University of Colorado Boulder

# CSCI 4502/5502

# Data Mining

Fall 2019
Lecture 14 (Oct 17)

# Announcements

- Homework 4
  - due at 9:30am, Thursday, Oct 17

- Homework 5
  - posted at moodle
  - due at 9:30am, Thursday, Oct 24

# Review

- Chapter 9: Advanced Classification
  - Bayesian belief networks
  - Backpropagation
  - Support vector machines
  - Classification using frequent patterns
  - Lazy learners (or learning from your neighbors)
  - Other classification methods
  - Additional topics regarding classification
  - Summary

# Chapter 10:
# Cluster Analysis

# What is Cluster Analysis?

- Cluster: a collection of data objects
  - similar to one another within a cluster
  - dissimilar to objects in other clusters
- Cluster analysis
  - group similar data objects into clusters
  - similarity measure & clustering algorithm
- Unsupervised learning
  - no predefined classes

# Requirements of Clustering

- Scalability
- Different types of attributes
- Clusters with arbitrary shape
- Minimal domain knowledge for parameters
- Noisy data
- Incremental, insensitive to input order
- High dimensionality
- Constraint-based clustering
- Interpretability and usability

# Major Clustering Methods (1)

- Partitioning methods
  - construct k partitions, iterative relocation
  - e.g., k-means, k-medoids, CLARANS
- Hierarchical methods
  - hierarchical decomposition, split/merge
  - e.g., BIRCH, ROCK, Chameleon
- Density-based methods
  - connectivity and density functions
  - e.g., DBSCAN, OPTICS, DENCLUE

# Major Clustering Methods (2)

- Grid-based methods
  - quantize into cells, multi-granularity grid
  - e.g., STING, WaveCluster
- Model-based methods
  - hypothesized cluster model, best fit
  - e.g., EM, COBWEB, SOM

# Major Clustering Methods (3)

- Clustering high-dimensional data
  - subspace clustering: CLIQUE, PROCLUS
  - frequent-pattern-based clustering: pCluster
- Constraint-based clustering
  - user-specified or application-oriented constraints
  - e.g., COD (obstacles), user-constrained clustering, semi-supervised clustering

# Chapter 10: Cluster Analysis

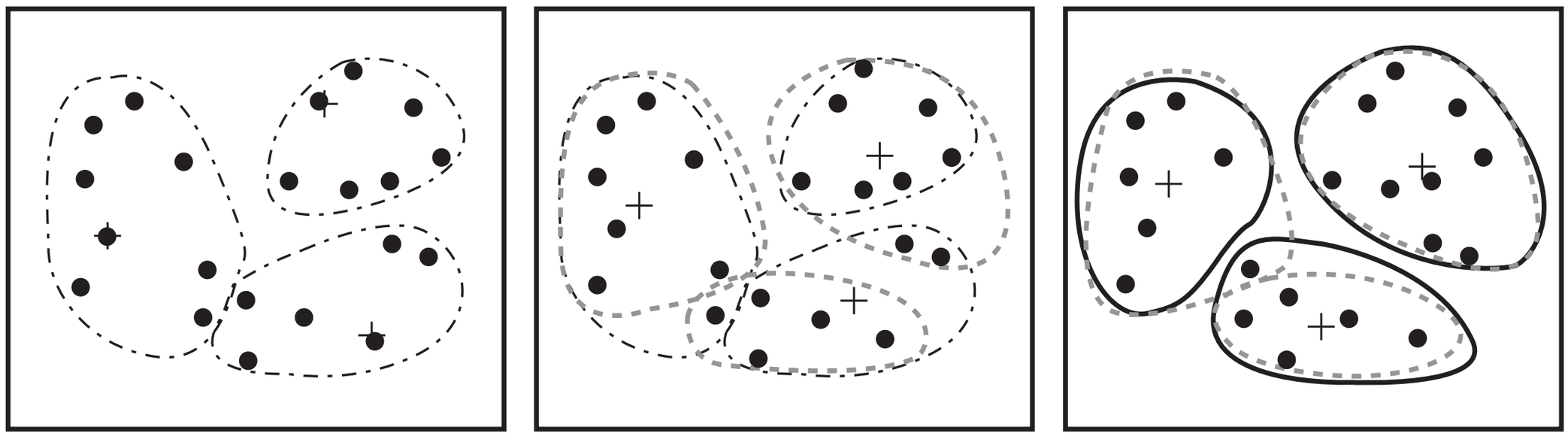
- Basic concepts
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Evaluation of clustering
- Summary

# Partitioning Methods

- Given a data set D of n objects
- Given k, find a partition of k clusters that optimizes the chosen partition criterion
- Global optimal: enumerate all partitions
- Heuristic methods
  - k-means: cluster represented by mean (centroid)
  - k-medoids: cluster represented by medoid (object closest to centroid)

# k-Means Clustering (1)

- Partition objects into k nonempty clusters
- Compute mean (centroid) of each cluster
- Assign each object to closest centroid
- Repeat till no more assignment changes



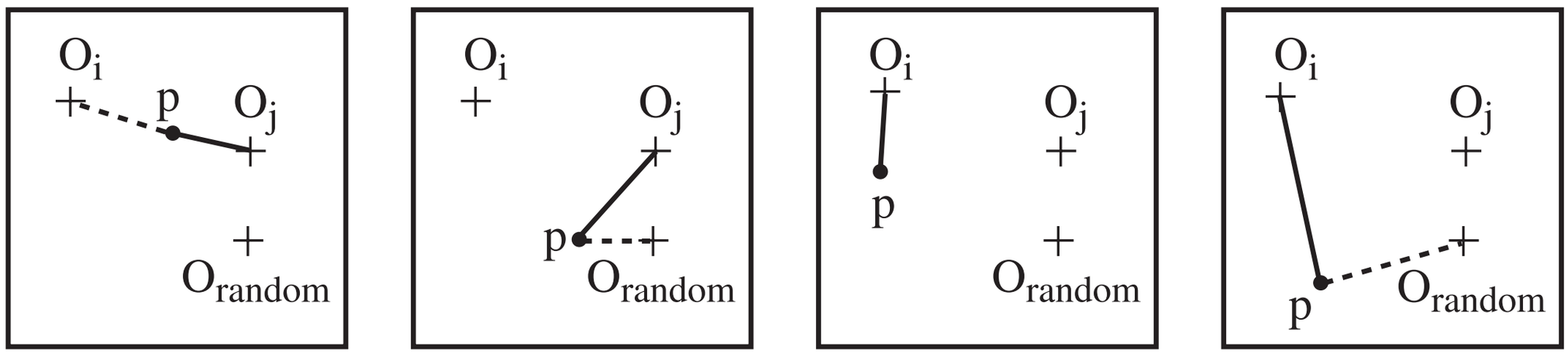
(a) (b) (c)

