Data Mining:

Concepts and Techniques

— Chapter 7 —

Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
www.cs.uiuc.edu/~hanj

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

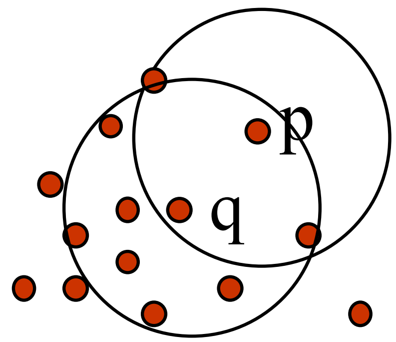
Density-Based Clustering Methods

- 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. (1996)
  - OPTICS: Ankerst, et al (1999).
  - DENCLUE: Hinneburg & D. Keim (1998)

Density-Based Clustering: Basic Concepts

- Two parameters:
  - *Eps*: Maximum radius of the neighborhood
  - *MinPts*: Minimum number of points in an Eps-neighborhood of that point
- $N_{Eps}(p)$: $\{q \text{ belongs to } D \mid d(p,q) <= Eps\}$
- 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:

    $|N_{Eps}(q)| >= MinPts$

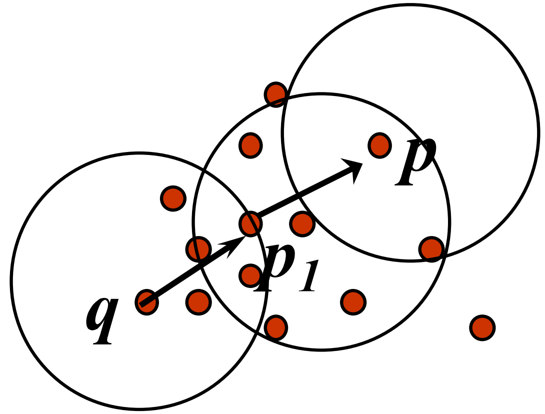


MinPts = 5

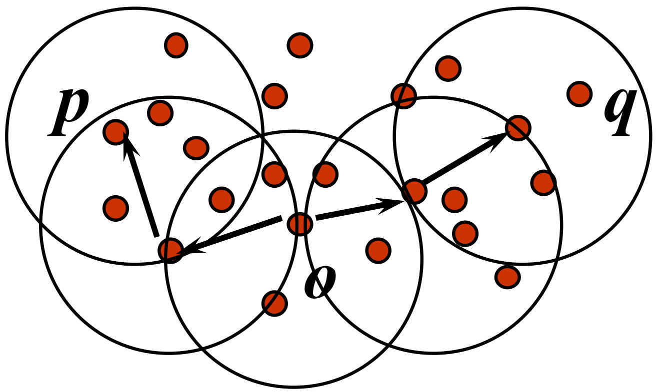
Eps = 1 cm

Density-Based Clustering: Basic Concepts

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



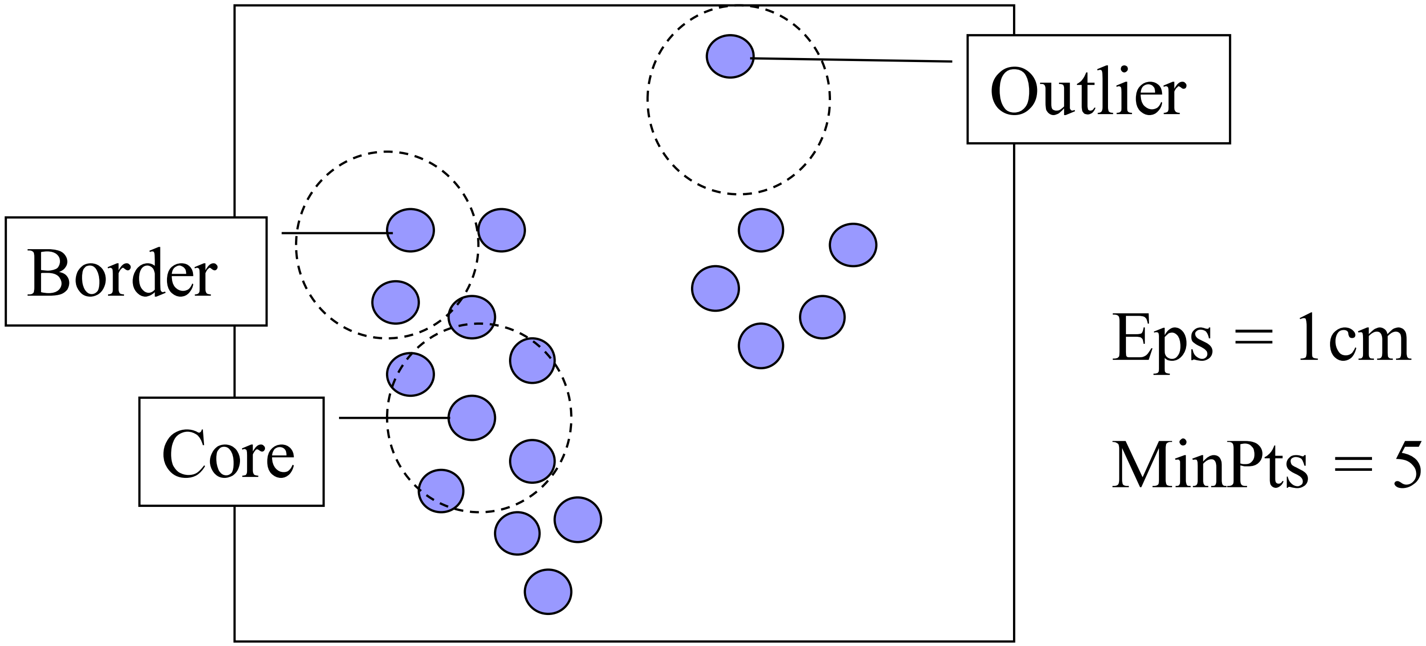
- Density-connected
  - A point p is density-connected to a point q w.r.t. *Eps*, *MinPts* if there is a point o such that both, p and q are density-reachable from o w.r.t. *Eps* and *MinPts*



Explanation on whiteboard

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 containing p and all the density-reachable points from *p*. Mark these points as processed.
- Mark *p* as processed.
- Continue this 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.

