University of Colorado Boulder

CSCI 4502/5502

Data Mining

Fall 2019
Lecture 15 (Oct 22)

Reminders

- Homework 5
  - due at 9:30am, Thursday, Oct 24
- Midterm Exam
  - Th Oct 31: midterm exam
- Homework grading
  - HW 1-3 are graded, please check moodle
  - HW 4 is being graded

Review

- Chapter 10: Cluster Analysis
  - Basic concepts
  - Partitioning methods
  - Hierarchical methods
  - Density-based methods
  - Grid-based methods
  - Evaluation of clustering
  - Summary

Chapter 11:
Advanced Cluster Analysis

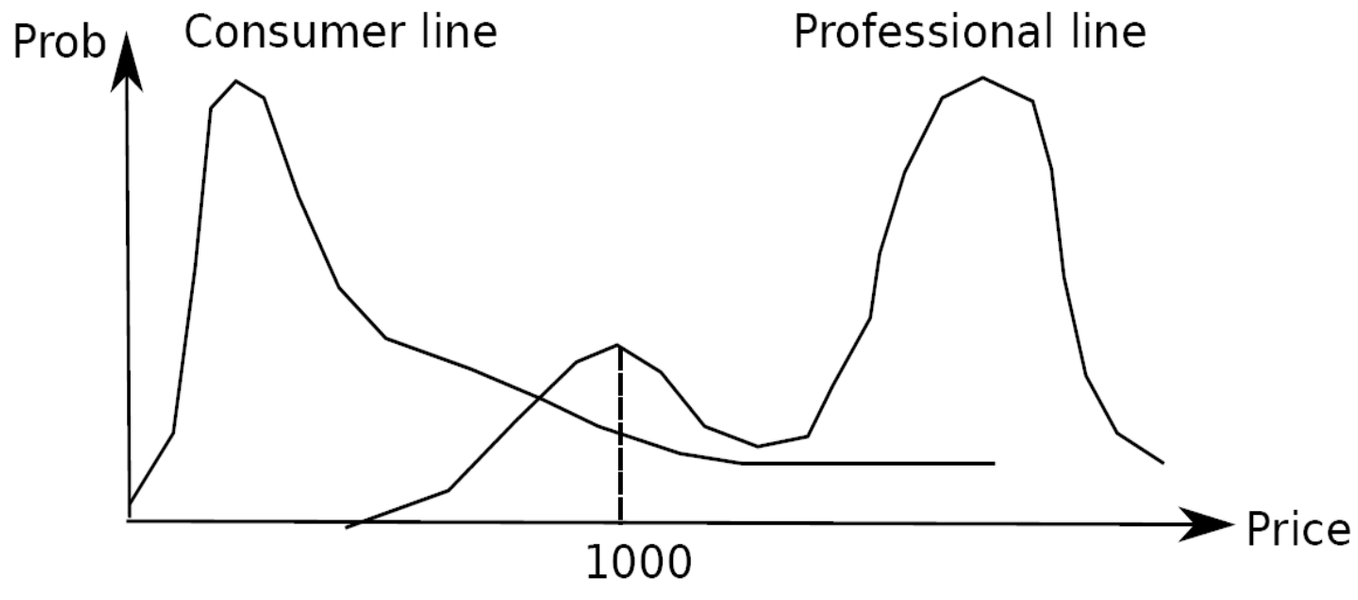
- Chapter 11: Advanced Cluster Analysis
  - probabilistic model-based clustering
  - clustering high-dimensional data
  - clustering graph and network data
  - clustering with constraints

Fuzzy Clusters

- Cluster membership of each object
  - belongs to a single cluster
  - weighted distribution in multiple clusters
- Fuzzy clusters (soft clusters)
  - n objects, k clusters
  - w_ij: probability of object i belonging to cluster j
  - 0 <= w_ij <= 1

Probabilistic Clusters

- Hidden categories (probabilistic clusters)
  - each represented by a probability density function over the data space
- Mixture model: observed data instances drawn independently from multiple clusters



Model-Based Clustering

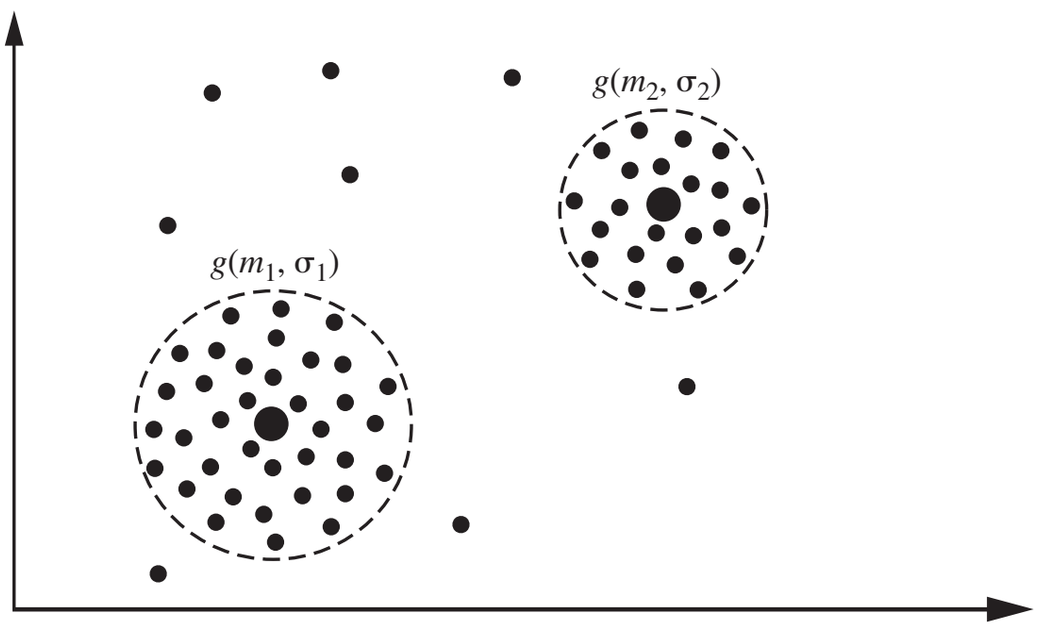
- Assumption: Data are generated by a mixture of underlying probability distributions

$$P(D|\boldsymbol{C}) = \prod_{i=1}^{n} P(o_i|\boldsymbol{C}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j f_j(o_i)$$