# k-Means Clustering (2)

- Relatively efficient: O(nkt)
  - n: #objects, k: #clusters, t: #iterations
- Often terminates at local optimal
- Applicable only when centroid is defined
- Need to specify k in advance
- Not suitable for discovering clusters with non-convex shapes
- Sensitive to noise and outliers

# k-Medoids Clustering (1)

- k-means method is sensitive to outliers
  - substantially distort the distribution of data
- k-medoids: find representative objects (medoids)




- data object
+ cluster center
— before swapping
--- after swapping

# k-Medoids Clustering (2)

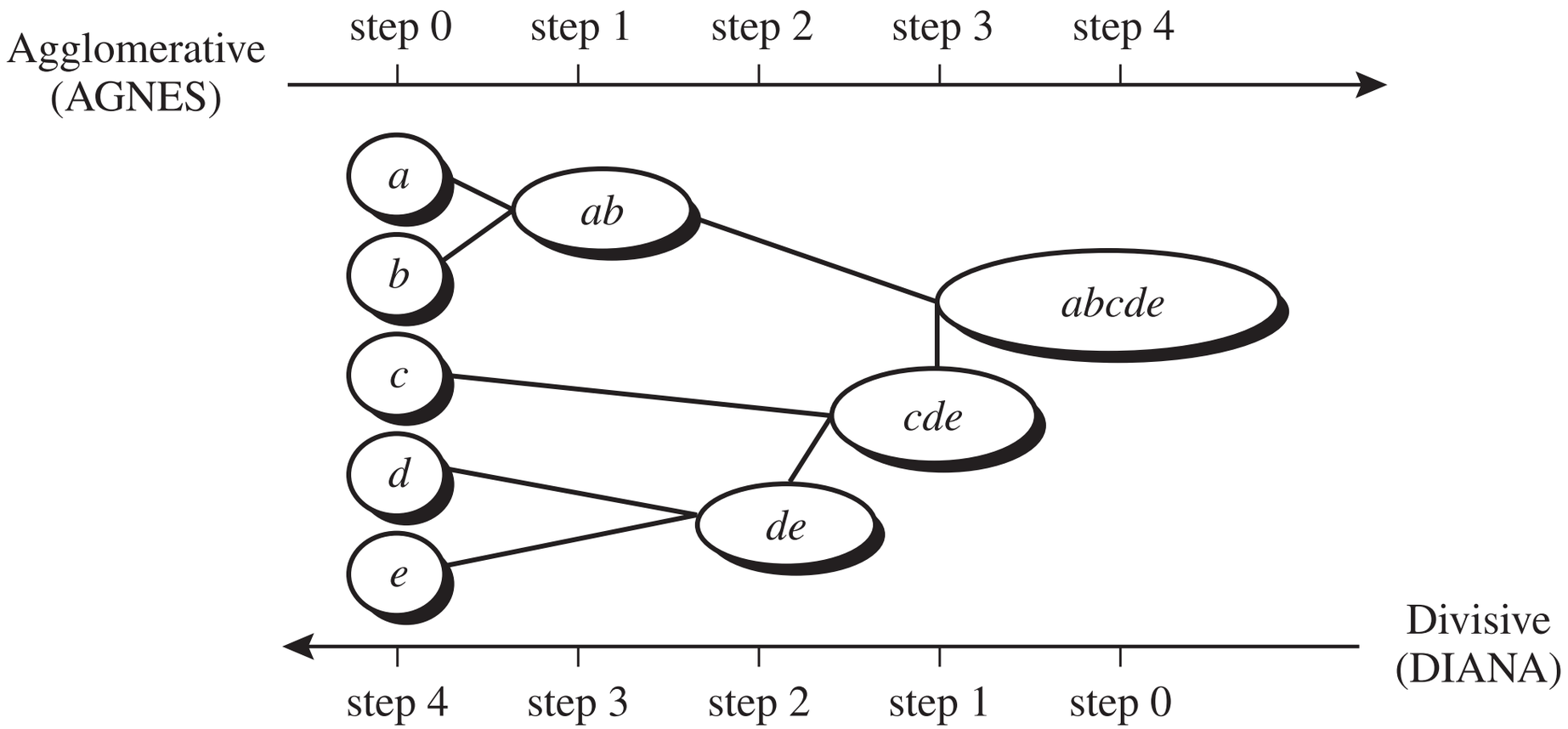
- PAM (Partitioning Around Medoids)
  - starts from an initial set of medoids
  - iteratively replace a medoid w/ a non-medoid if it reduces the total distance
    - effective for small data sets, does not scale
      - $O(k(n-k)^2)$ for each iteration
- CLARA: apply PAM on multiple sampled sets
- CLARANS: use randomized sample to search for neighboring solutions

# Chapter 10: Cluster Analysis

- Basic concepts
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Evaluation of clustering
- Summary