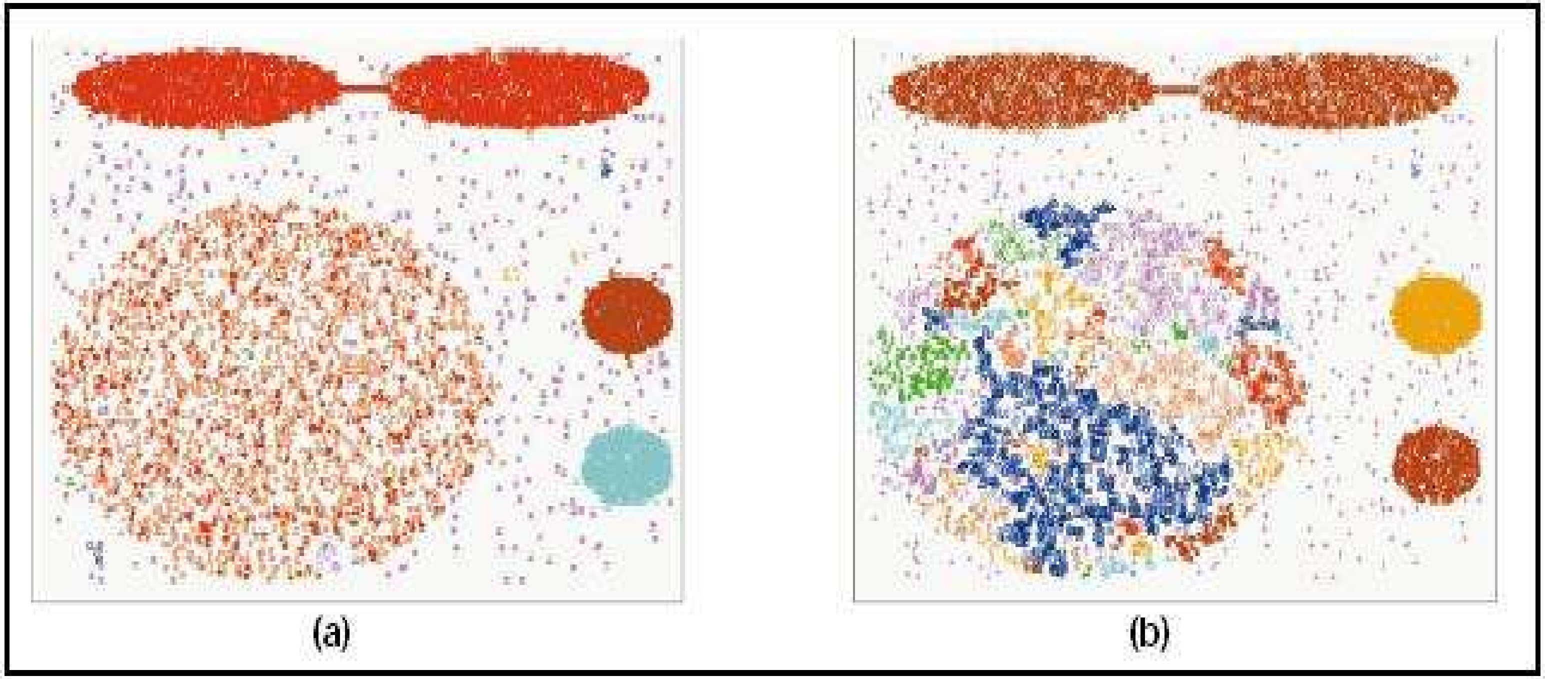
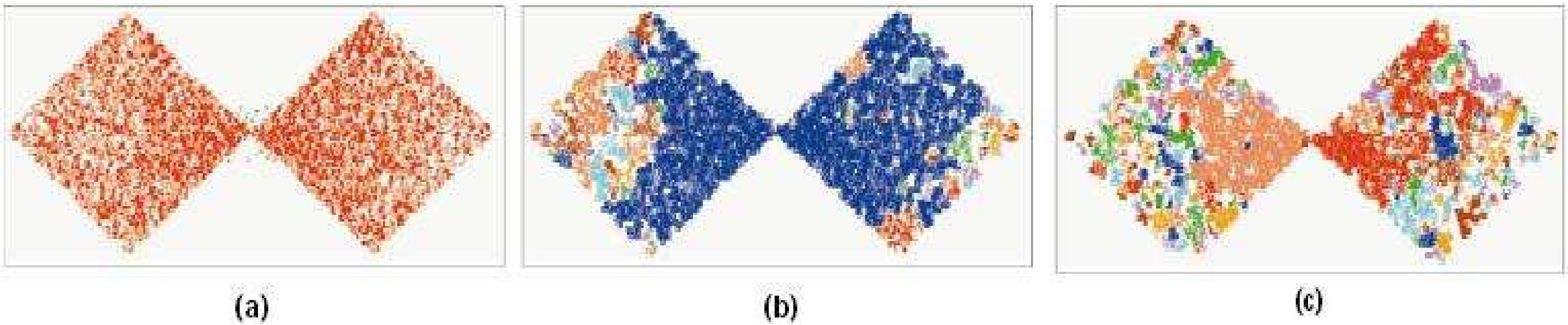


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.

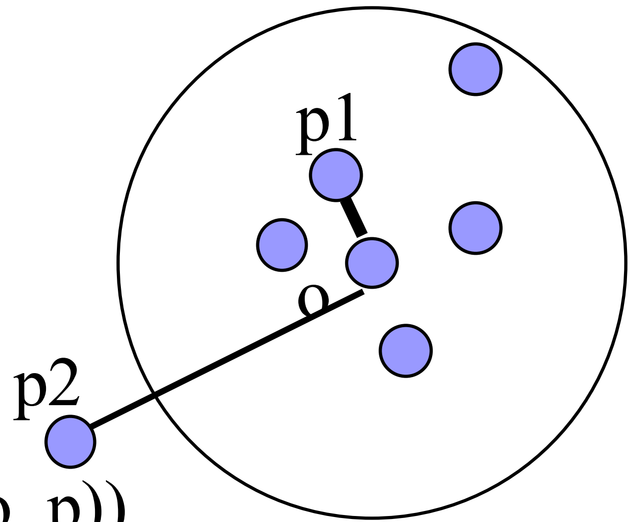


OPTICS: A Cluster-Ordering Method

- OPTICS: Ordering Points To Identify the Clustering Structure
  - Ankerst, Breunig, Kriegel, and Sander (1999)
  - Produces a special order of the database w.r.t. its density-based clustering structure
  - This cluster-ordering contains info equivalent 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 basic concepts