- Attempt to optimize the fit between data and some mathematical model
  - find a set C of k probabilistic clusters s.t. P(D|C) is maximized

EM: Expectation Maximization

- A popular iterative refinement algorithm
- An extension to k-means
  - assign each object to a cluster according to a weight (probability distribution)
  - new means computed on weighted sum
- Mixture of k distributions
  - distribution => cluster
  - e.g., Gaussian distr.
  - θ_j = (μ_j, σ_j)



The EM Algorithm (1)

- Expectation step (E-step)

$$P(\Theta_j|o_i,\mathbf{\Theta}) = \frac{P(o_i|\Theta_j)}{\sum_{l=1}^{k} P(o_i|\Theta_l)}$$

- Maximization step (M-step)

$$\mu_j = \sum_{i=1}^{n} o_i \frac{P(\Theta_j|o_i,\mathbf{\Theta})}{\sum_{l=1}^{n} P(\Theta_j|o_l,\mathbf{\Theta})} = \frac{\sum_{i=1}^{n} o_i P(\Theta_j|o_i,\mathbf{\Theta})}{\sum_{i=1}^{n} P(\Theta_j|o_i,\mathbf{\Theta})}$$

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{n} P(\Theta_j|o_i,\mathbf{\Theta})(o_i-u_j)^2}{\sum_{i=1}^{n} P(\Theta_j|o_i,\mathbf{\Theta})}}$$

The EM Algorithm (2)

- Simple, easy to implement
- Can be characterized by a few parameters
- Generally converges quickly but may not reach the global optima
- Computationally expensive if number of distributions is large or data set contains very few observed data points

Clustering High-D Data

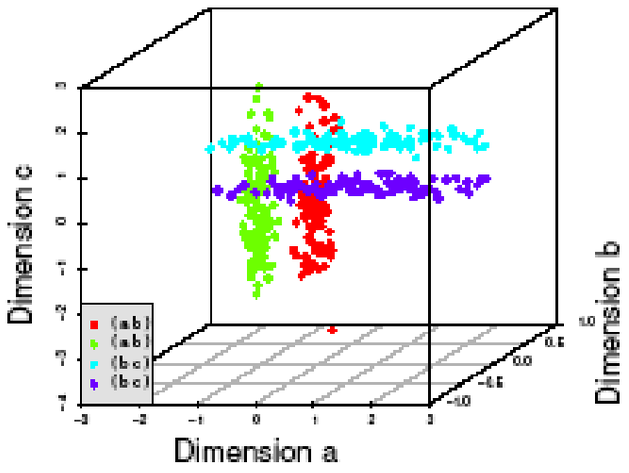
- High-dimensional data
  - e.g., text documents, DNA micro-array data
- Challenges
  - many irrelevant dimensions may mask clusters
  - distance measure dominated by noises
  - clusters may exist only in some subspaces
- Methods
  - subspace clustering, dimensionality reduction

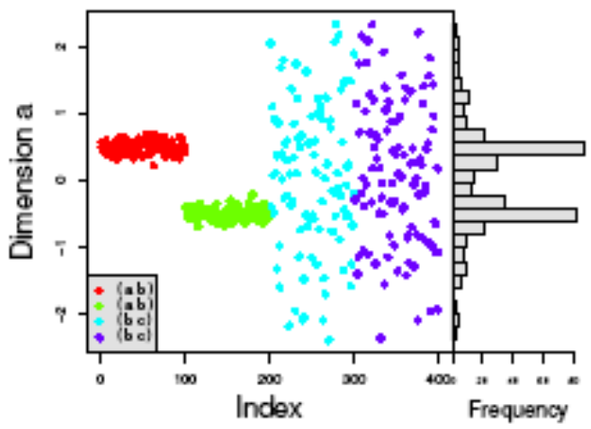
The Curse of Dimensionality

- Data in only one dimension is relatively packed
- Adding a dimension "stretch" the points across that dimension, making them further apart
- Adding more dimensions will make the points further apart
  - high-dimensional data is extremely sparse
- Distance measure becomes meaningless
  - due to equi-distance

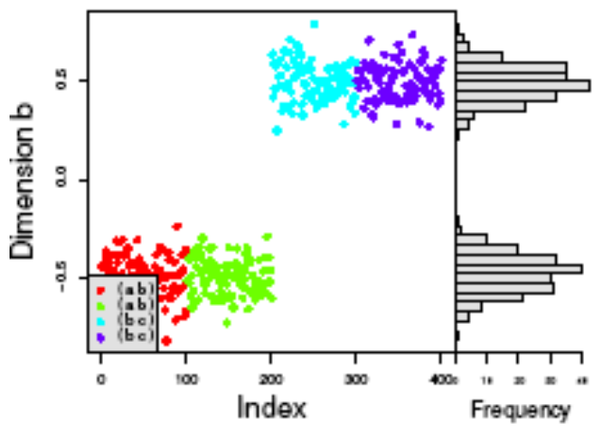
Why SubSpace Clustering?

