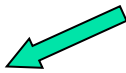


Unit IV : Clustering

Cluster Analysis – Partitioning Methods – Hierarchical
Methods – **Density Based Methods** – Grid Based
Methods – Outlier Analysis

Cluster Analysis

1. What is Cluster Analysis?
2. Types of Data in Cluster Analysis
3. A Categorization of Major Clustering Methods
4. Partitioning Methods
5. Hierarchical Methods
6. **Density-Based Methods** 
7. Grid-Based Methods
8. Outlier Analysis



Density-Based Clustering Methods

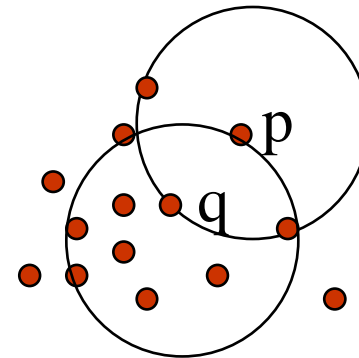
- Clustering based on density (local cluster criterion), such as density-connected points
- Major features:
 - Discover clusters of arbitrary shape
 - Handle noise
 - One scan
 - Need density parameters as termination condition
- Several interesting studies:
 - DBSCAN: Ester, et al. (KDD'96)
 - OPTICS: Ankerst, et al (SIGMOD'99).
 - DENCLUE: Hinneburg & D. Keim (KDD'98)
 - CLIQUE: Agrawal, et al. (SIGMOD'98) (more grid-based)

Density-Based Clustering: Basic Concepts

- Two parameters:
 - **Eps**: Maximum radius of the neighbourhood
 - **MinPts**: Minimum number of points in an Eps-neighbourhood of that point
- $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid \text{dist}(p,q) \leq Eps\}$
- **Directly density-reachable**: A point p is directly density-reachable from a point q w.r.t. Eps , $MinPts$ if

- p belongs to $N_{Eps}(q)$
- core point condition:

$$|N_{Eps}(q)| \geq MinPts$$



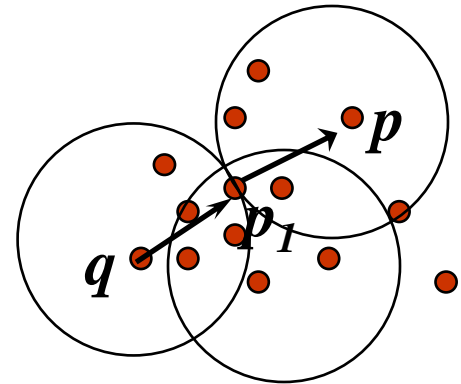
MinPts = 5

Eps = 1 cm

Density-Reachable and Density-Connected

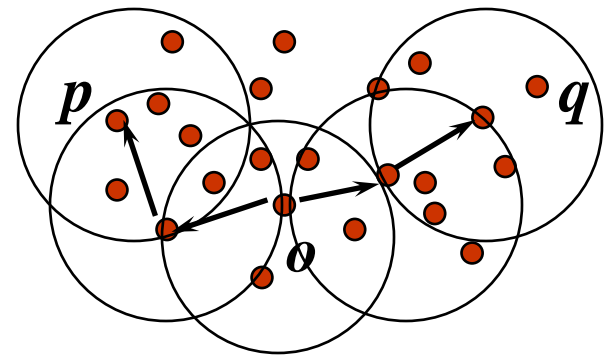
- Density-reachable:

- A point p is **density-reachable** from a point q w.r.t. $Eps, MinPts$ if there is a chain of points $p_1, \dots, p_n, p_1 = q, p_n = p$ such that p_{i+1} is directly density-reachable from p_i



