



Lecture 5. Density-Based and Grid-Based Clustering Methods

Lecture 5. Density-Based and Grid-Based Clustering Methods

□ Density-Based Clustering

- Basic Concepts
- DBSCAN: A Density-Based Clustering Algorithm
- OPTICS: Ordering Points To Identify Clustering Structure

□ Grid-Based Clustering Methods

- Basic Concepts
- STING: A Statistical Information Grid Approach
- CLIQUE: Grid-Based Subspace Clustering



Session 1: Basic Concepts of Density-Based Clustering

Density-Based Clustering Methods

- ❑ Clustering based on density (a local cluster criterion), such as density-connected points
- ❑ Major features:
 - ❑ Discover clusters of arbitrary shape
 - ❑ Handle noise
 - ❑ One scan (only examine the local region to justify density)
 - ❑ Need density parameters as termination condition
- ❑ Several interesting studies:
 - ❑ DBSCAN: Ester, et al. (KDD'96) To be covered in this lecture
 - ❑ OPTICS: Ankerst, et al (SIGMOD'99) To be covered in this lecture
 - ❑ DENCLUE: Hinneburg & D. Keim (KDD'98)
 - ❑ CLIQUE: Agrawal, et al. (SIGMOD'98) (also, grid-based) To be covered in this lecture



Session 2: DBSCAN: A Density-Based Clustering Algorithm

DBSCAN: A Density-Based Spatial Clustering Algorithm

- DBSCAN (M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, KDD'96)

- Discovers clusters of arbitrary shape: Density-Based Spatial Clustering of Applications with Noise

- A *density-based* notion of cluster

- A *cluster* is defined as a maximal set of density-connected points

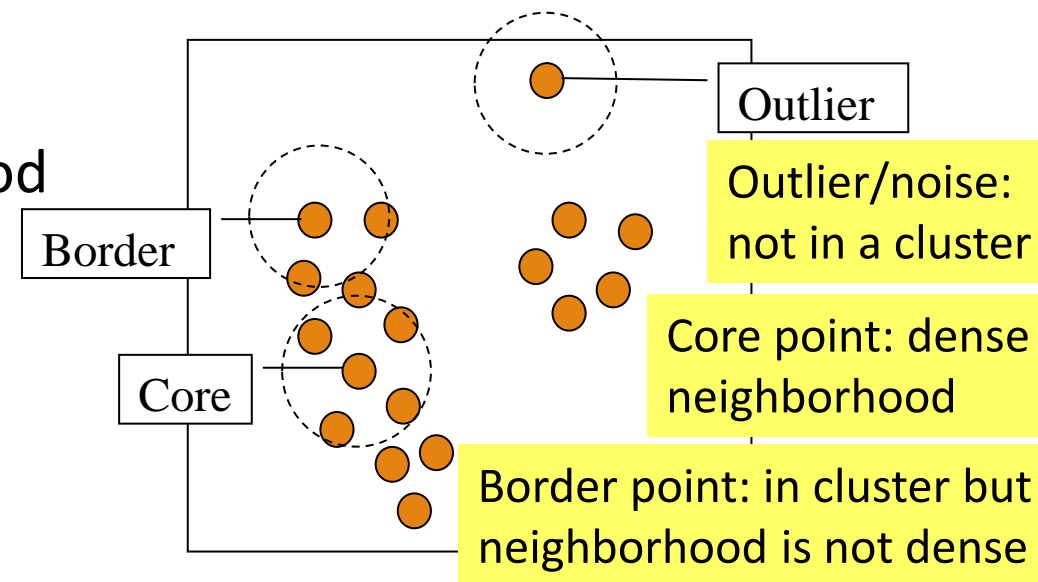
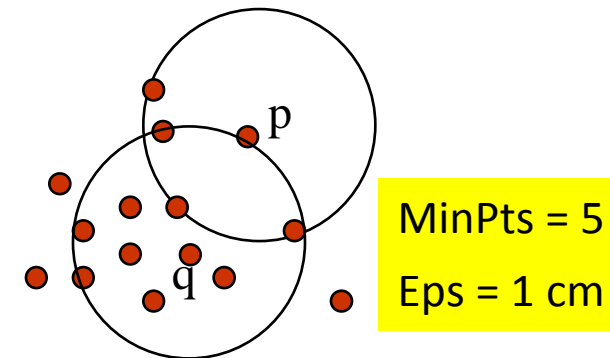
- Two parameters:

- Eps* (ϵ)**: Maximum radius of the neighborhood

- MinPts***: Minimum number of points in the Eps -neighborhood of a point

- The $Eps(\epsilon)$ -neighborhood of a point q :

- $N_{Eps}(q)$: $\{p \text{ belongs to } D \mid \text{dist}(p, q) \leq Eps\}$



DBSCAN: Density-Reachable and Density-Connected

□ Directly density-reachable:

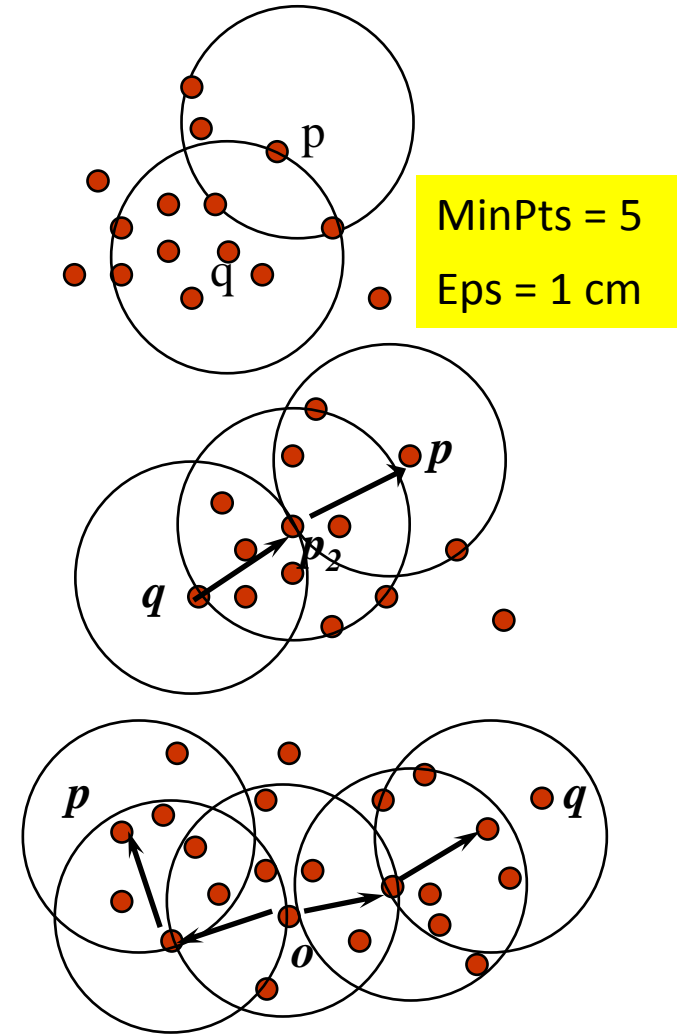
- A point p is **directly density-reachable** from a point q w.r.t. Eps (ϵ), $MinPts$ if
 - p belongs to $N_{Eps}(q)$
 - **core point** condition: $|N_{Eps}(q)| \geq MinPts$

□ 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 $MinPts$



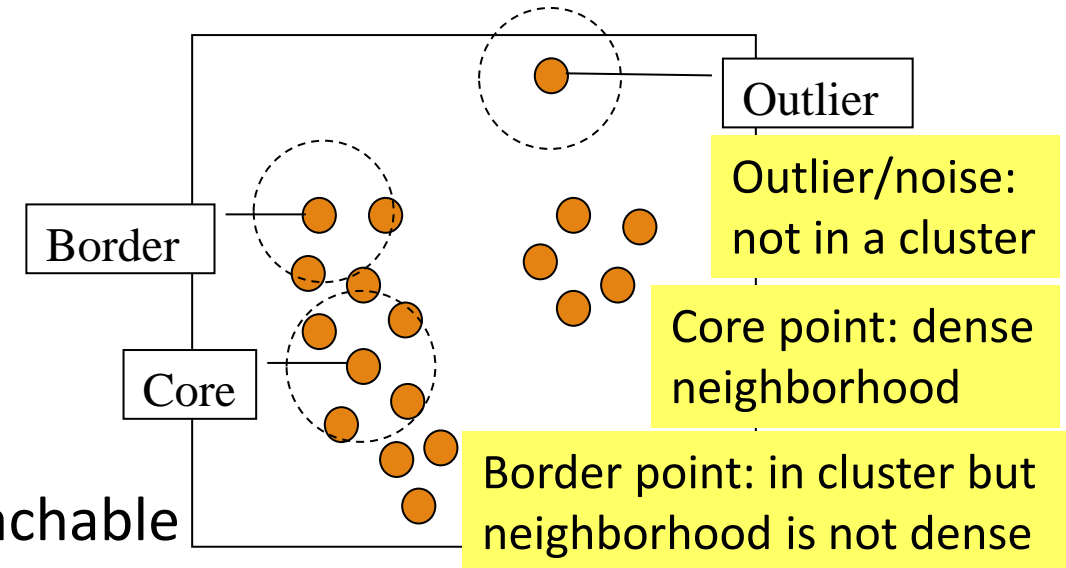
DBSCAN: The Algorithm

Algorithm

- Arbitrarily 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

Computational complexity

- If a spatial index is used, the computational complexity of DBSCAN is $O(n \log n)$, where n is the number of database objects
- Otherwise, the complexity is $O(n^2)$