- Clusters may exist only in some subspaces
- Subspace clustering: find clusters in all the subspaces

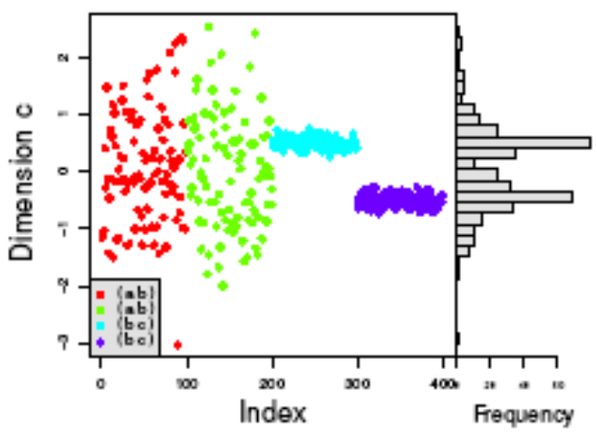




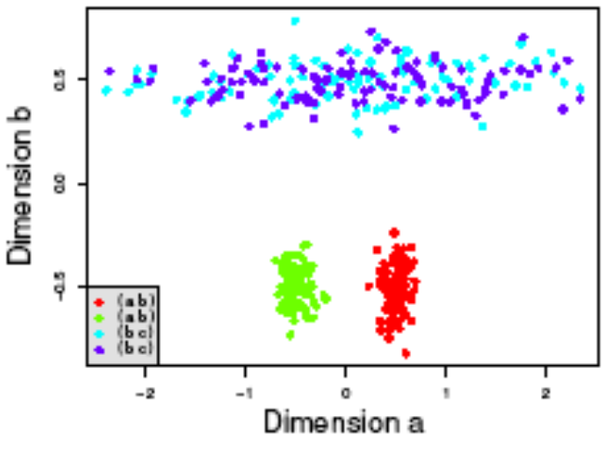
(a) Dimension a



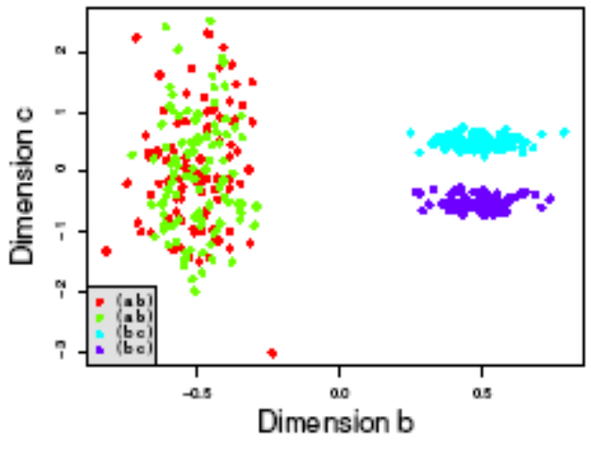
(b) Dimension b



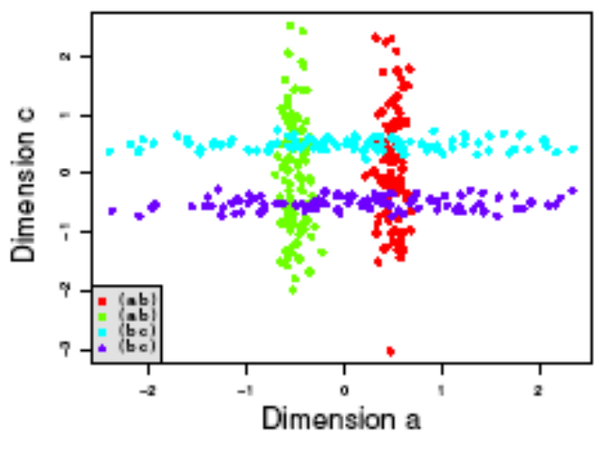
(c) Dimension c



(a) Dims a & b



(b) Dims b & c



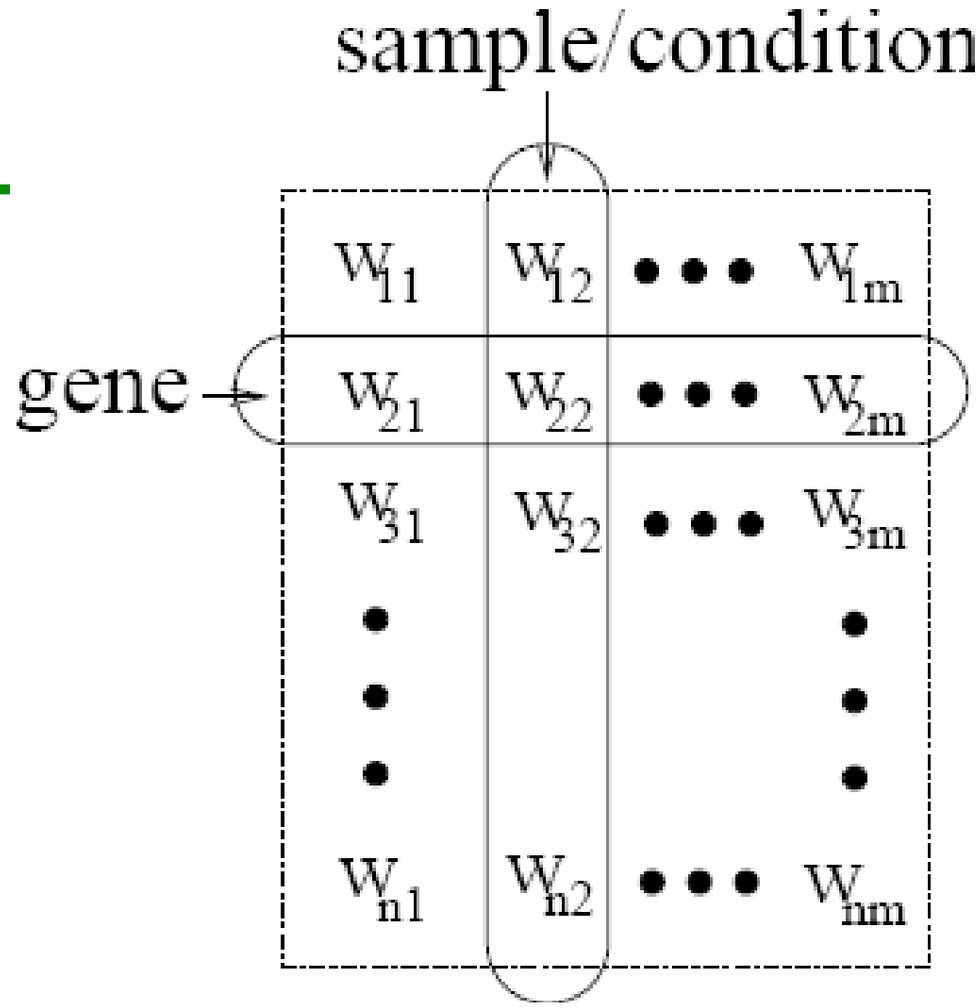
(c) Dims a & c

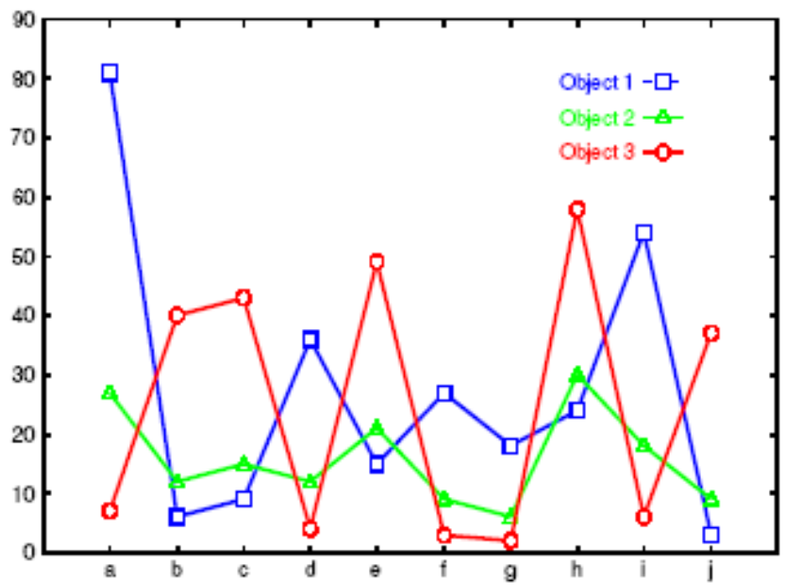
Subspace Clustering