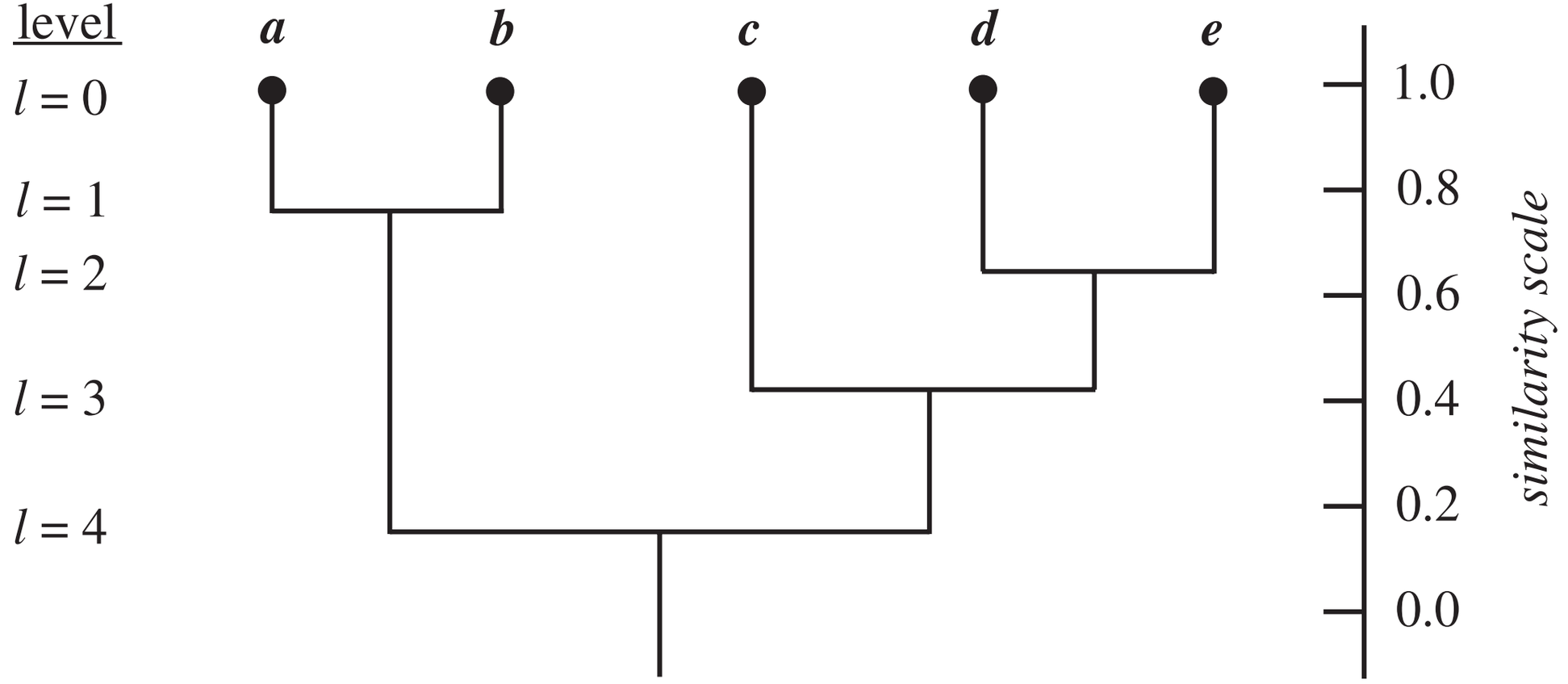
# Hierarchical Clustering

- Groups data objects into a tree of clusters
- Agglomerative: bottom-up merging
- Decisive: top-down splitting



# Dendrogram

- Represents the process of hierarchical clustering
- Clustering: cut dendrogram at a certain level



# Distance between Clusters

- Minimum distance

$$d_{min}(C_i, C_j) = min_{p \in C_i, p' \in C_j} |p - p'|$$

- Maximum distance

$$d_{max}(C_i, C_j) = max_{p \in C_i, p' \in C_j} |p - p'|$$

- Mean distance

$$d_{mean}(C_i, C_j) = |m_i - m_j|$$

- Average distance

$$d_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i} \sum_{p' \in C_j} |p - p'|$$

# Hierarchical Clustering

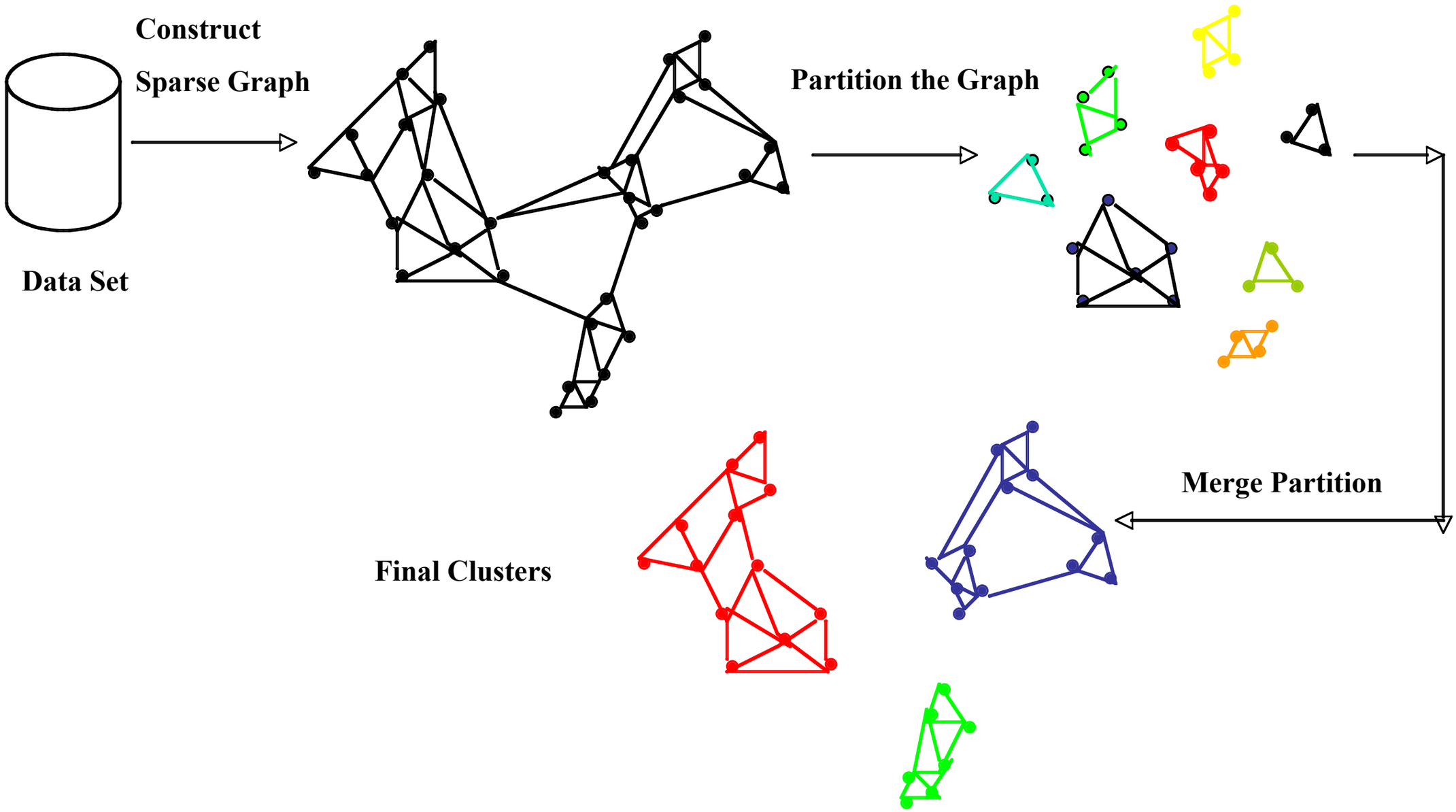
- Uses distance matrix as clustering criteria
- Does not need to specify k (#clusters)
- Termination condition
  - e.g., cluster distance exceeds a threshold
  - Cannot undo previous merge/split decisions
- BIRCH (1996): CF-tree, microclusters
- CHAMELEON (1999): dynamic modeling

# BIRCH

- Clustering large amount of numerical data
- Multi-phase clustering
  - Phase 1: microclustering
    - hierarchical clustering
  - Phase 2: macroclustering
    - e.g., iterative partitioning
- Clustering feature
  - CF = <n, LS, SS> $\sum_{i=1}^{N} X_i$ $\sum_{i=1}^{N} X_i^2$