- Core Distance of p wrt MinPts: smallest distance eps' between p and an object in its eps-neighborhood such that p would be a core object for eps' and MinPts. Otherwise, undefined.
- Reachability Distance of p wrt o:
  Max (core-distance (o), d (o, p)) if o is core object.
  Undefined otherwise



Max (core-distance (o), d (o, p))

MinPts = 5

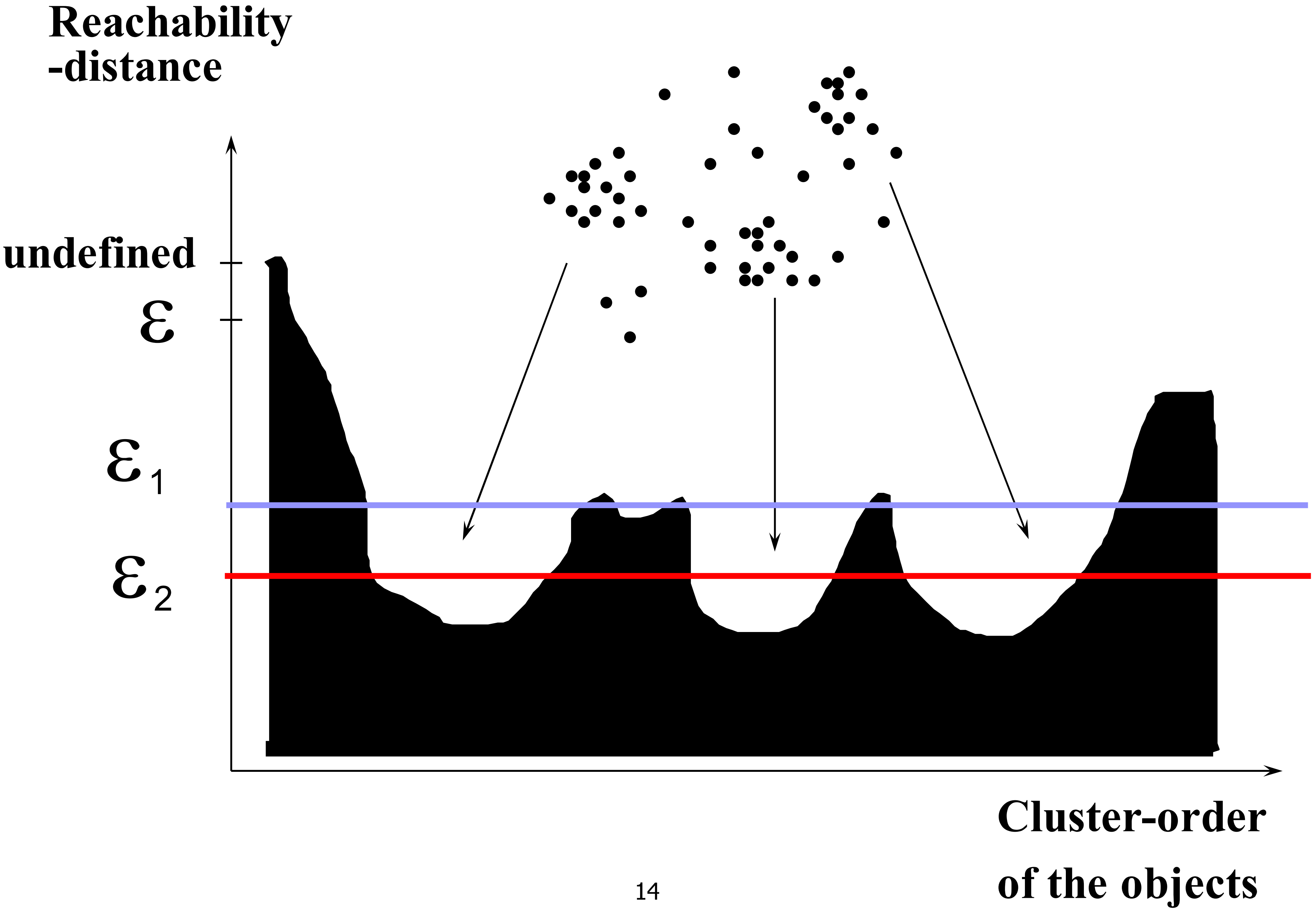
r(p1, o) = 1.5cm. r(p2,o) = 4cm

$\varepsilon = 3$ cm

OPTICS

- (1) Select non-processed object o
- (2) Find neighbors (eps-neighborhood)
- Compute core distance for o
- Write object o to ordered file and mark o as processed
- If o is not a core object, restart at (1)
- (o is a core object …)
- Put neighbors of o in Seedlist and order
  - If neighbor n is not yet in SeedList then add (n, reachability from o) else if reachability from o < current reachability, then update reachability + order SeedList wrt reachability
- Take new object from Seedlist with smallest reachability and restart at (2)

Example on whiteboard