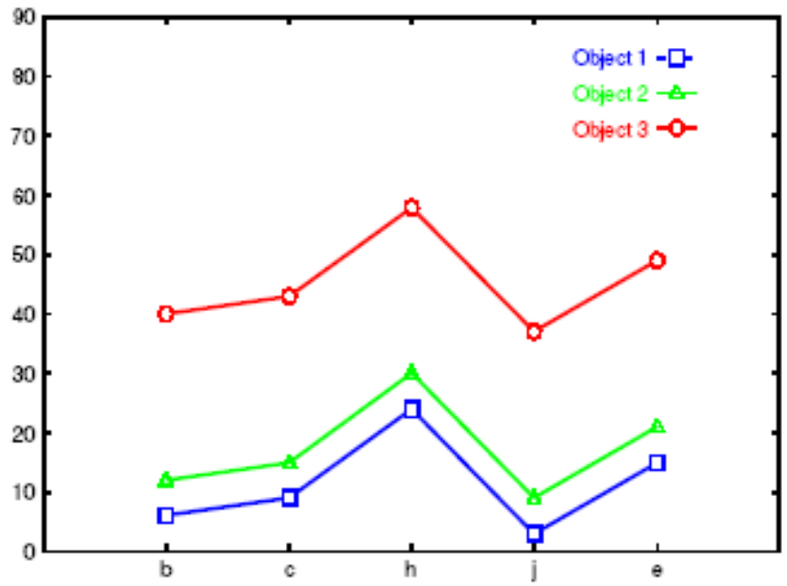
- Dimension-growth: CLIQUE
- Dimension-reduction: PROCLUS
  - mediods, subset of dimensions with small distance, k * L dimensions for k clusters
- Frequent pattern-based clustering
  - frequent term-based document clustering
  - clustering by pattern similarity in micro-array data (pClustering)

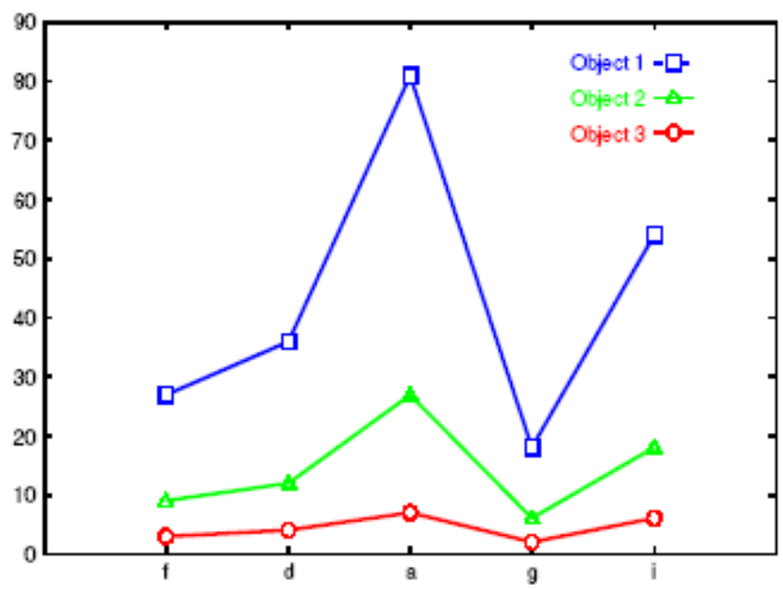
Bi-Clustering

- Cluster both objects and attributes simultaneously
- Examples
  - micro-array data analysis
  - customers and products







