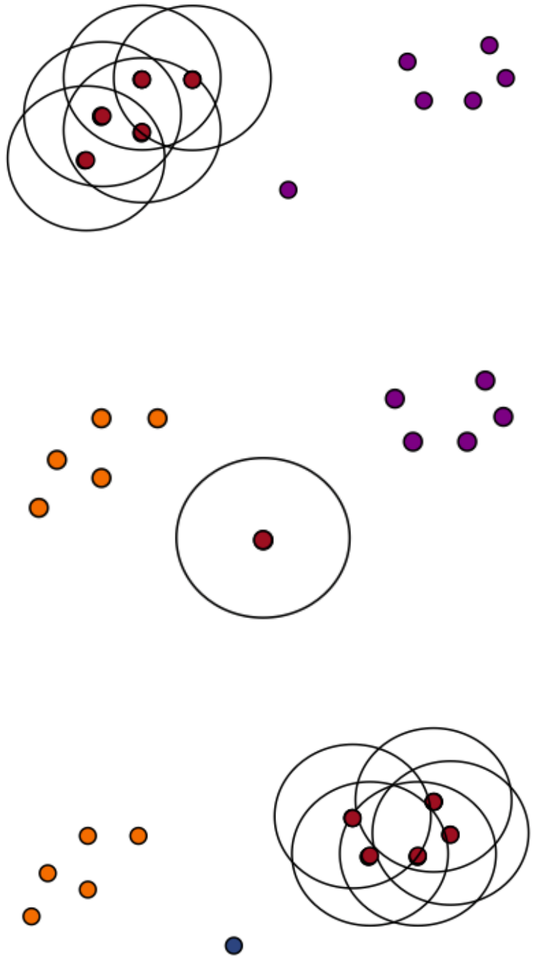
- Overview
- Basic concepts
- **The DBSCAN Algorithm**
- Analysis of DBSCAN

The DBSCAN algorithm

```
DBSCAN(D, eps, MinPts)
C = 0
for each unvisited point P in dataset D
        mark P as visited
        NeighborPts = regionQuery(P, eps)
        if sizeof(NeighborPts) < MinPts
                mark P as NOISE
        else
                C = next cluster
                expandCluster(P, NeighborPts, C, eps, MinPts)

expandCluster(P, NeighborPts, C, eps, MinPts)
        add P to cluster C
        for each point P' in NeighborPts
                if P' is not visited
                        mark P' as visited
                        NeighborPts' = regionQuery(P', eps)
                        if sizeof(NeighborPts') >= MinPts
                                NeighborPts = NeighborPts joined with NeighborPts‘
                if P' is not yet member of any cluster
                        add P' to cluster C

regionQuery(P, eps) return all points within P's eps-neighborhood (including P)
```



https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Outline

- Overview
- Basic concepts
- The DBSCAN Algorithm
- Analysis of DBSCAN

DBSCAN is sensitive to parameters

Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.