DBSCAN Is Sensitive to the Setting of Parameters

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

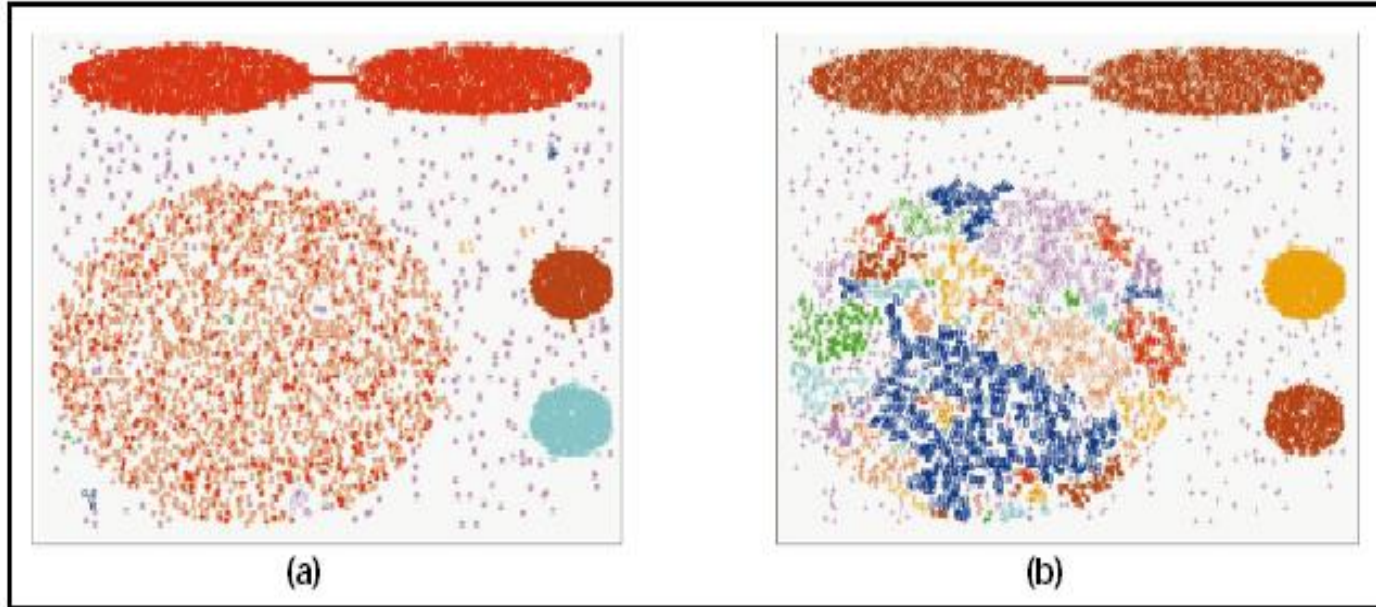
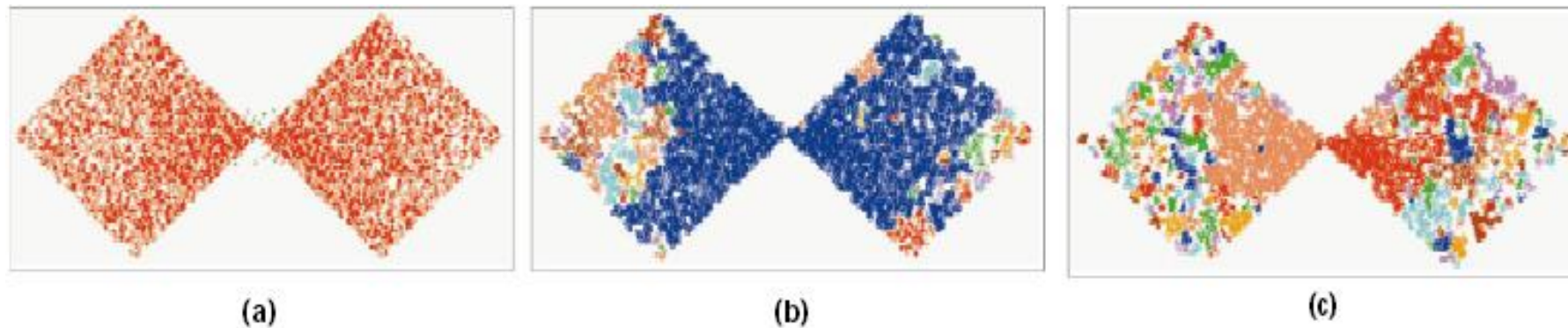


Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.