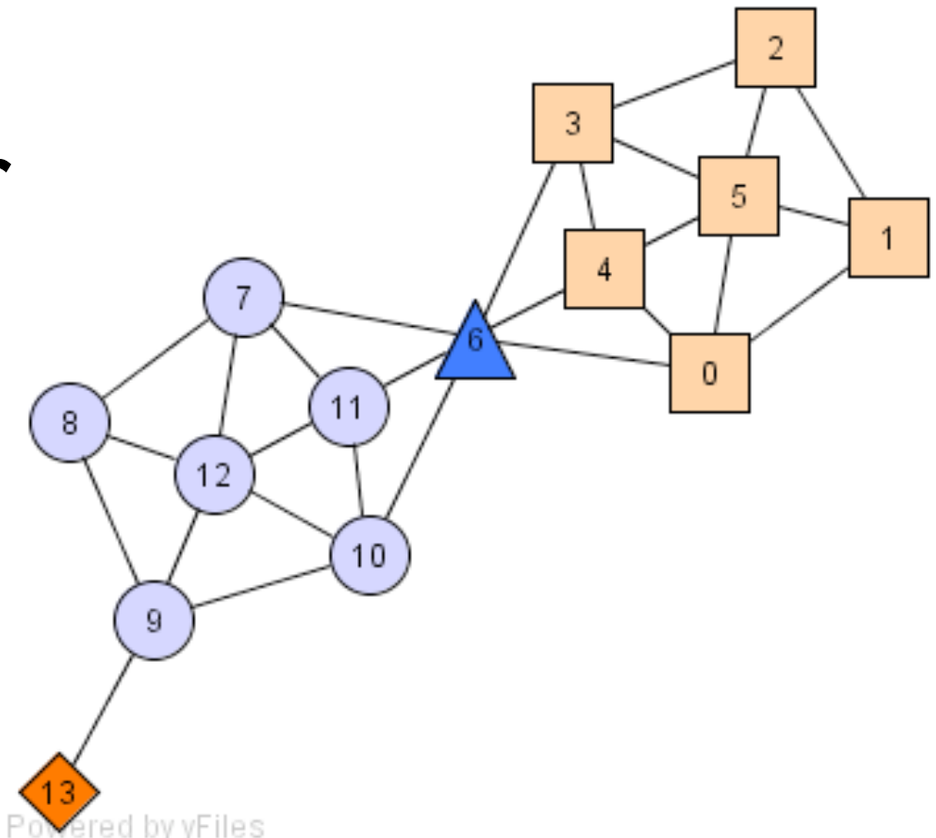


Clustering Graphs

- Applications
  - bi-partite graphs: customers and products, authors and conferences
  - web search engines: web graphs, click through graphs
  - social networks, friendship/coauthor graphs
- Graph clustering methods
  - generic clustering or graph specific (e.g., minimum cuts, density based)

Density-based Graph Clustering

- Structure-connected cluster C
  - connectivity $\forall v, w \in C : CONNECT_{\varepsilon,\mu}(v, w)$
  - maximality $\forall v, w \in V : v \in C \wedge REACH_{\varepsilon,\mu}(v, w) \Rightarrow w \in C$
- Hubs
  - not belong to any cluster
  - bridge to many clusters
- Outliers
  - connect to few clusters