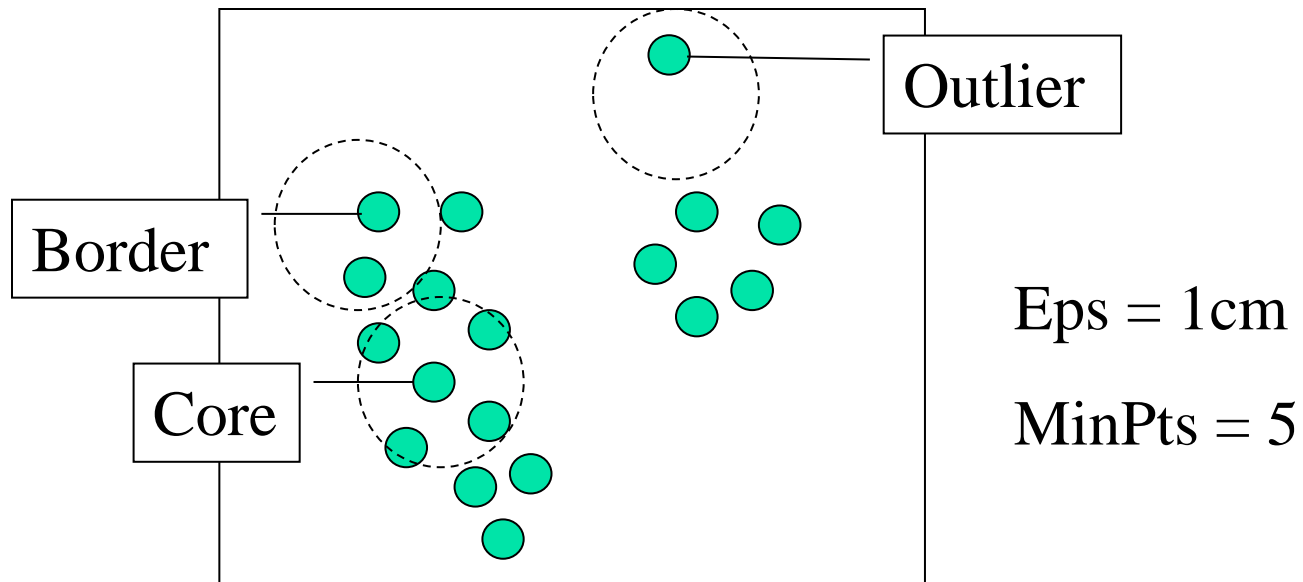
- Density-connected

- A point p is **density-connected** to a point q w.r.t. $Eps, MinPts$ if there is a point o such that both, p and q are density-reachable from o w.r.t. Eps and $MinPts$



DBSCAN: Density Based Spatial Clustering of Applications with Noise

- Relies on a *density-based* notion of cluster: A *cluster* is defined as a maximal set of density-connected points
- Discovers clusters of arbitrary shape in spatial databases with noise





DBSCAN: The Algorithm

- Arbitrary select a point p
- Retrieve all points density-reachable from p w.r.t. Eps and $MinPts$.
- If p is a core point, a cluster is formed.
- If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.
- Continue the process until all of the points have been processed.