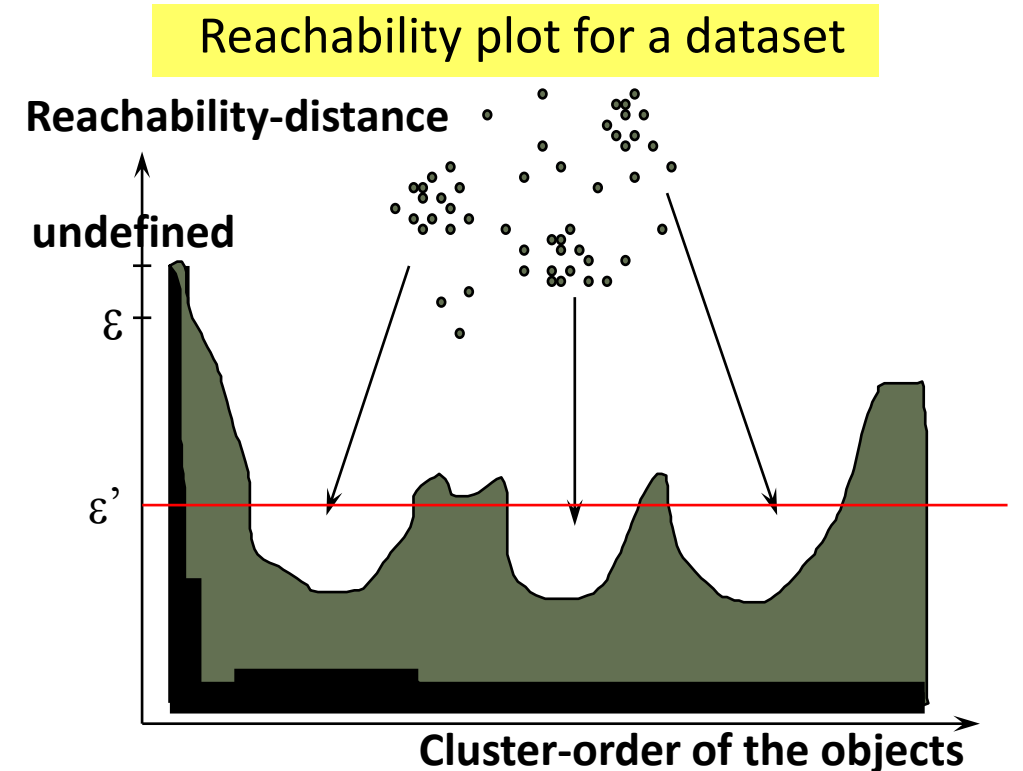
Ack. Figures from G. Karypis, E.-H. Han, and V. Kumar, *COMPUTER*, 32(8), 1999



Session 3: OPTICS: Ordering Points To Identify Clustering Structure

OPTICS: Ordering Points To Identify Clustering Structure

- ❑ OPTICS (Ankerst, Breunig, Kriegel, and Sander, SIGMOD'99)
 - ❑ DBSCAN is sensitive to parameter setting
 - ❑ An extension: finding clustering structure
- ❑ Observation: Given a *MinPts*, density-based clusters w.r.t. a higher density are completely contained in clusters w.r.t. to a lower density
- ❑ Idea: Higher density points should be processed first—find high-density clusters first
- ❑ OPTICS stores such a clustering order using two pieces of information:
 - ❑ *Core distance* and *reachability distance*



- ❑ Since points belonging to a cluster have a low reachability distance to their nearest neighbor, valleys correspond to clusters
- ❑ The deeper the valley, the denser the cluster

OPTICS: An Extension from DBSCAN

- Core distance of an object p : The smallest value ε such that the ε -neighborhood of p has at least $MinPts$ objects

Let $N_\varepsilon(p)$: ε -neighborhood of p

ε is a distance value

Core-distance $_{\varepsilon, MinPts}(p)$ = Undefined if $\text{card}(N_\varepsilon(p)) < MinPts$

$MinPts$ -distance(p), otherwise

- Reachability distance of object p from core object q is the min. radius value that makes p density-reachable from q

Reachability-distance $_{\varepsilon, MinPts}(p, q)$ =

Undefined, if q is not a core object

$\max(\text{core-distance}(q), \text{distance}(q, p))$, otherwise

- Complexity: $O(N \log N)$ (if index-based)