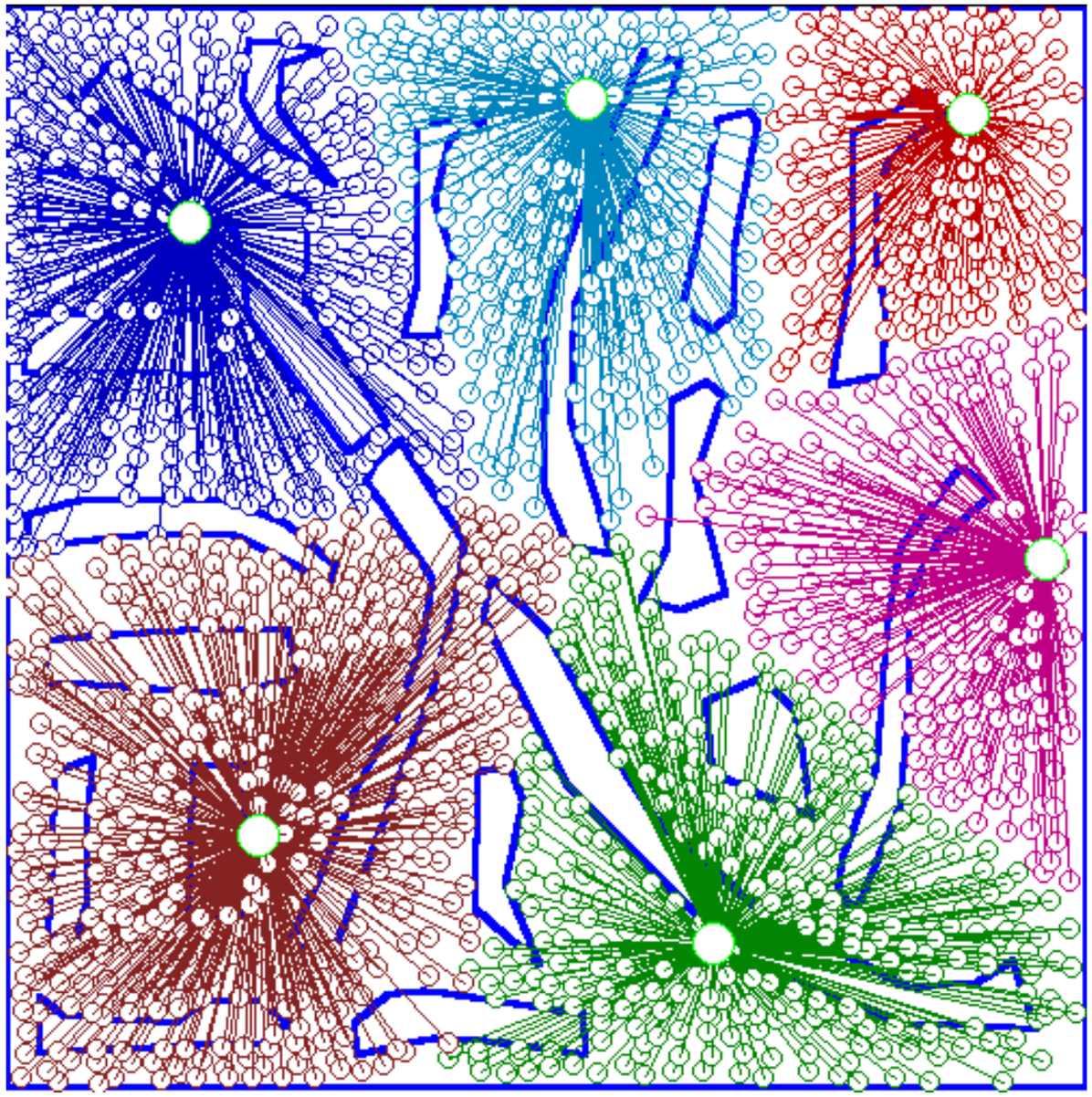
Categorization of Constraints

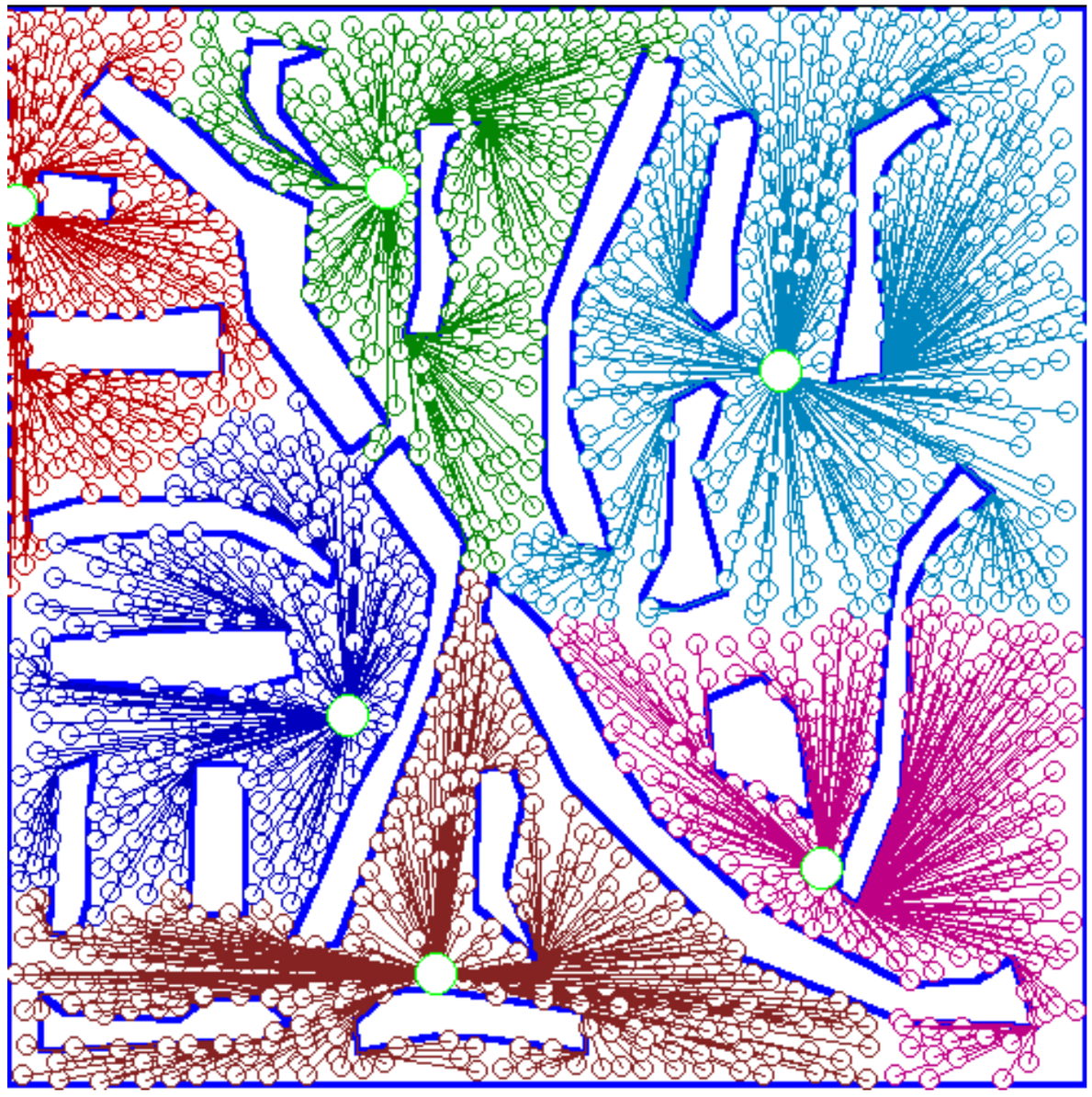
- Constraints on individual objects
  - cluster houses worth over $300K
- Constraints on the selection of clustering parameters
  - #clusters, MinPts, etc.
- Constraints on distance or similarity function
  - weighted functions
  - obstacles (e.g., river, lake)

Categorization of Constraints

- User-specific constraints on the properties of individual clusters
  - contain at least 500 valued customers and 5000 ordinary ones
- Semi-supervised clustering based on "partial" supervision
  - pairs of objects labeled as belonging to the same or different clusters

Clustering w/ Obstacle Objs





- Chapter 11: Advanced Cluster Analysis
  - probabilistic model-based clustering
  - clustering high-dimensional data
  - clustering graph and network data
  - clustering with constraints

Chapter 12:
Outlier Analysis

- Chapter 12: Outlier Analysis
  - Outlier and outlier analysis
  - Statistical approaches
  - Proximity-based approaches
  - Clustering-based approaches
  - Classification-based approaches
  - Mining contextual and collective outliers
  - Outlier detection in high dimensional

Types of Outliers

- Global outliers (point anomaly)
  - different from the rest of the data set
- Contextual outliers (conditional outlier)
  - contextual vs. behavioral attributes
    - e.g., 30-minute commute to work
- Collective outliers
  - a subset of objects
  - e.g., over 100 delayed flights