DBSCAN: Sensitive to Parameters

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

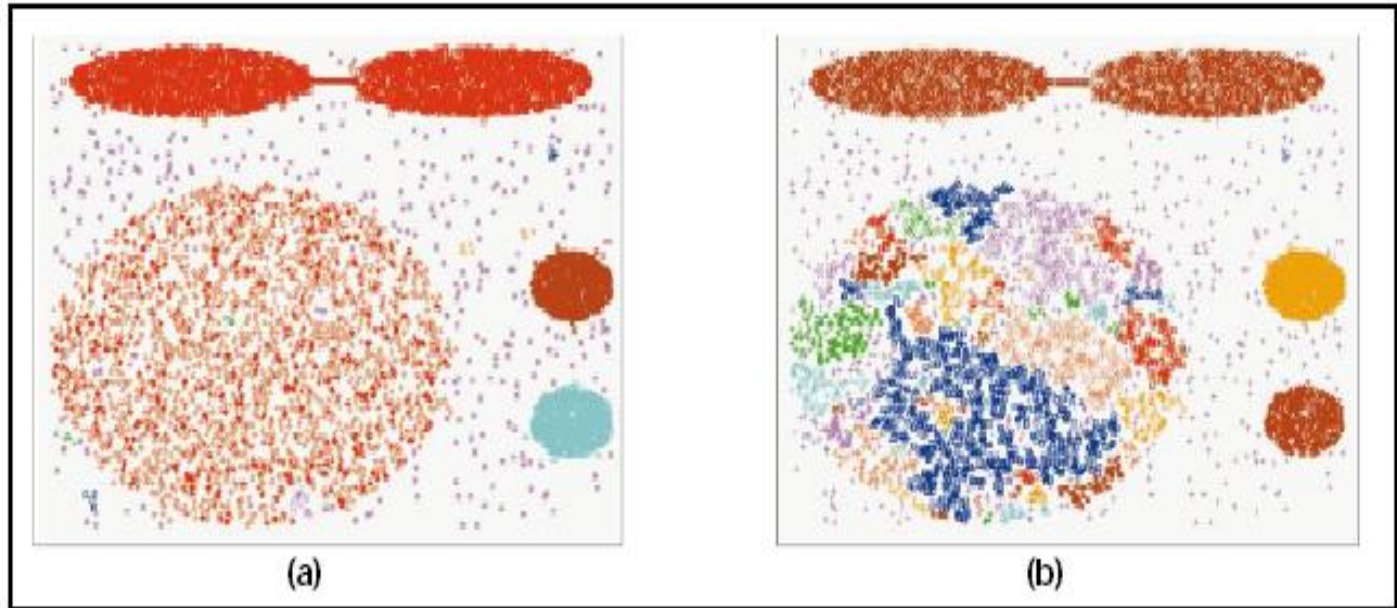
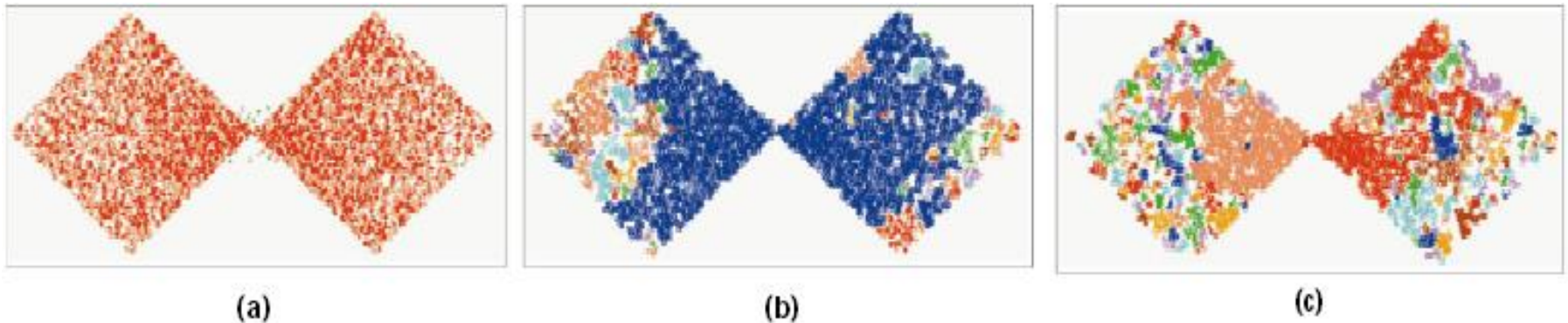
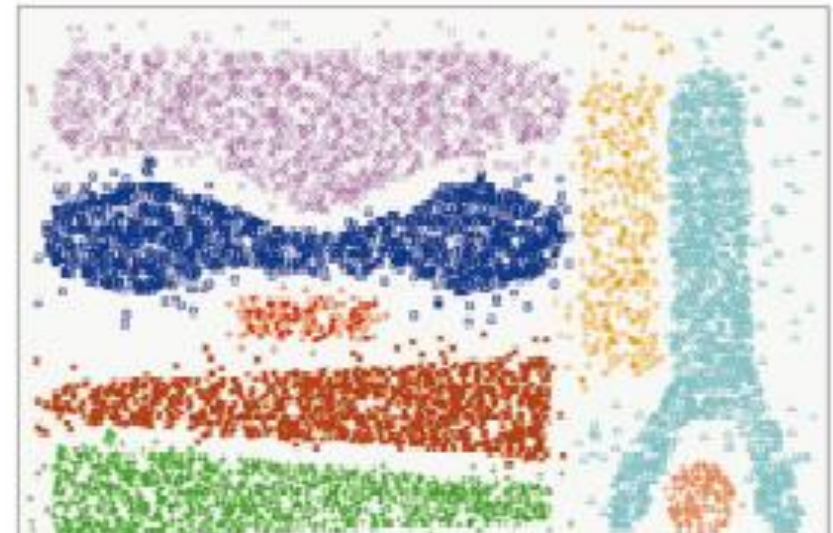
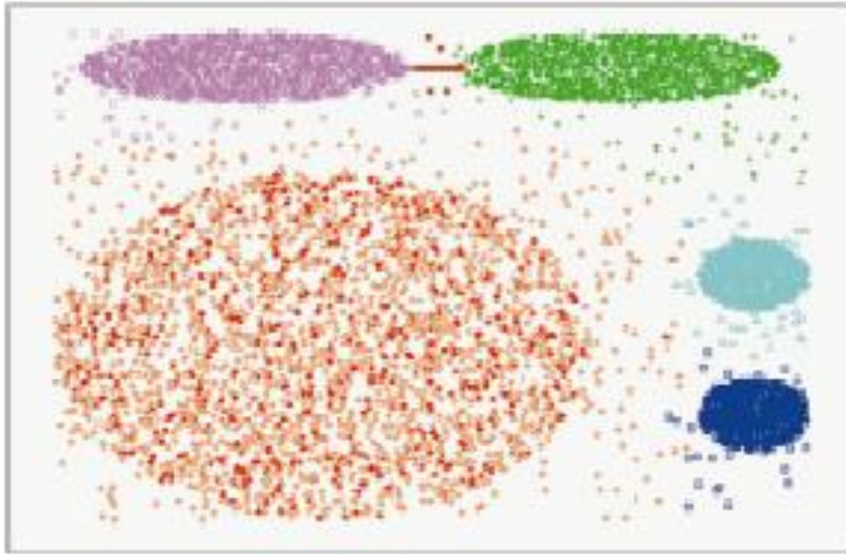