where N : # of points

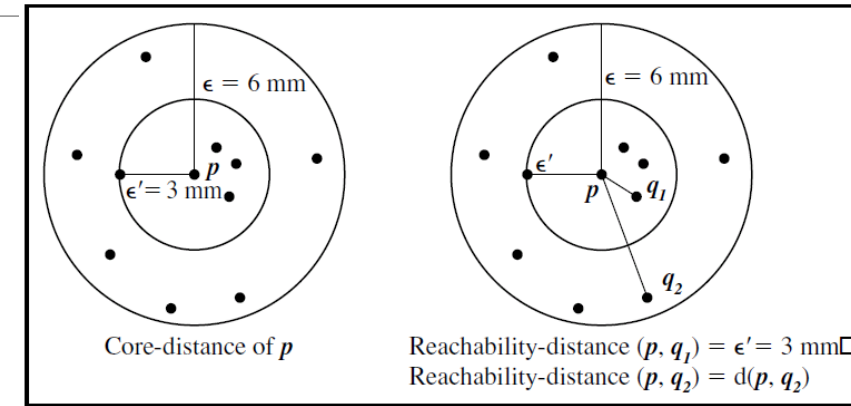
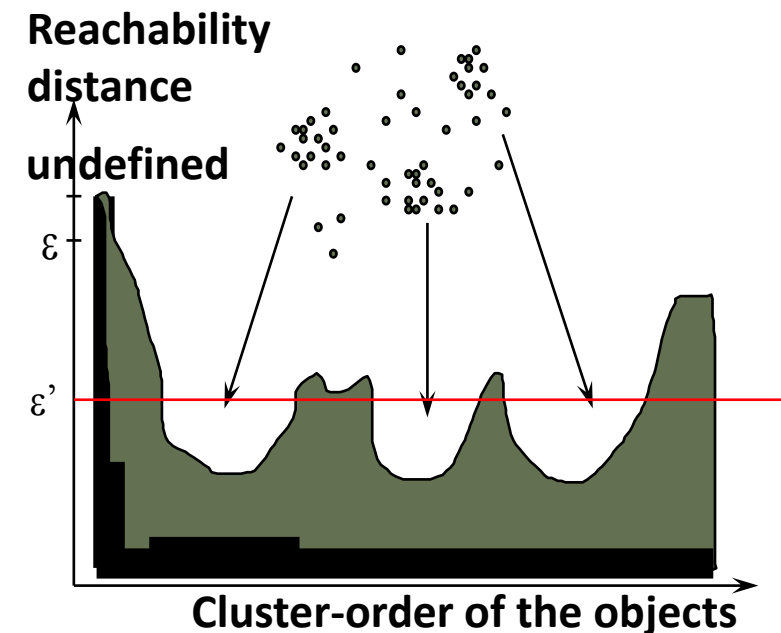
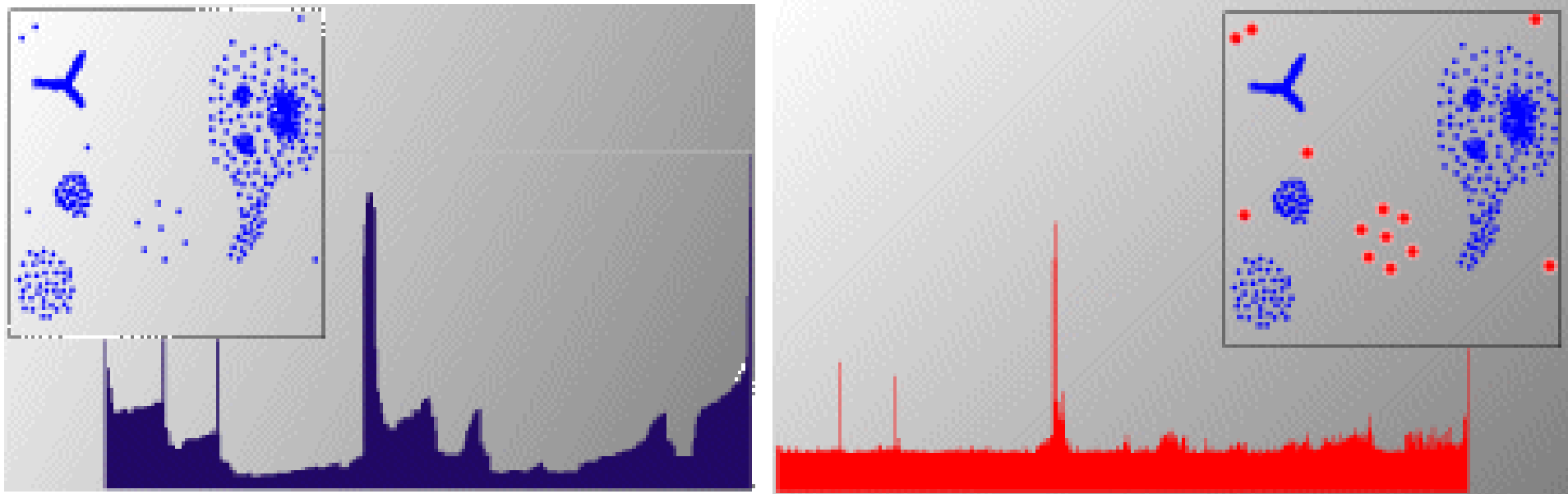


Figure 10.16: OPTICS terminology. Based on [ABKS99].



OPTICS: Finding Hierarchically Nested Clustering Structures

- OPTICS produces a special cluster-ordering of the data points with respect to its density-based clustering structure
- The cluster-ordering contains information equivalent to the density-based clusterings corresponding to a broad range of parameter settings
- Good for both automatic and interactive cluster analysis—finding intrinsic, even hierarchically nested clustering structures