# CHAMELEON (1)

## Overall framework



# CHAMELEON (2)

- k-nearest-neighbor graph
- Graph partition: edge cut EC($C_i$, $C_j$)
- Relative interconnectivity (#edge cuts)

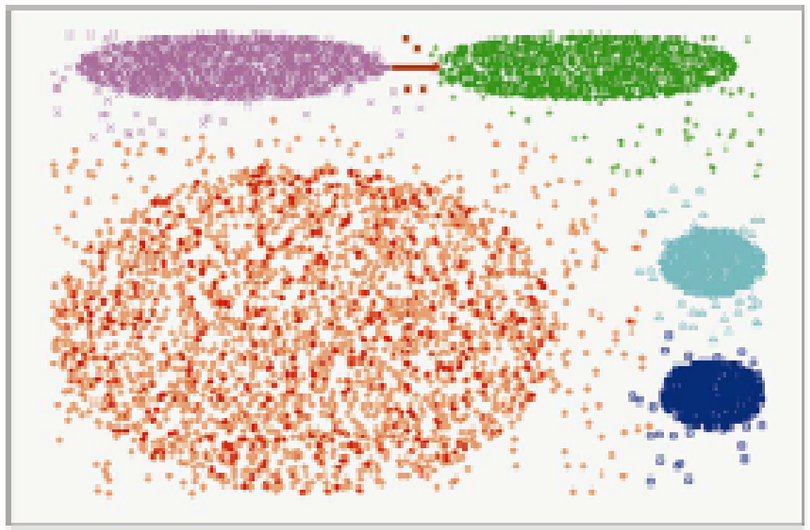
$$RI(C_i, C_j) = \frac{|EC_{\{C_i,C_j\}}|}{\frac{1}{2}(|EC_{C_i}| + |EC_{C_j}|)}$$

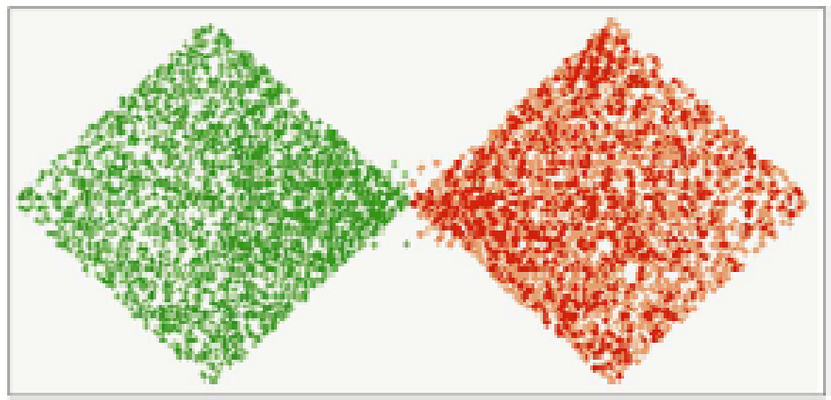
- Relative closeness (avg weight of edge

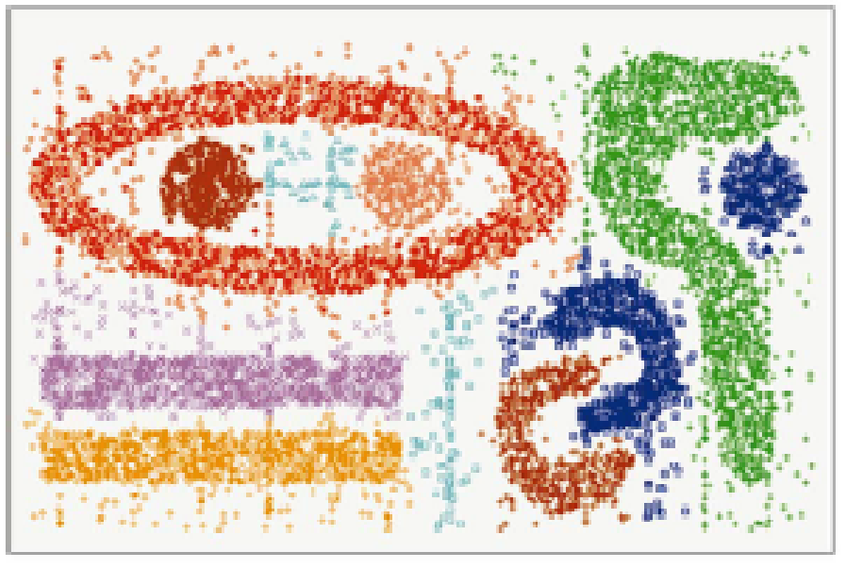
$$RC(C_i, C_j) = \frac{\bar{S}_{EC_{\{C_i,C_j\}}}}{\frac{|C_i|}{|C_i|+|C_j|}\bar{S}_{EC_{C_i}} + \frac{|C_j|}{|C_i|+|C_j|}\bar{S}_{EC_{C_j}}}$$

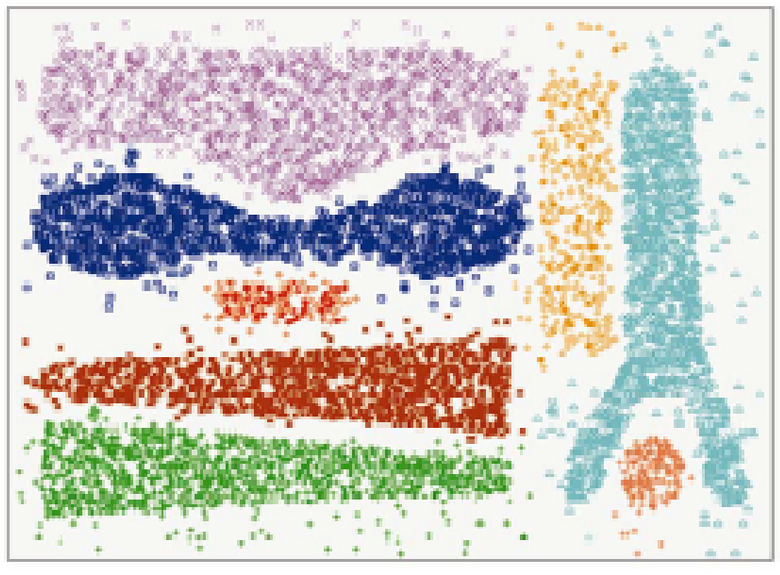
# CHAMELEON (3)