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



CHAMELEON (Clustering Complex Objects)





OPTICS: A Cluster-Ordering Method (1999)

- OPTICS: Ordering Points To Identify the Clustering Structure
 - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
 - Produces a special order of the database wrt its density-based clustering structure
 - This cluster-ordering contains info equiv to the density-based clusterings corresponding to a broad range of parameter settings
 - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
 - Can be represented graphically or using visualization techniques

OPTICS: Some Extension from DBSCAN

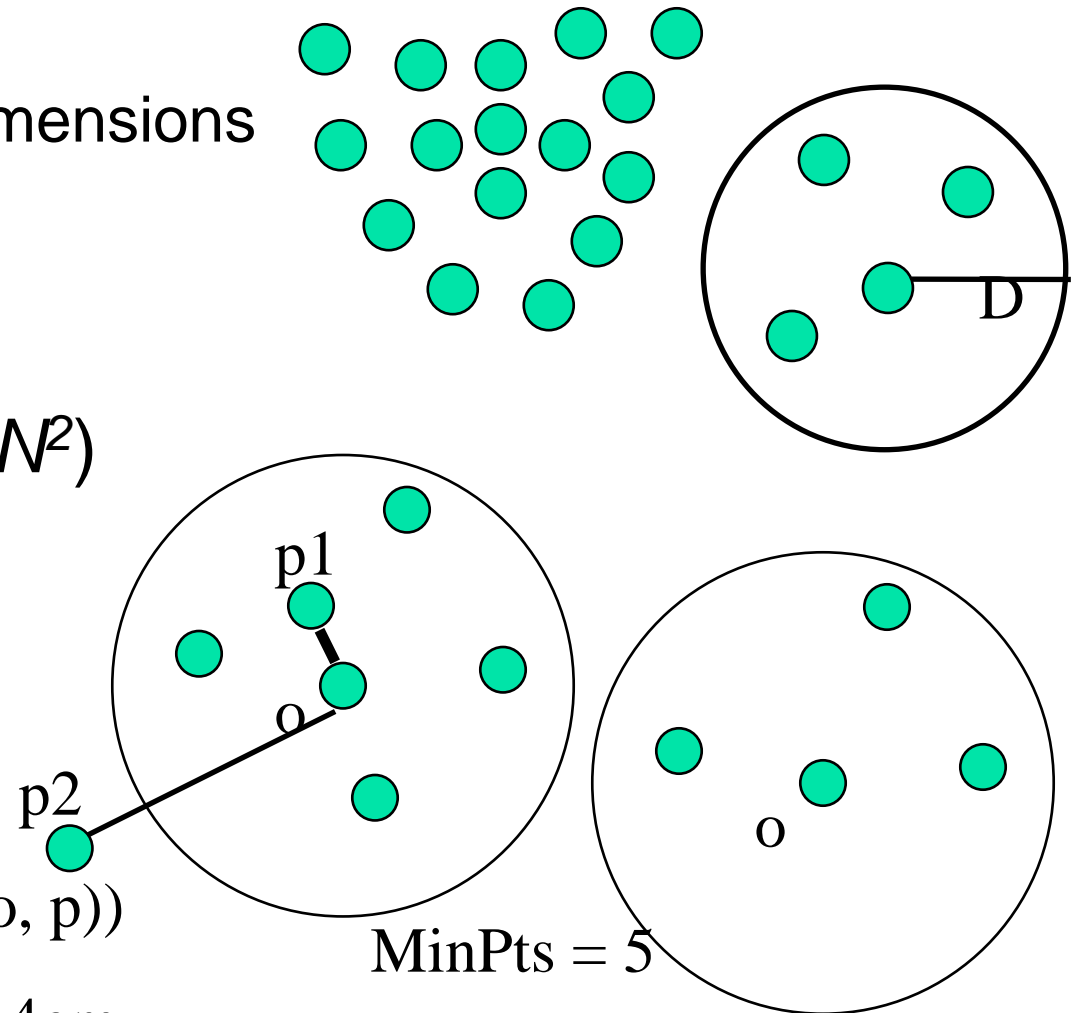
- Index-based:
 - k = number of dimensions
 - $N = 20$
 - $p = 75\%$
 - $M = N(1-p) = 5$
- Complexity: $O(kN^2)$