Finding nested clustering structures with different parameter settings



Session 4: Grid-Based Clustering Methods

Grid-Based Clustering Methods

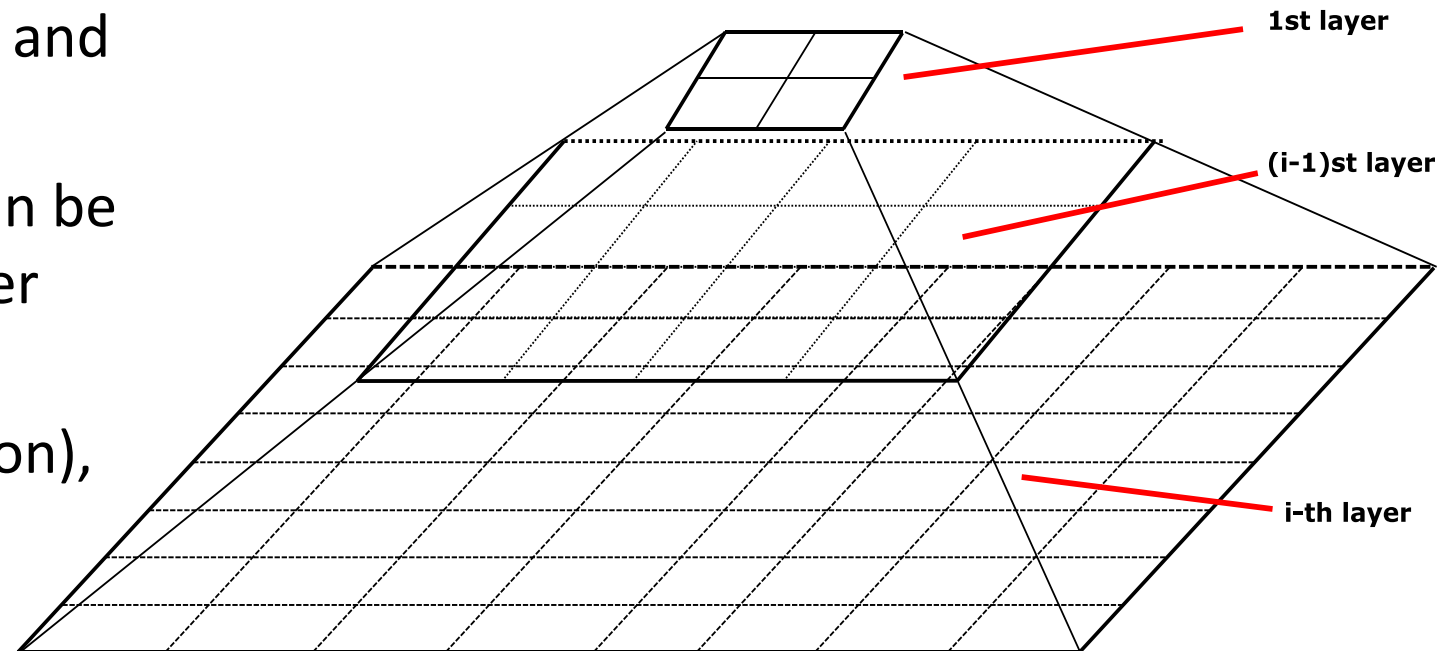
- ❑ Grid-Based Clustering: Explore multi-resolution grid data structure in clustering
 - ❑ Partition the data space into a finite number of cells to form a grid structure
 - ❑ Find clusters (dense regions) from the cells in the grid structure
- ❑ Features and challenges of a typical grid-based algorithm
 - ❑ Efficiency and scalability: # of cells \ll # of data points
 - ❑ Uniformity: Uniform, hard to handle highly irregular data distributions
 - ❑ Locality: Limited by predefined cell sizes, borders, and the density threshold
 - ❑ Curse of dimensionality: Hard to cluster high-dimensional data
- ❑ Methods to be introduced
 - ❑ **STING** (a Statistical INformation Grid approach) (Wang, Yang and Muntz, VLDB'97)
 - ❑ **CLIQUE** (Agrawal, Gehrke, Gunopulos, and Raghavan, SIGMOD'98)
 - ❑ Both grid-based and subspace clustering