Reachability
-distance
undefined
ε
ε 1
ε 2
Cluster-order
of the objects

DENCLUE: Using Statistical Density Functions

- DENsity-based CLUstEring by Hinneburg & Keim (1998)
- Using statistical density functions
- Major features
  - 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 DBSCAN
  - But needs a large number of parameters

Denclue: 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

$$f_{Gaussian}(x,y) = e^{-\frac{d(x,y)^2}{2\sigma^2}}$$

OBS: minus

- Overall density of the data space can be calculated as the sum of the influence function of all data points

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

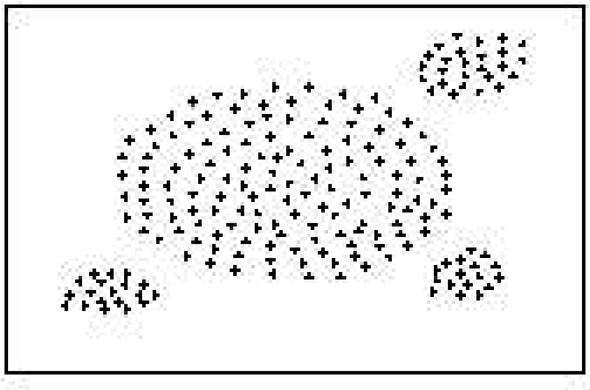
OBS: minus

- Clusters can be determined mathematically by identifying density attractors. Density attractors are local maxima of the overall density function

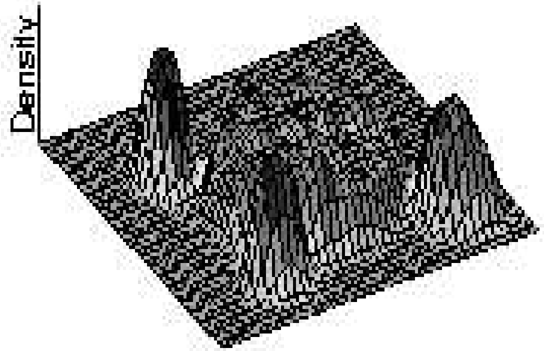
$$\nabla f_{Gaussian}^{D}(x,x_i) = \sum_{i=1}^{N} (x_i - x) \cdot e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