- Core Distance

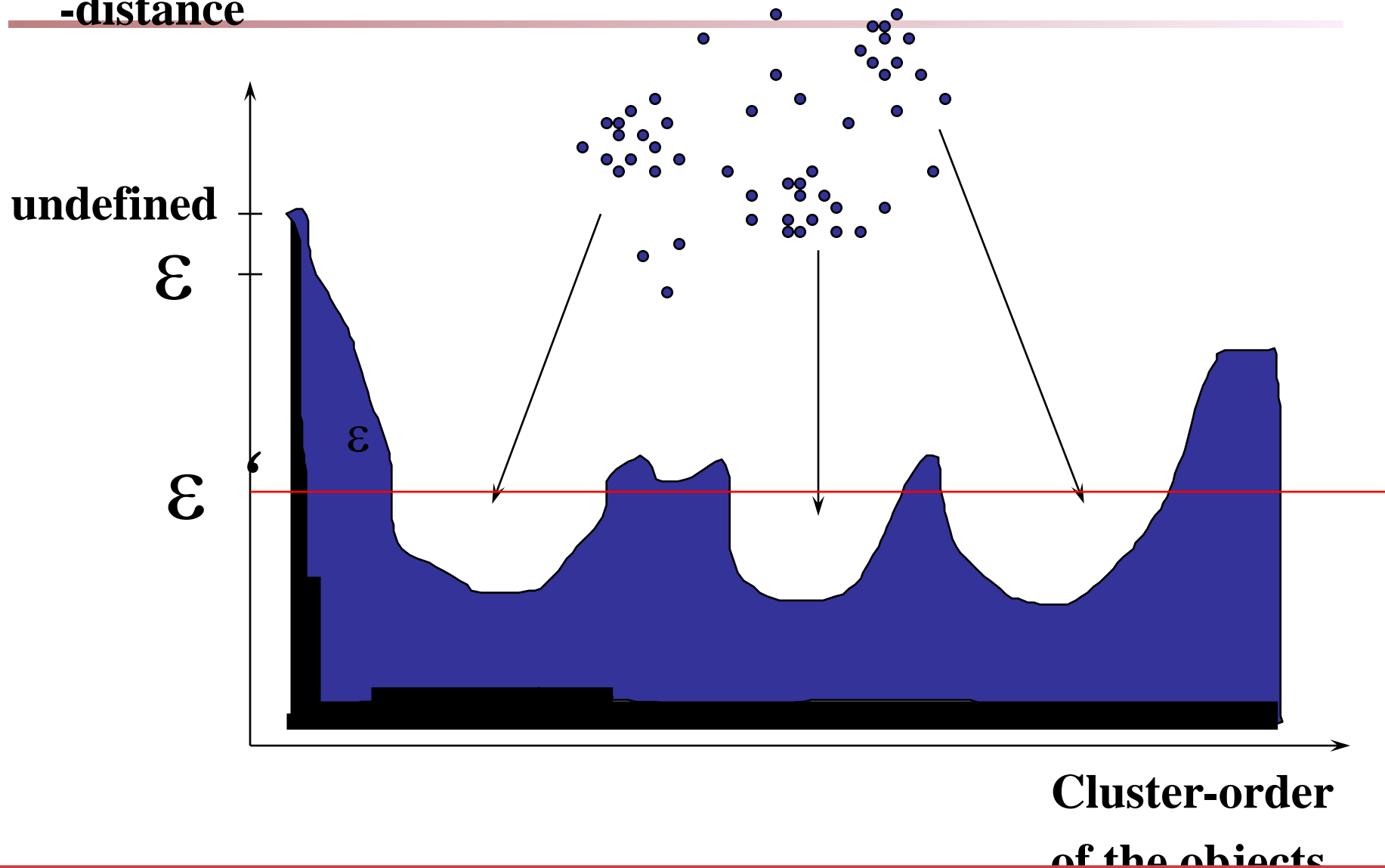
- Reachability Distance

$\text{Max}(\text{core-distance}(o), d(o, p))$

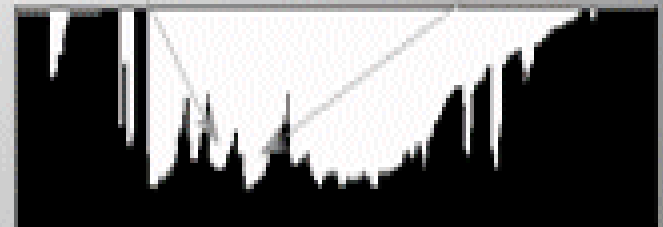