Outlier Detection Challenges

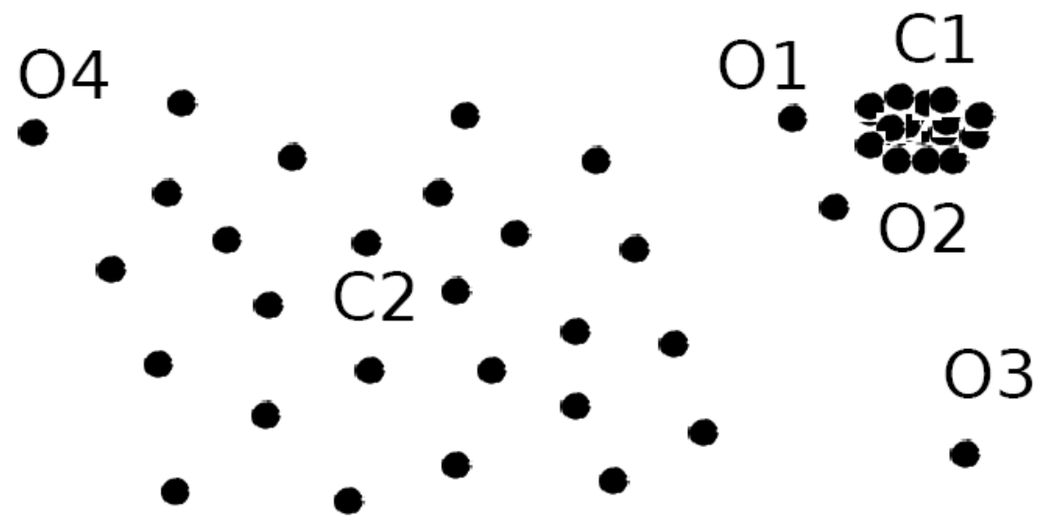
- Modeling normal objects and outliers
  - difficult to enumerate all normal cases
- Application-specific outlier detection
  - e.g., clinic data vs. marketing analysis
- Handling noise in outlier detection
  - blur the normal vs. outlier distinction
- Understandability
  - justification, degree of outlier

Outlier Detection Methods

- Whether user labels are available
  - supervised, semi-supervised, unsupervised
- Assumptions about normal data and outliers
  - statistical (model-based)
  - proximity-based
  - clustering-based

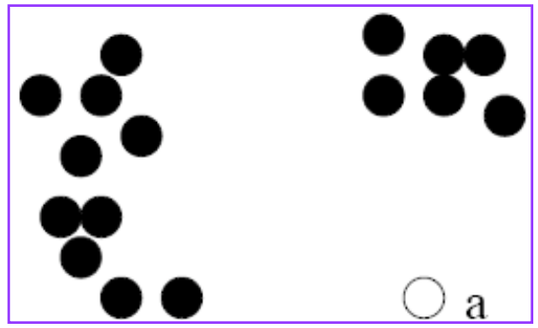
Proximity-based Approaches

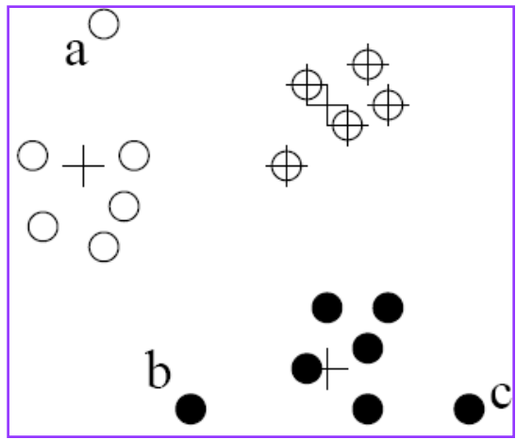
- Intuition: Objects that are far away from the others are outliers
- Distance-based
  - distance of the k-th nearest neighbor
  - or fraction of objs within distance r
- Density-based
  - density around outlier
  - density of neighbors



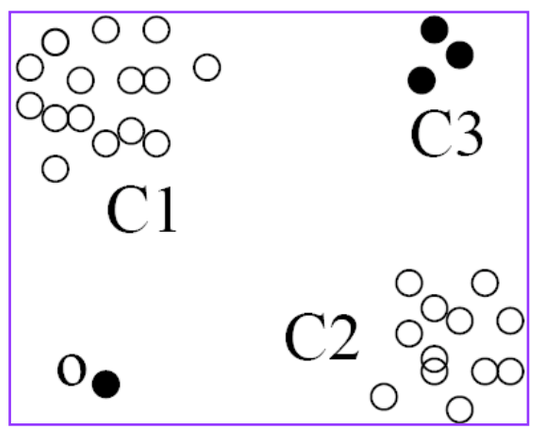
Clustering-based Approaches

- An object is outlier if
  - does not belong to any cluster
  - far from its closest cluster
  - belongs to a small or sparse cluster





- Using training set to find patterns of normal data
- Compare new obj w/ clusters