- Clustering complex objects









# Chapter 10: Cluster Analysis

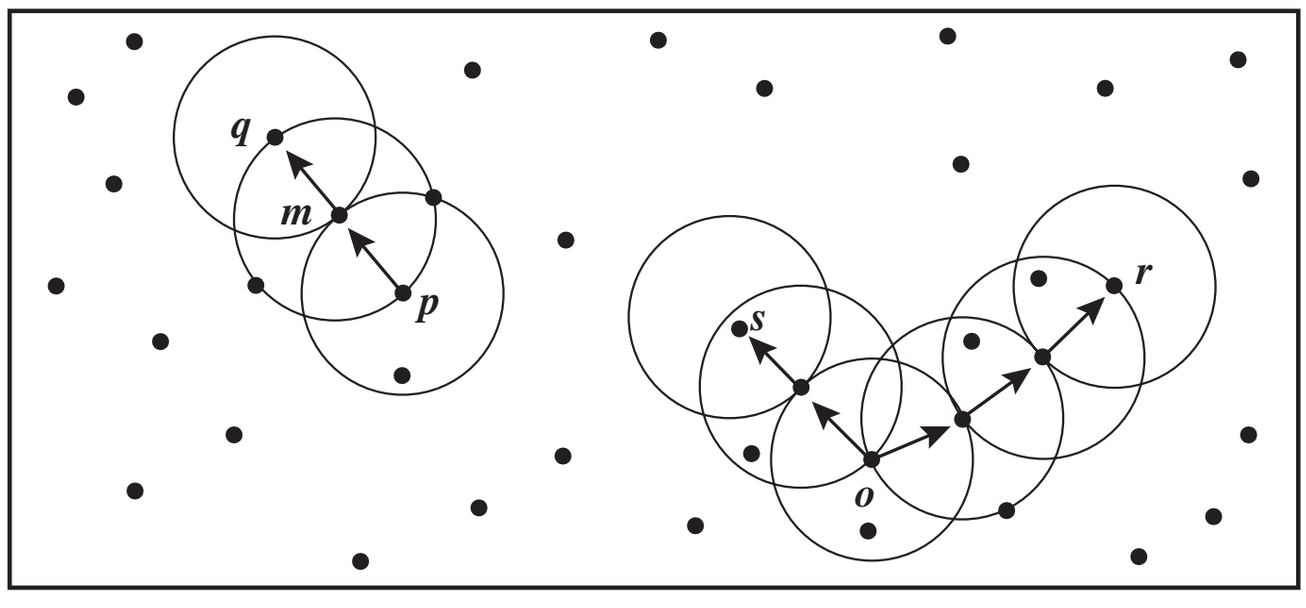
- Basic concepts
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Evaluation of clustering
- Summary

# Density-Based Clustering

- Clustering based on density (local cluster criterion), e.g., density-connected points
- Major features
  - clusters of arbitrary shape, handles noise, single scan, density parameter for termination
- Typical methods
  - DBSCAN, DENCLUE

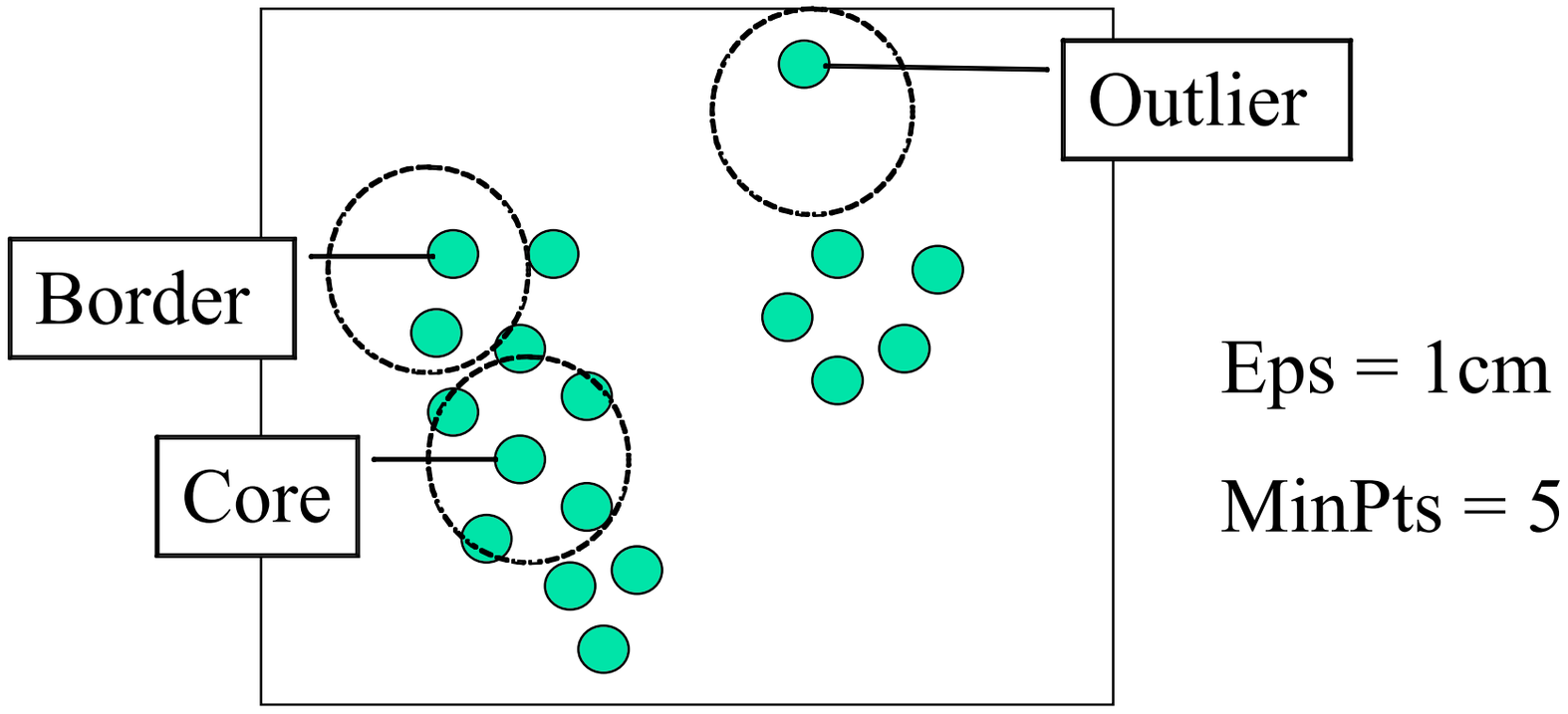
# DBSCAN (1)

- ε-neighborhood of p: within radius ε of p
- Core object p: at least MinPts points in the ε-neighborhood of p
- Directly density reachable
- Density reachable
- Density connected



# DBSCAN (2)

- Cluster: a maximal set of density-connected points
- Check ε-neighborhood of each point p
- Core object? Border object?