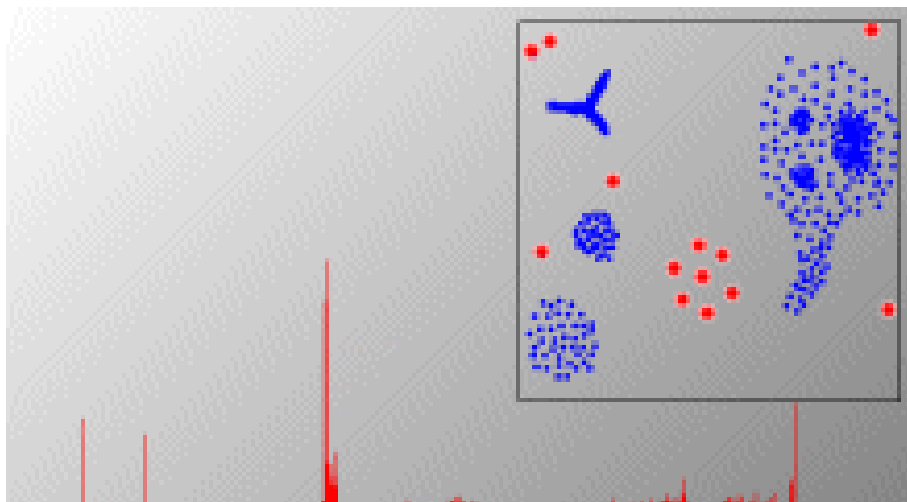
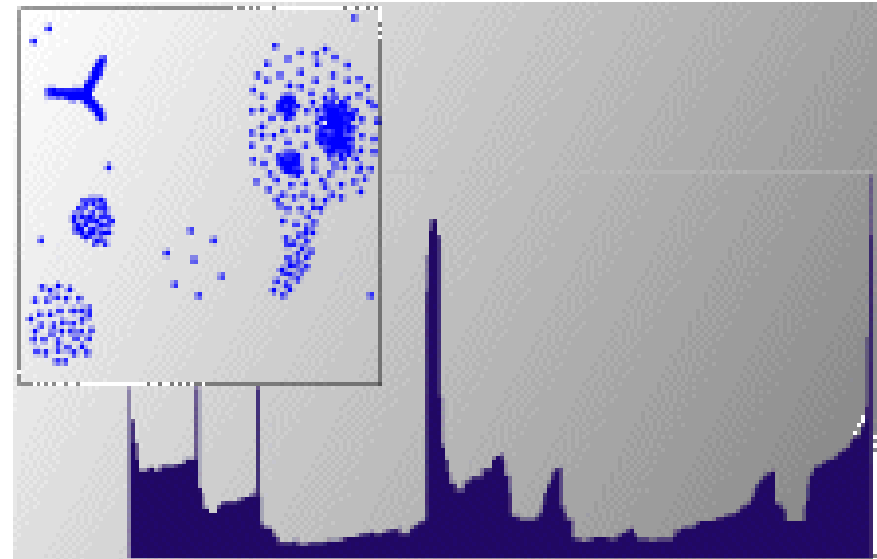
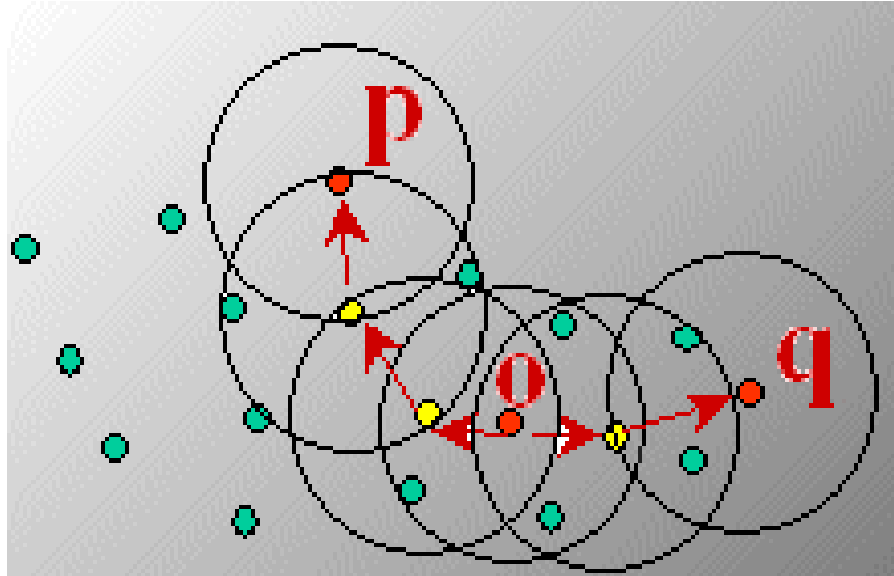
$r(p1, o) = 2.8\text{cm}$. $r(p2, o) = 4\text{cm}$