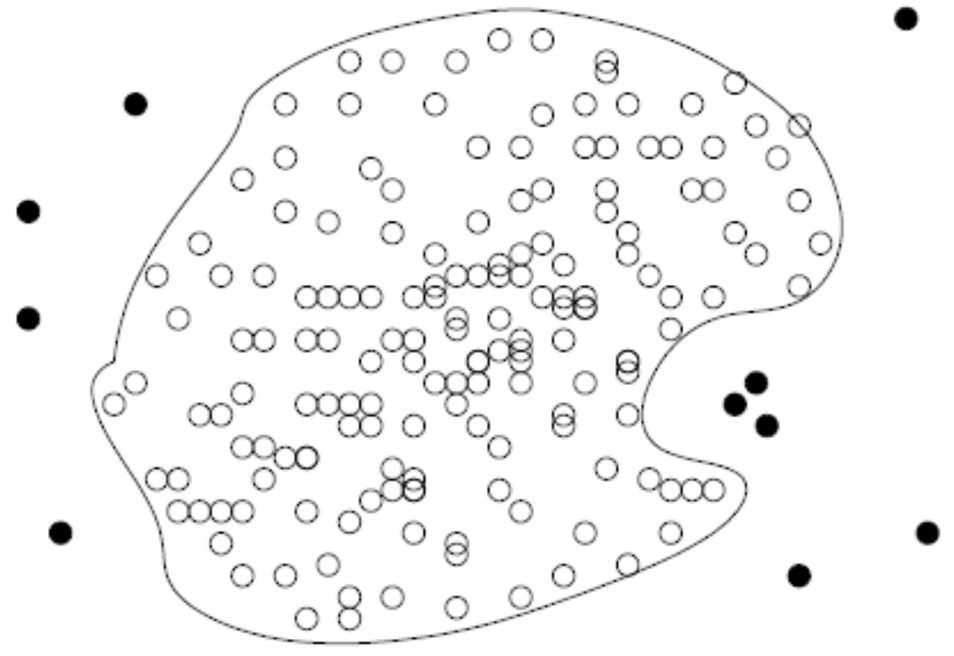


Clustering-based Approaches

- Strength
  - no labels required, support many data types
  - clusters are summaries of data
  - only need to compare obj to clusters (fast)
- Weakness
  - effectiveness depends on clustering method
  - high computation cost: first find clusters
  - to reduce cost: e.g., fixed-width clustering

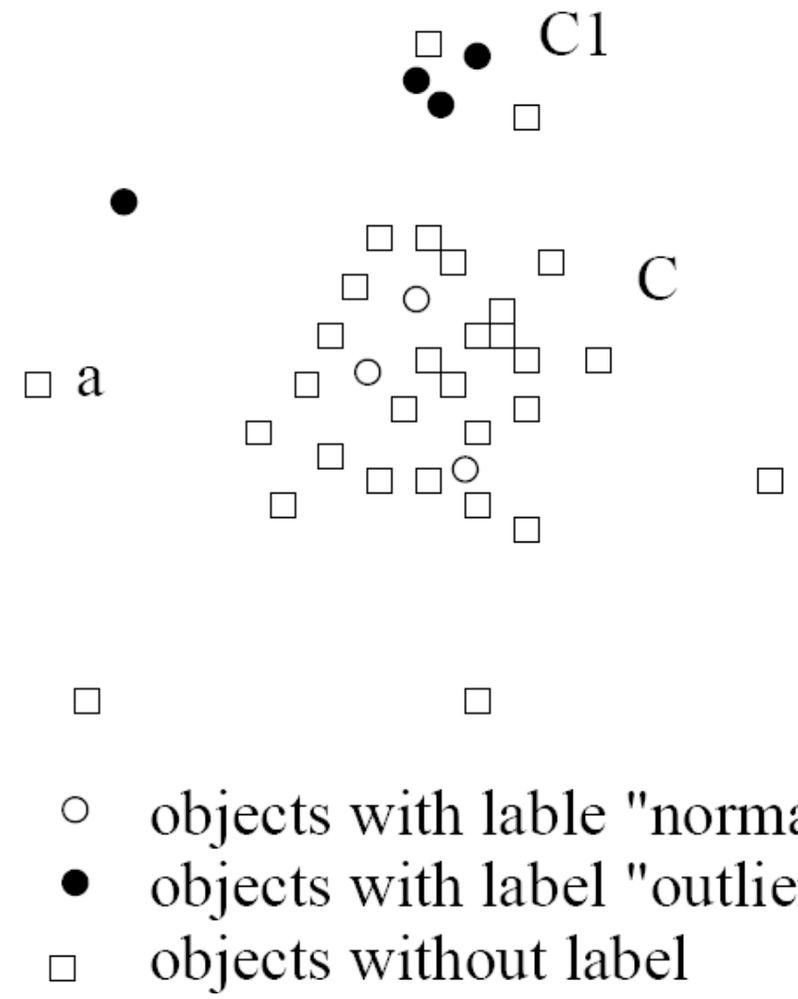
Classification-based Methods

- Train a classification model that can distinguish normal data from outliers
- Brute-force approach
  - labeled normal and outlier objects
  - skewed data, cannot detect unseen anomaly
- One-class model
  - classifier for normal data



Semi-supervised Learning

- Combine classification and clustering
- Clustering to find large cluster C and small cluster C1
- C: objs w/ normal label
- One-class model from C
- C1: obs w/ outlier label
- Any obj not in C's model
  - identify as outlier



Contextual Outliers (1)

- Transform to conventional outlier detection if the contexts can be clearly identified
  - identify the context of o
  - compare o w/ others in the same context
  - use conventional outlier detection
- Generalize context if too few objects in the same context
  - e.g., customers by age group
  - e.g., by geographic region

Contextual Outliers (2)

- Modeling normal behavior with respect to contexts
- Contextual attributes => behavior attributes
  - mixture model U on contextual attributes
  - mixture model V on behavior attributes
  - p(V_i|U_j)
- Approaches
  - regression, Markov models, finite state automaton

Collective Outliers (1)

- Objects as a group deviate significantly from the entire data set
- Need to examine structures, i.e., relationships between multiple data objects
  - e.g., temporal, spatial, graph/network
- Different from contextual outlier detection
  - structures are often not explicitly defined
- Reduce to conventional outlier detection
  - structure unit ==> "super" data object

Collective Outliers (2)

- Model the expected behavior of structure unites directly
- E.g., online social networks of customers
  - structure unit: subgraphs of the network
    - small subgraphs with very low frequency
    - large subgraphs that are surprisingly frequent
- E.g., temporal sequences
  - learn a Markov model from the sequences; a subsequence that deviates from the model

Outliers in High-D Data

- Data in high-D spaces are often sparse
  - object distances dominated by noise
- Interpretation of outliers
  - too many features, need to identify subspaces of the outliers
- Subspaces that signifying the outliers
  - capture the local behavior of data
- Scalable: #subspaces increases exponentially

- Chapter 12: Outlier Analysis
  - Outlier and outlier analysis
  - Statistical approaches
  - Proximity-based approaches
  - Clustering-based approaches
  - Classification-based approaches
  - Mining contextual and collective outliers
  - Outlier detection in high dimensional data