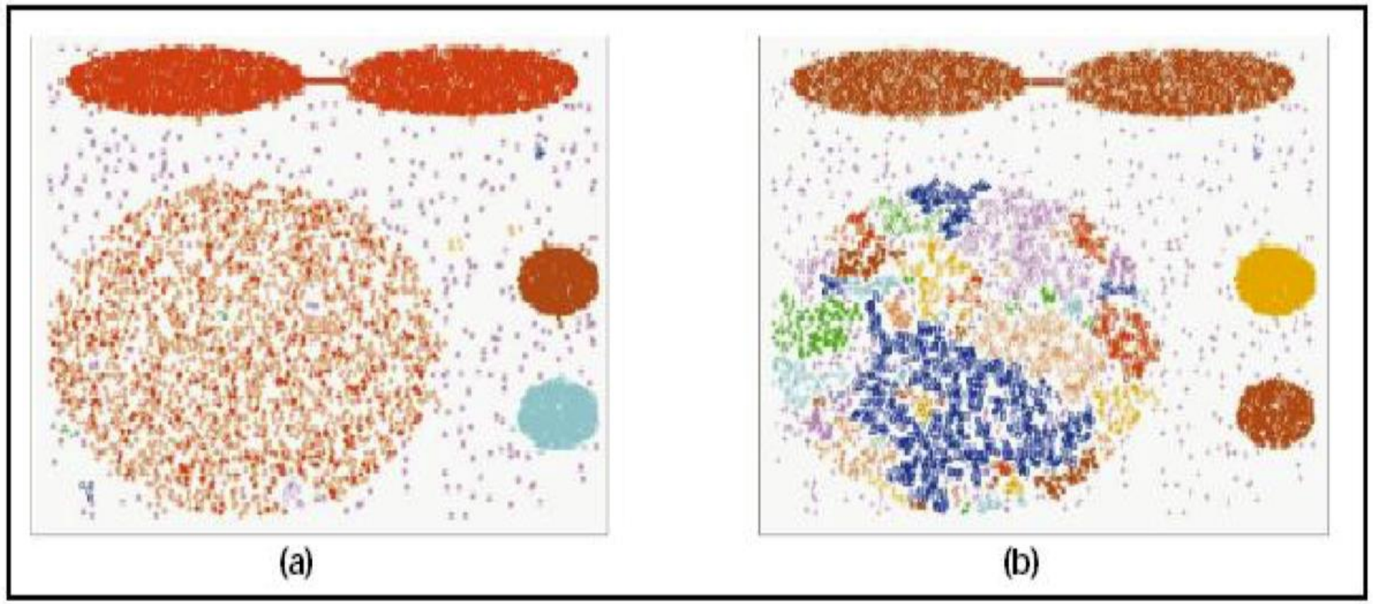


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.

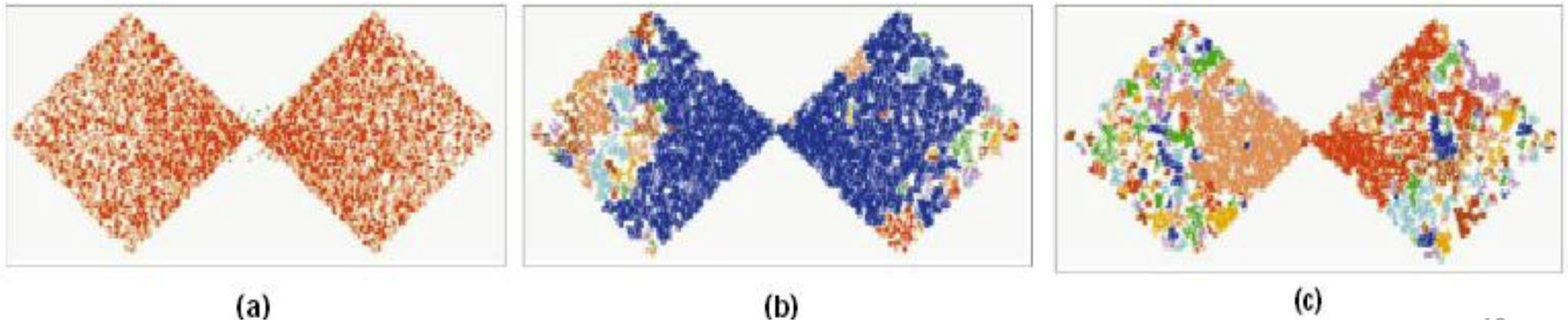


Image credit: George Karypis.

Effect of eps

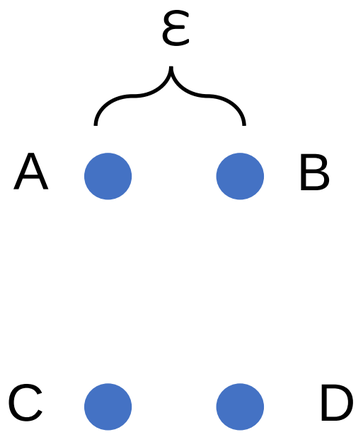
- High value (what will happen?)
  - Clusters will merge and the majority of data points will be in the same cluster
- Low value (what will happen?)
  - A large part of data won't be clustered and considered as outliers. Because, they won't satisfy the number of pints to create a dense region