# DBSCAN (3)

- Sensitive to parameters: ε and MinPts

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

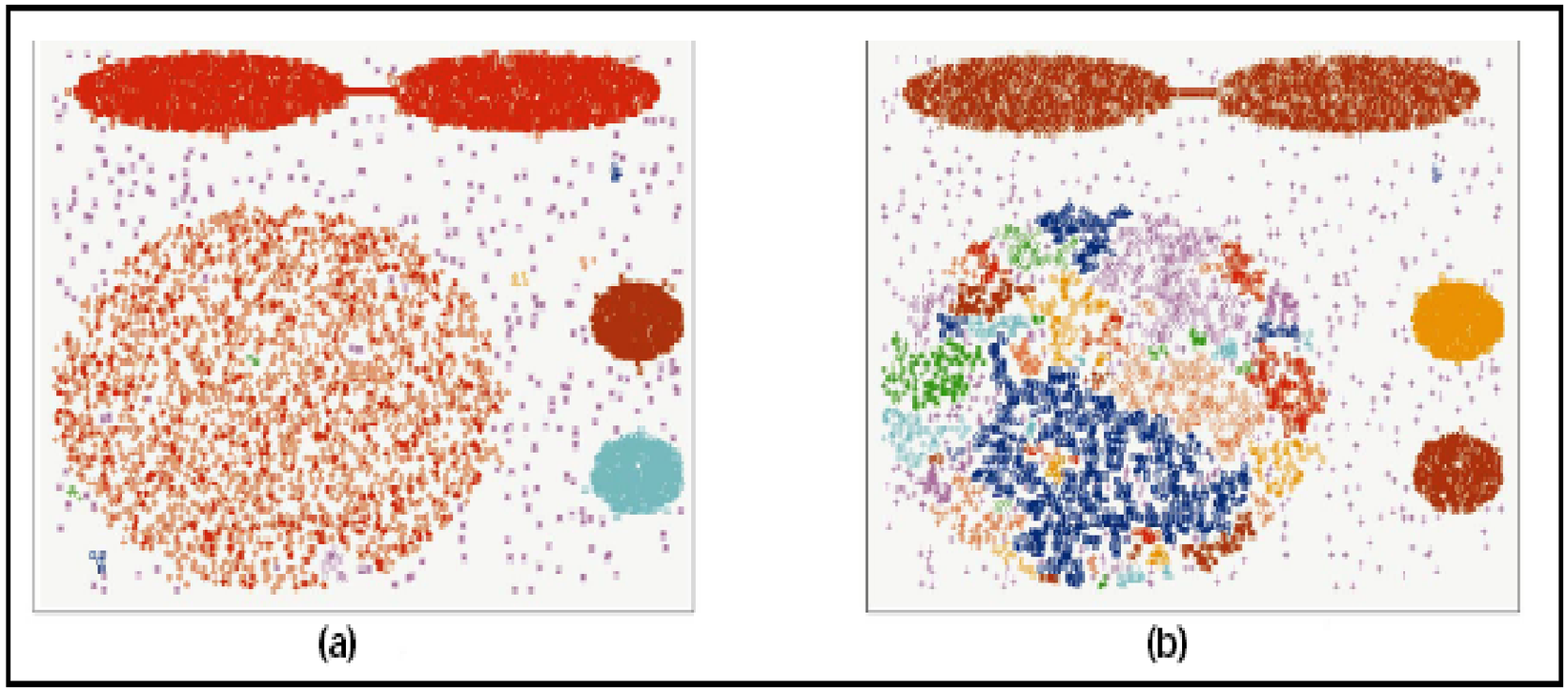
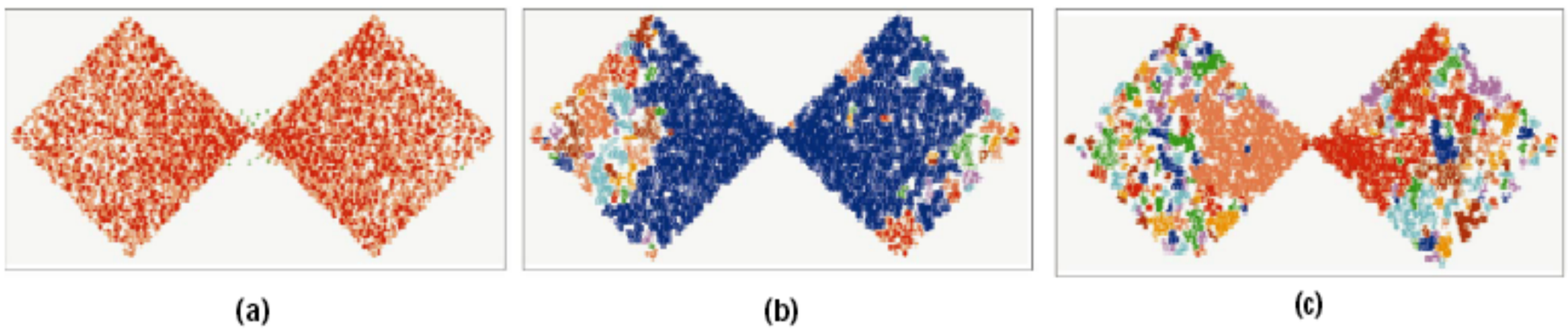



*Figure 9. DBScan results for DS2 with MinPts at 4 and Eps at (a) 5.0, (b) 3.5, and (c) 3.0.*



# DENCLUE (1)

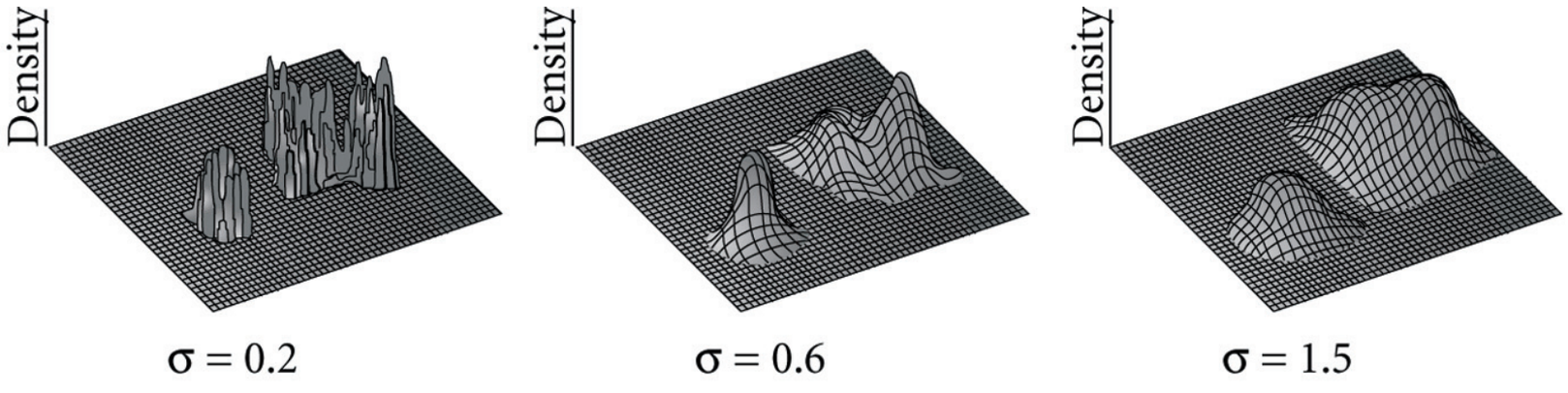
- Uses statistical density functions
- Major features
  - solid mathematical foundation
  - good for data sets with large amounts of noise
  - compact description of arbitrarily-shaped clusters in high-dimensional data sets
  - significantly faster than existing algorithm
  - needs a large number of parameters