Session 5: STING: A Statistical Information Grid Approach

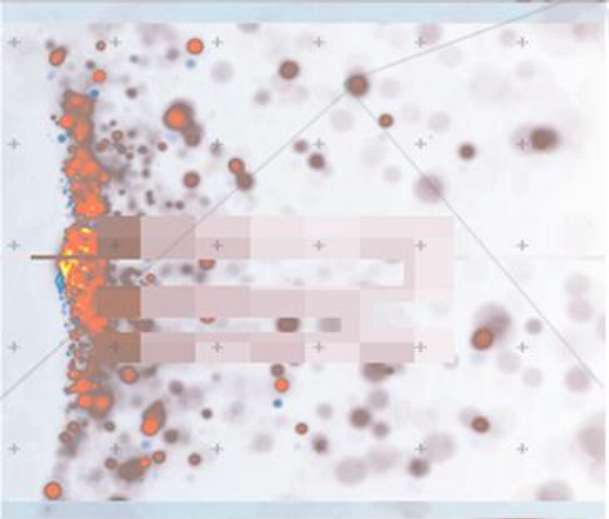
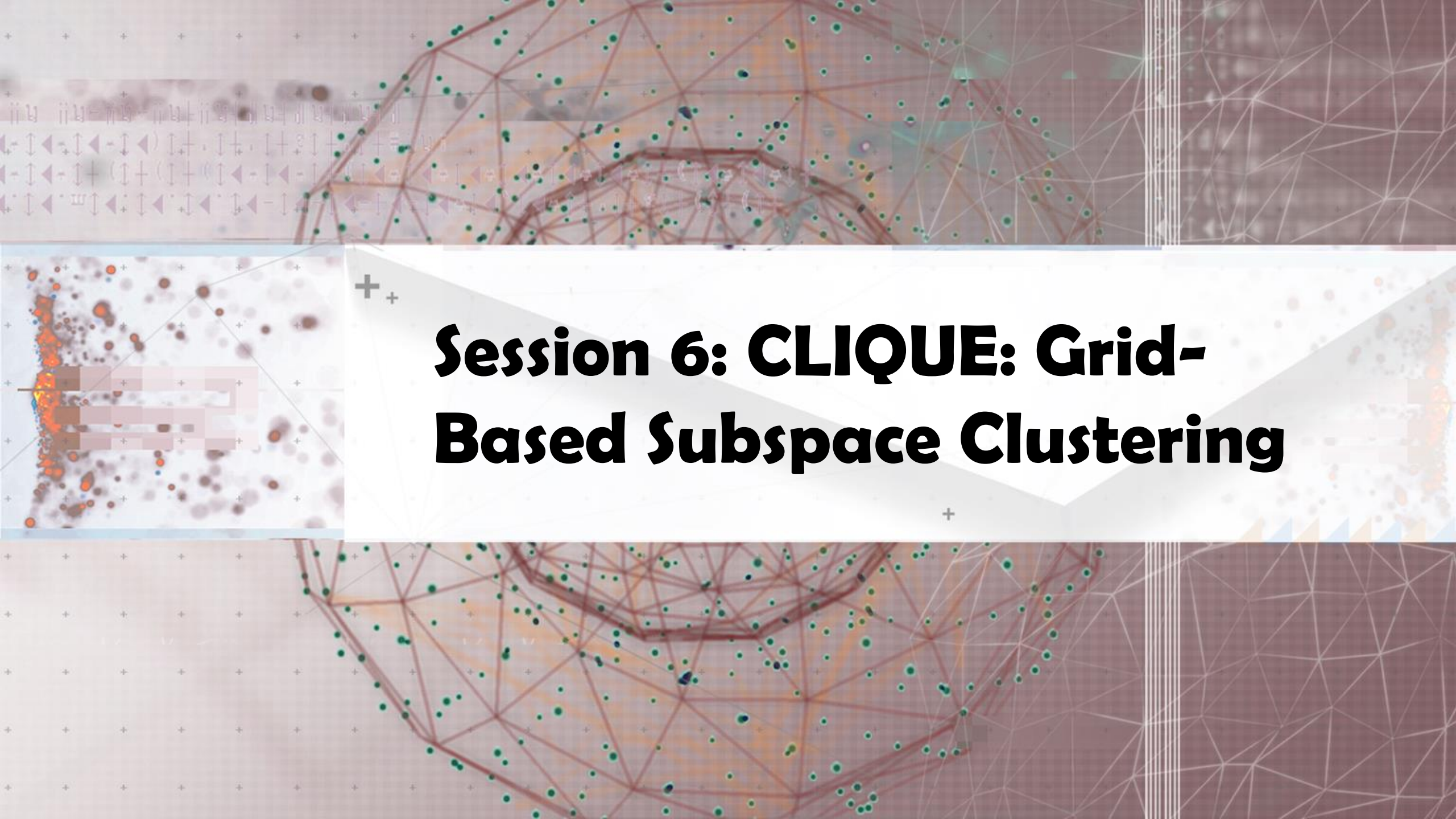
STING: A Statistical Information Grid Approach

- ❑ STING (Statistical Information Grid) (Wang, Yang and Muntz, VLDB'97)
- ❑ The spatial area is divided into rectangular cells at different levels of resolution, and these cells form a tree structure
- ❑ A cell at a high level contains a number of smaller cells of the next lower level
- ❑ Statistical information of each cell is calculated and stored beforehand and is used to answer queries
- ❑ Parameters of higher level cells can be easily calculated from that of lower level cell, including
 - ❑ *count, mean, s*(standard deviation), *min, max*
 - ❑ type of distribution—*normal, uniform, etc.*



Query Processing in STING and Its Analysis

- ❑ To process a region query
 - ❑ Start at the root and proceed to the next lower level, using the STING index
 - ❑ Calculate the likelihood that a cell is relevant to the query at some confidence level using the statistical information of the cell
 - ❑ Only children of likely relevant cells are recursively explored
 - ❑ Repeat this process until the bottom layer is reached
- ❑ Advantages
 - ❑ Query-independent, easy to parallelize, incremental update
 - ❑ Efficiency: Complexity is $O(K)$
 - ❑ K : # of grid cells at the lowest level, and $K \ll N$ (i.e., # of data points)
- ❑ Disadvantages
 - ❑ Its probabilistic nature may imply a loss of accuracy in query processing