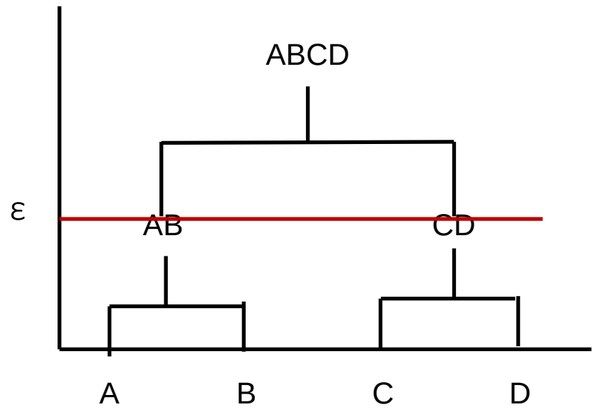
Do we need to define the number of clusters in DBSCAN?

Effect of minimum number of points *minPts*

- **minPts = 1?** Every point will be a cluster on its own, Why? Don't forget, in DBSCAN, a core point is counted as the number of neighboring points

- **minPts = 2?**





Dendogram cut at height ε

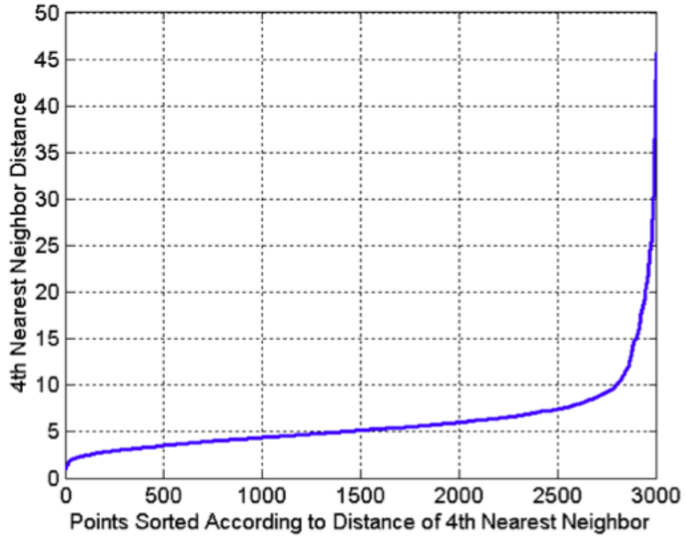
- So, **minPts** should be at least 3

Rule of thumb, minPts >= D+1;

For noisy data => minPts = 2*D (yield more significant clusters)

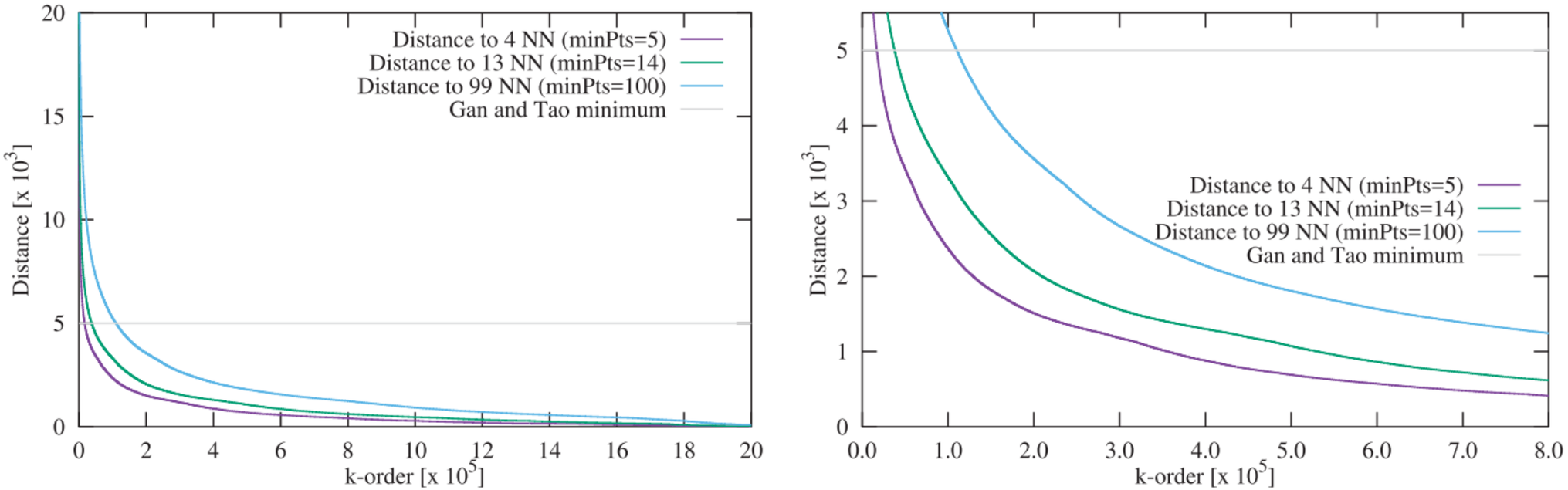
How about eps? (Elbow effect)