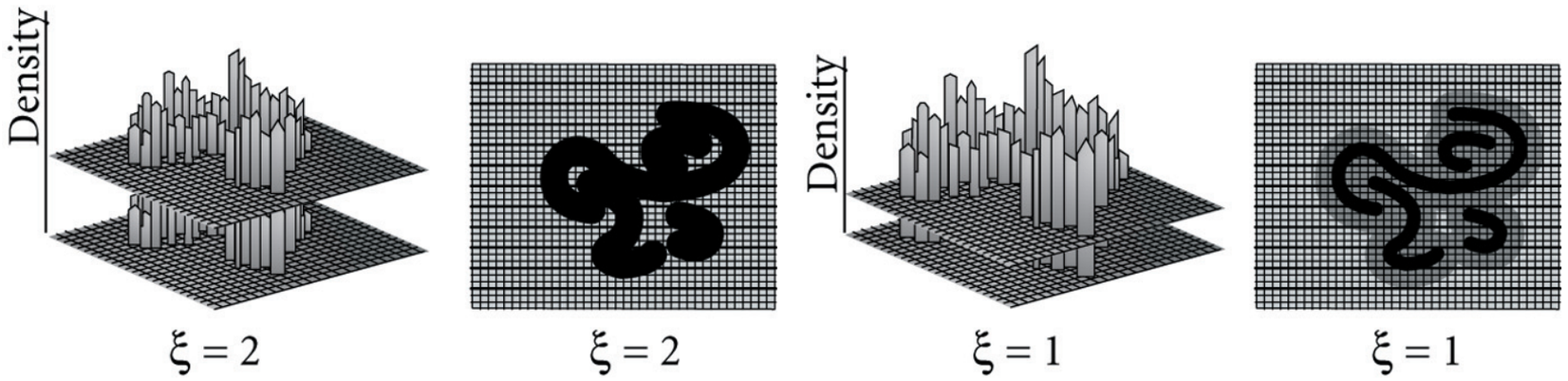
# DENCLUE (2)

- Influence function: impact of a data point within its neighborhood
- Overall density: sum of the influence function of all data points
- Density attractors: local maximal of overall density function
- Clusters can be determined mathematically by identifying density attractors

# DENCLUE (3)

- Center-defined or arbitrary-shaped clusters



# Chapter 10: Cluster Analysis

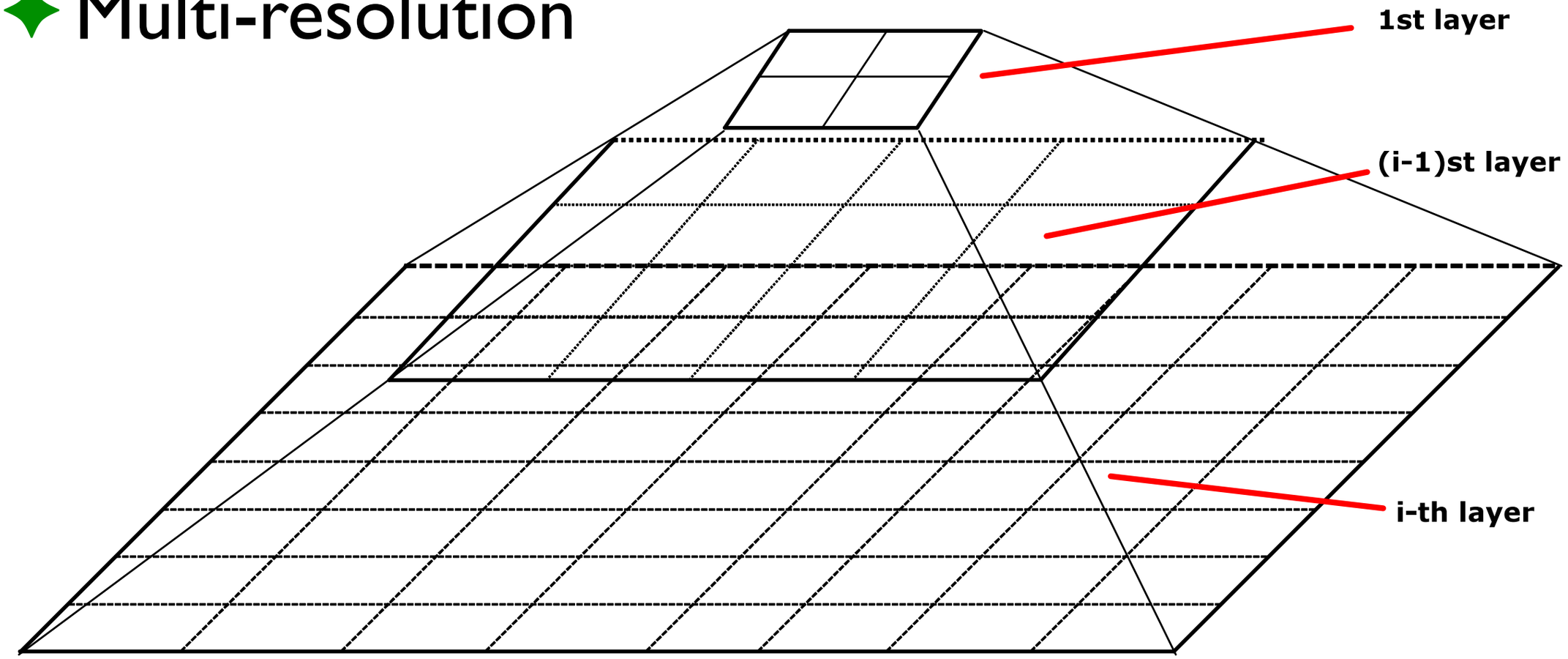
- Basic concepts
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Evaluation of clustering
- Summary

# Grid-Based Clustering

- Uses multi-resolution grid data structure
- Quantizes object space to cells in the grid
- Fast processing time
  - depends on #cells, not #objects
- Typical methods
  - STING: statistical info of grid cells
  - CLIQUE: high-dimensional data