Reachability -distance



Density-Based Clustering: OPTICS & Its Applications



DENCLUE: Using Statistical Density Functions

- DENSity-based CLUstEring by Hinneburg & Keim (KDD'98)

- Using statistical density functions: $f_{Gaussian}(x, y) = e^{-\frac{d(x,y)^2}{2\sigma^2}}$

$$f_{Gaussian}^D(x) = \sum_{i=1}^N e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

- Major features

$$\nabla f_{Gaussian}^D(x, x_i) = \sum_{i=1}^N (x_i - x) \cdot e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

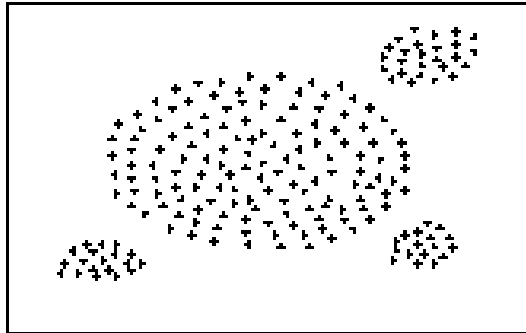
- Solid mathematical foundation
- Good for data sets with large amounts of noise
- Allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets
- Significant faster than existing algorithm (e.g., DBSCAN)
- But needs a large number of parameters



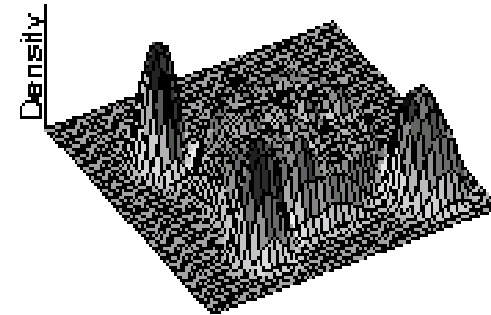
Dendclue: Technical Essence

- Uses grid cells but only keeps information about grid cells that do actually contain data points and manages these cells in a tree-based access structure
- Influence function: describes the impact of a data point within its neighborhood
- Overall density of the data space can be calculated as the sum of the influence function of all data points
- Clusters can be determined mathematically by identifying density attractors
- Density attractors are local maximal of the overall density function