- Idea is that for points in a cluster, their k^{th} nearest neighbors are at roughly the same distance
- Noise points have the k^{th} nearest neighbors at farther distance
- So, plot sorted distance of every point to its k^{th} nearest



Here we have 3000 points and x-axis shows just a point index.
Point indices are sorted in ascending order based on their 4^{th} nearest neighbor distance

Elbow effect another example



(a) k-distance plots

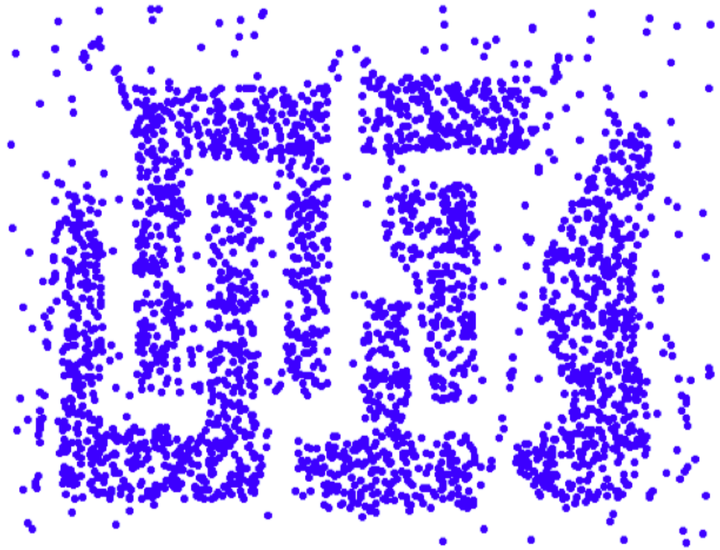
(b) k-distance plots (magnified region)