OBS: minus

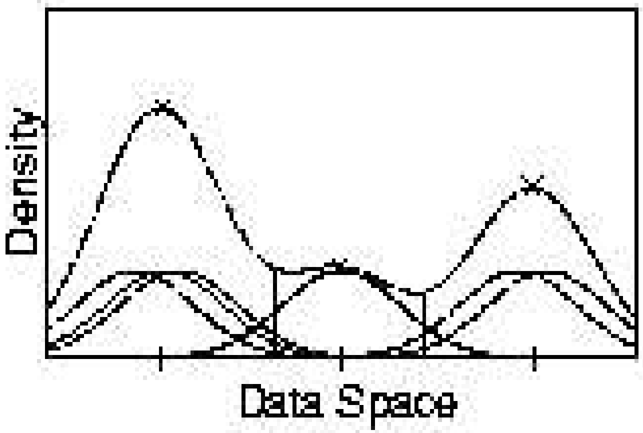
Density Attractor



(a) Data Set



(c) Gaussian

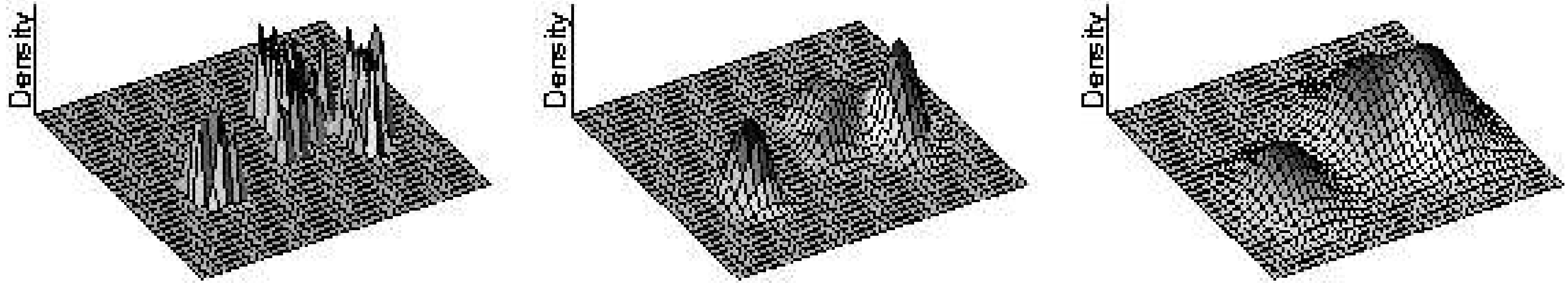


Denclue: Technical Essence

- Significant density attractor for threshold k: density attractor with density larger than or equal to k
- Center-defined cluster for a significant density attractor x for threshold k: points that are density attracted by x
  - Points that are attracted to a density attractor with density less than k are called outliers

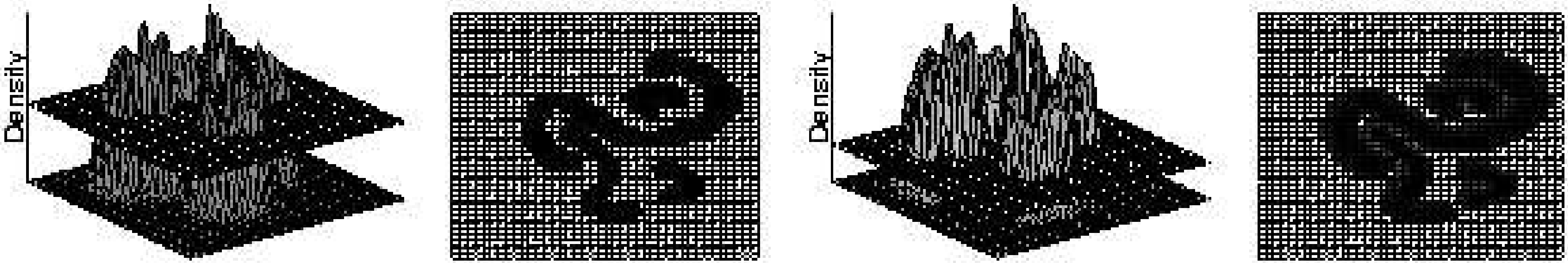
- Set of significant density attractors X for threshold k: for each pair of density attractors x1, x2 in X there is a path from x1 to x2 such that each point on the path has density larger than or equal to k
- Arbitrary-shape cluster for a set of significant density attractors X for threshold k: points that are density attracted to some density attractor in X

Center-Defined and Arbitrary-shape clusters



(a) $\sigma = 0.2$ (b) $\sigma = 0.6$ (d) $\sigma = 1.5$

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



(a) $\xi = 2$ (b) $\xi = 2$ (c) $\xi = 1$ (d) $\xi = 1$

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