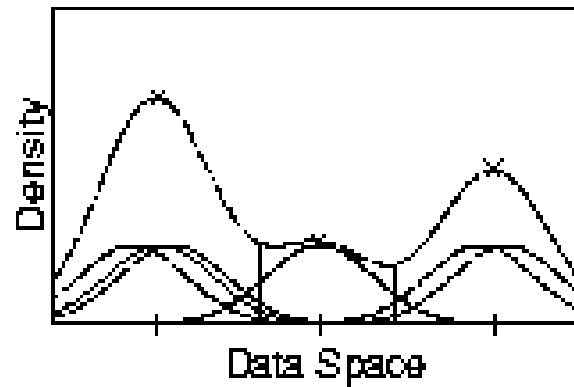
Density Attractor



(a) Data Set



(c) Gaussian



Center-Defined and Arbitrary

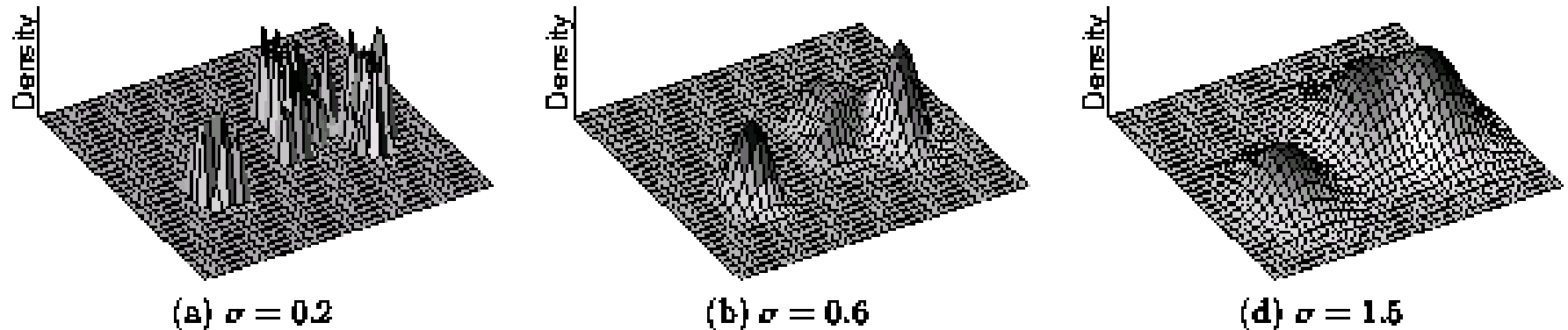


Figure 3: Example of Center-Defined Clusters for different σ

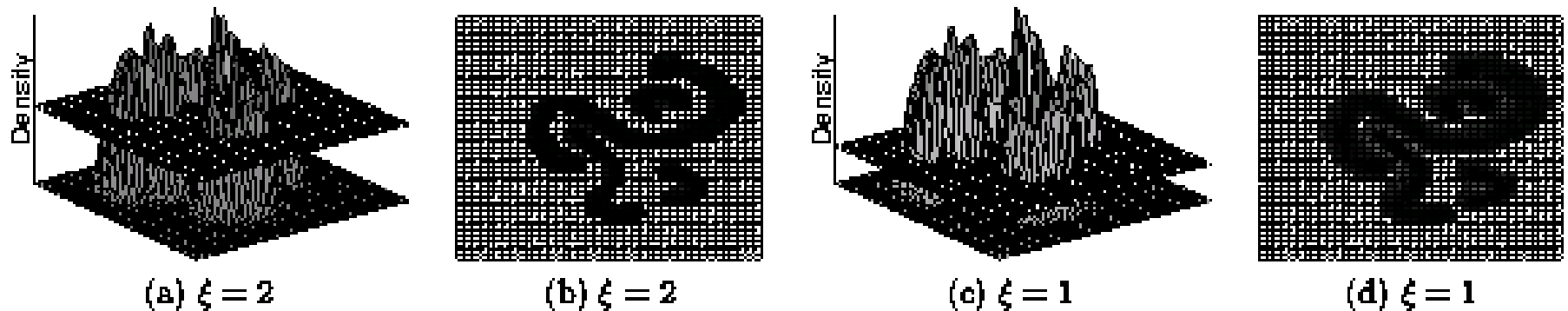


Figure 4: Example of Arbitrary-Shape Clusters for different ξ