# STatistical INformation Grid

- Wang, Yang, and Muntz (VLDB'97)
- Spatial area => rectangular cells
- Multi-resolution



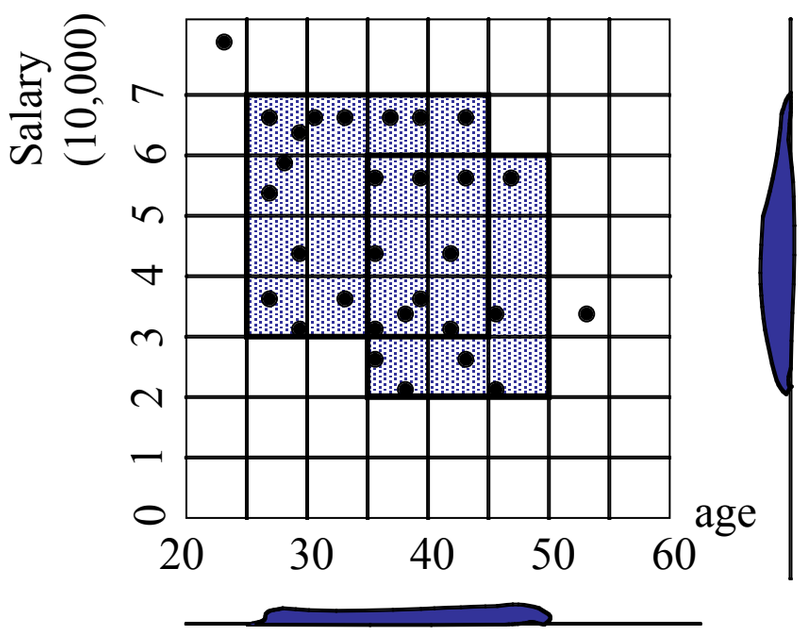
# STING: Pros and Cons

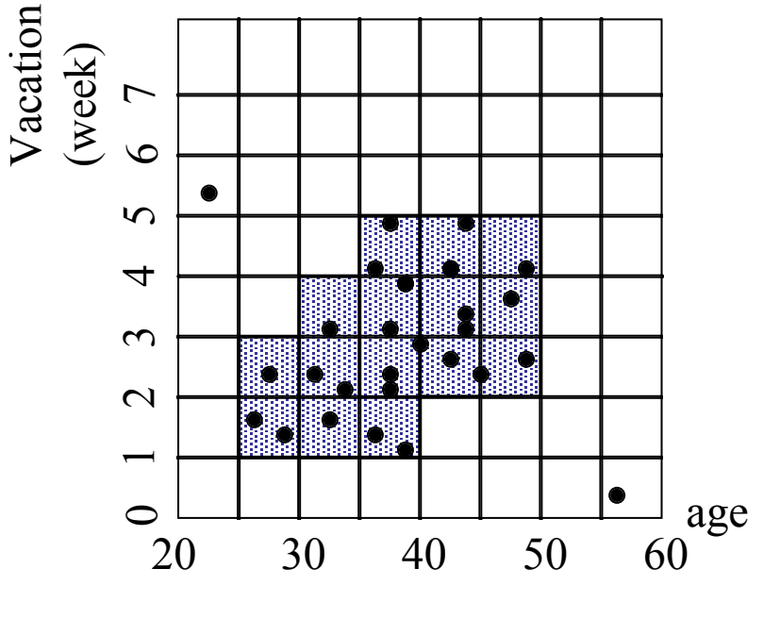
- Pros
  - O(g): number of grid cells at bottom level
  - query-independent, easy to parallelize, incremental update
- Cons
  - finer granularity vs. coarser granularity
  - only horizontal or vertical cluster boundaries, no diagonal boundaries

# CLIQUE (1)

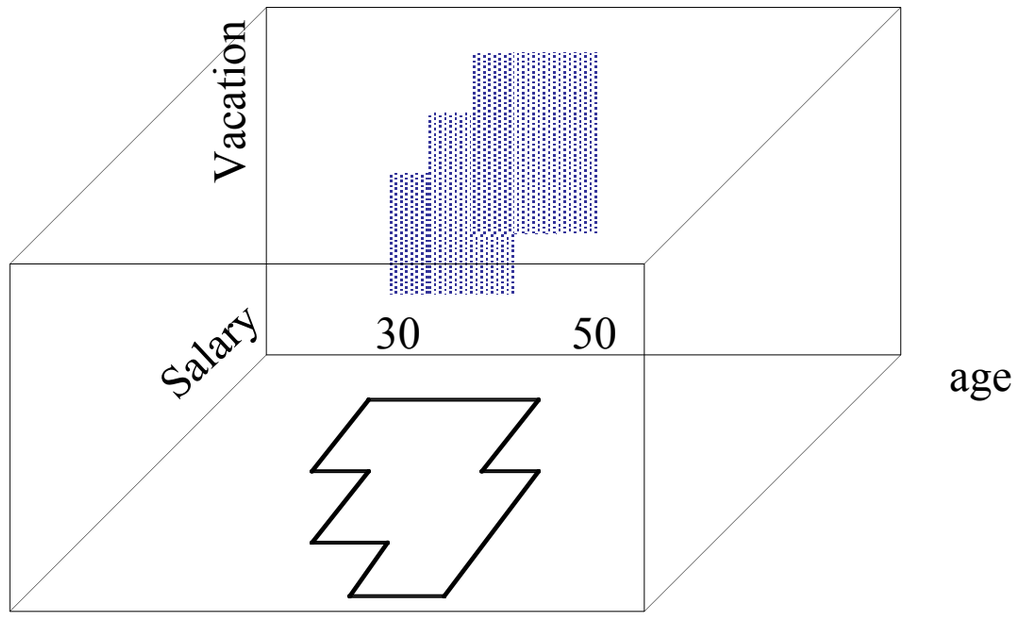
- Agrawal, Gehrke, Gunopulos, Raghavan (SIGMOD'98)
- Dimension-growth subspace clustering
  - grow from single dimensions to high dims
- Both density-based and grid-based
  - each dimension => equal-width intervals
  - non-overlapping rectangular units
  - cluster: a set of connected dense units

# CLIQUE (2)




$\tau = 3$



# Chapter 10: Cluster Analysis

- Basic concepts
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Evaluation of clustering
- Summary

# Evaluation of Clustering

- Major tasks
  - clustering tendency
  - number of clusters
  - clustering quality

- Clustering quality
  - extrinsic methods (w/ ground truth)
  - intrinsic methods (w/o ground truth)

# Chapter 10: Cluster Analysis

- Basic concepts
- Partitioning methods
- Hierarchical methods
- Density-based methods
- Grid-based methods
- Evaluation of clustering
- Summary