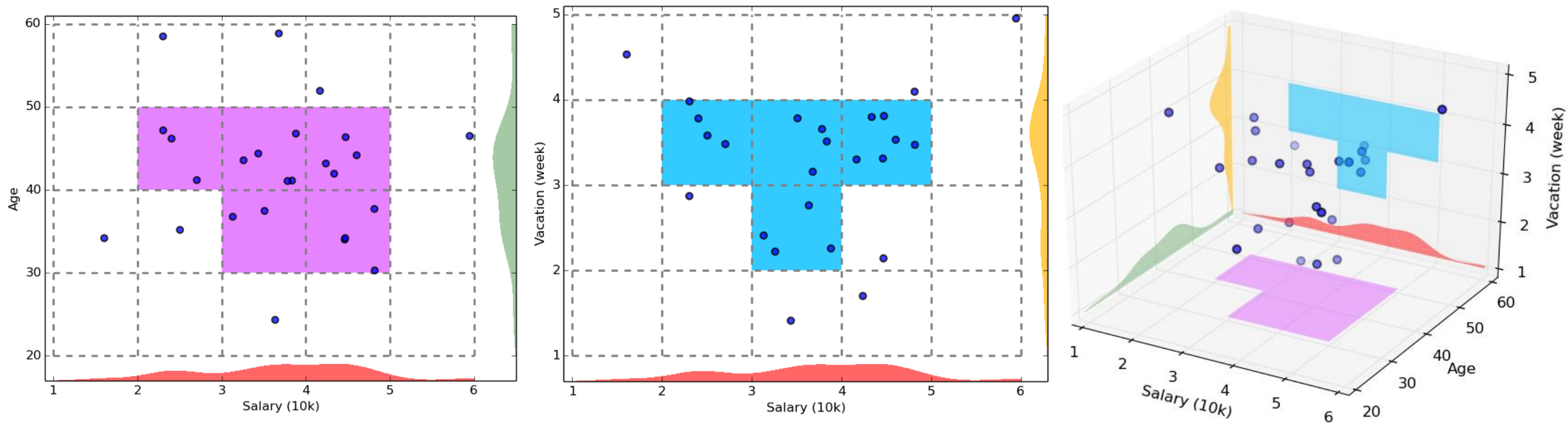


Session 6: CLIQUE: Grid- Based Subspace Clustering

CLIQUE: Grid-Based Subspace Clustering

- ❑ CLIQUE (Clustering In QUest) (Agrawal, Gehrke, Gunopulos, Raghavan: SIGMOD'98)
- ❑ CLIQUE is a **density-based** and **grid-based** **subspace clustering** algorithm
 - ❑ **Grid-based**: It discretizes the data space through a grid and estimates the density by counting the number of points in a grid cell
 - ❑ **Density-based**: A cluster is a maximal set of connected dense units in a subspace
 - ❑ A unit is dense if the fraction of total data points contained in the unit exceeds the input model parameter
 - ❑ **Subspace clustering**: A subspace cluster is a set of neighboring dense cells in an arbitrary subspace. It also discovers some minimal descriptions of the clusters
- ❑ It automatically identifies subspaces of a high dimensional data space that allow better clustering than original space using the Apriori principle

CLIQUE: SubSpace Clustering with Aprori Pruning



- ❑ Start at 1-D space and discretize numerical intervals in each axis into grid
- ❑ Find dense regions (clusters) in each subspace and generate their minimal descriptions
- ❑ Use the dense regions to find promising candidates in 2-D space based on the Apriori principle
- ❑ Repeat the above in level-wise manner in higher dimensional subspaces

Major Steps of the CLIQUE Algorithm

- ❑ Identify subspaces that contain clusters
 - ❑ Partition the data space and find the number of points that lie inside each cell of the partition
 - ❑ Identify the subspaces that contain clusters using the Apriori principle
- ❑ Identify clusters
 - ❑ Determine dense units in all subspaces of interests
 - ❑ Determine connected dense units in all subspaces of interests
- ❑ Generate minimal descriptions for the clusters
 - ❑ Determine maximal regions that cover a cluster of connected dense units for each cluster
 - ❑ Determine minimal cover for each cluster

Additional Comments on *CLIQUE*

□ Strengths

- *Automatically* finds subspaces of the highest dimensionality as long as high density clusters exist in those subspaces
- *Insensitive* to the order of records in input and does not presume some canonical data distribution
- Scales *linearly* with the size of input and has good scalability as the number of dimensions in the data increases

□ Weaknesses

- As in all grid-based clustering approaches, the quality of the results crucially depends on the appropriate choice of the number and width of the partitions and grid cells

The background of the slide is a complex, abstract composition. It features a central white banner with a subtle geometric pattern of thin lines. To the left of the banner is a rectangular inset showing a dense cluster of orange and red dots, resembling a galaxy or a data visualization. The main background is a mix of muted colors (pinks, purples, greys) with a grid of small white plus signs and a network of thin, intersecting lines in various colors (red, green, blue, orange).

Session 7: Summary

Summary: Density-Based and Grid-Based Clustering Methods

□ Density-Based Clustering

- Basic Concepts
- DBSCAN: A Density-Based Clustering Algorithm
- OPTICS: Ordering Points To Identify Clustering Structure

□ Grid-Based Clustering Methods

- Basic Concepts
- STING: A Statistical Information Grid Approach
- CLIQUE: Grid-Based Subspace Clustering

Recommended Readings

- ❑ M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases. KDD'96
- ❑ W. Wang, J. Yang, R. Muntz, STING: A Statistical Information Grid Approach to Spatial Data Mining, VLDB'97
- ❑ R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. SIGMOD'98
- ❑ A. Hinneburg and D. A. Keim. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. KDD'98
- ❑ M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: Ordering Points to Identify the Clustering Structure. SIGMOD'99
- ❑ M. Ester. Density-Based Clustering. In (Chapter 5) Aggarwal and Reddy (eds.), Data Clustering: Algorithms and Applications . CRC Press. 2014
- ❑ W. Cheng, W. Wang, and S. Batista. Grid-based Clustering. In (Chapter 6) Aggarwal and Reddy (eds.), Data Clustering: Algorithms and Applications. CRC Press. 2014