minPts often does not have a significant impact on the clustering results

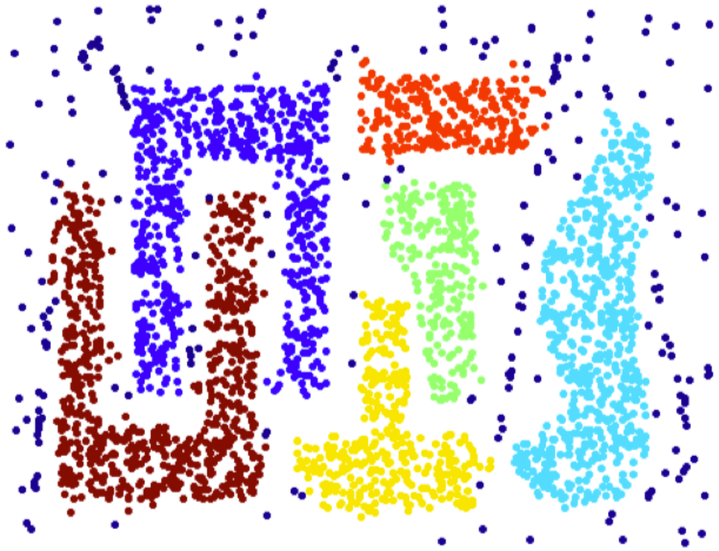
Erich Schuber et al

When DBSCAN works well

- Robust to noise
- Can detect arbitrarily-shaped clusters



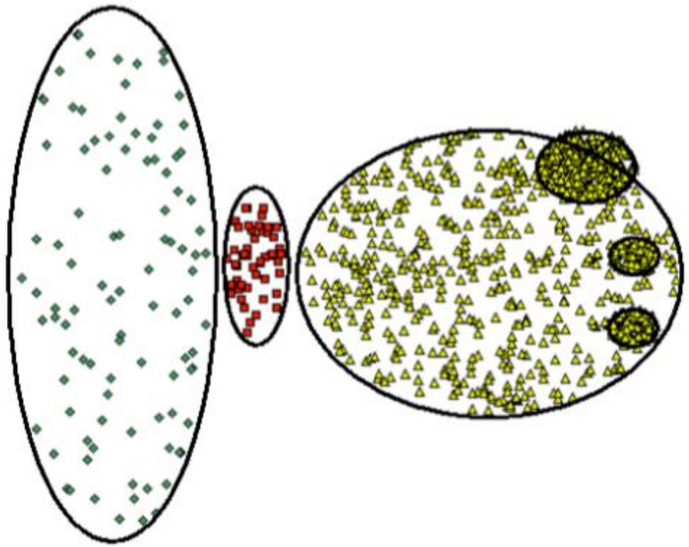
Original Points



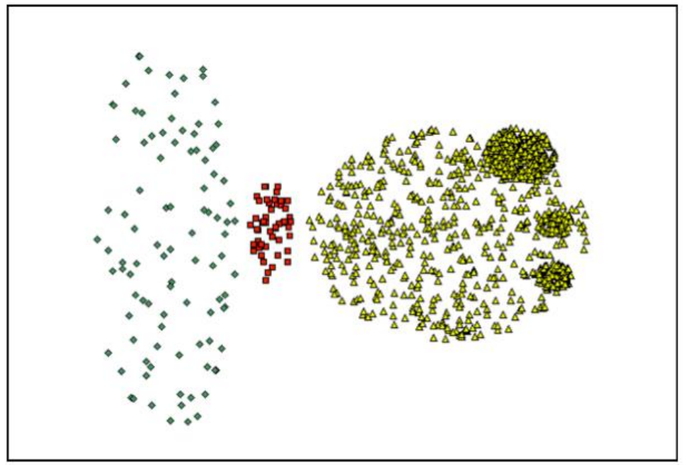
Clusters

When DBSCAN does NOT work well

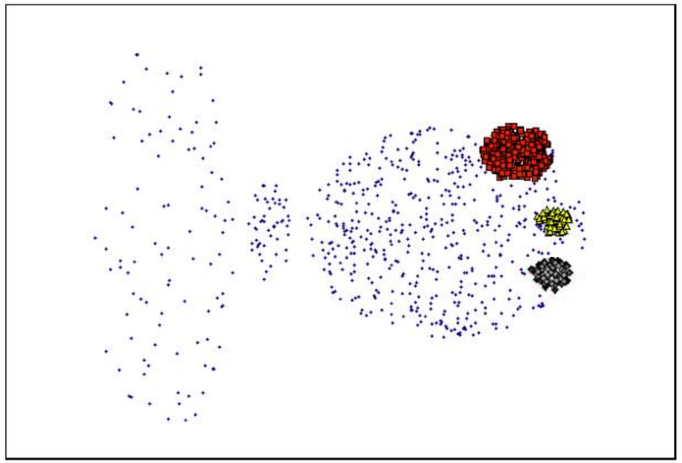
- Cannot handle varying densities
- Sensitive to parameters—hard to determine the best setting of parameters



Original Points



(MinPts=4, Eps=9.92).



(MinPts=4, Eps=9.75)