資料分析與學習基石
(*Fundamental of Data Analytics and Learning*)

-- Unsupervised Learning (1/2)

Hung-Yu Kao (高宏宇)
***I**ntelligent **K**nowledge **M**anagement Lab*

Master Program of Artificial Intelligence
Institute of Medical Informatics,
Dept. of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan

Unsupervised Learning

- Learning without a teacher
- Self-organization

Clustering
- K-mean
- Hierarchical clustering
- DBSCAN
- …

Anomaly Detection
- Outlier detection

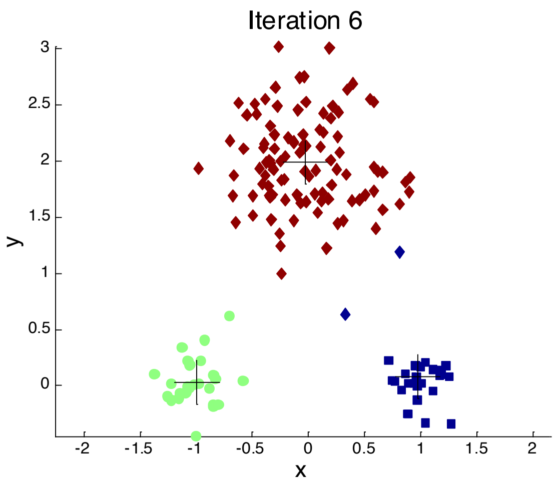
Neural Network
- Autoencoder
- Generative Adversarial Network (GAN)
- SOM

Learning approach
- Expectation Maximization (EM)
- PCA, MF, SVD

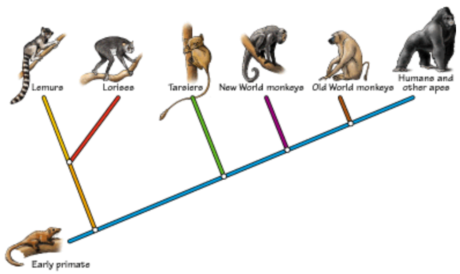
Clustering

- Clustering: process of grouping a set of physical or abstract objects into classes of similar objects
- Cluster: a collection of data objects that
  - are similar to one another within the same cluster
  - are dissimilar to the objects in other clusters
- Clustering: unsupervised classification
  - supervised classification: known #cluster & cluster labels
  - unsupervised classification: unknown #cluster & cluster labels

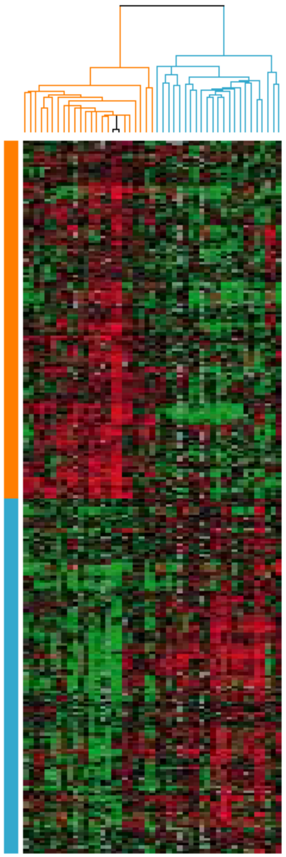
Cluster



Iteration 6
3
2.5
2
1.5
1
0.5
0
y
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
x



Lemurs
Lorises
Tarsiers
New World monkeys
Old World monkeys
Humans and other apes
Early primate



Good Clustering

- Good clustering (produce high quality clusters)
  - intra-cluster similarity is high
  - inter-cluster class similarity is low
- Quality factors
  - similarity measure and its implementation
  - definition and *representation* of cluster chosen
  - clustering algorithm

What is Similarity?

Similarity is hard to define, but…   "We know it when we see it"

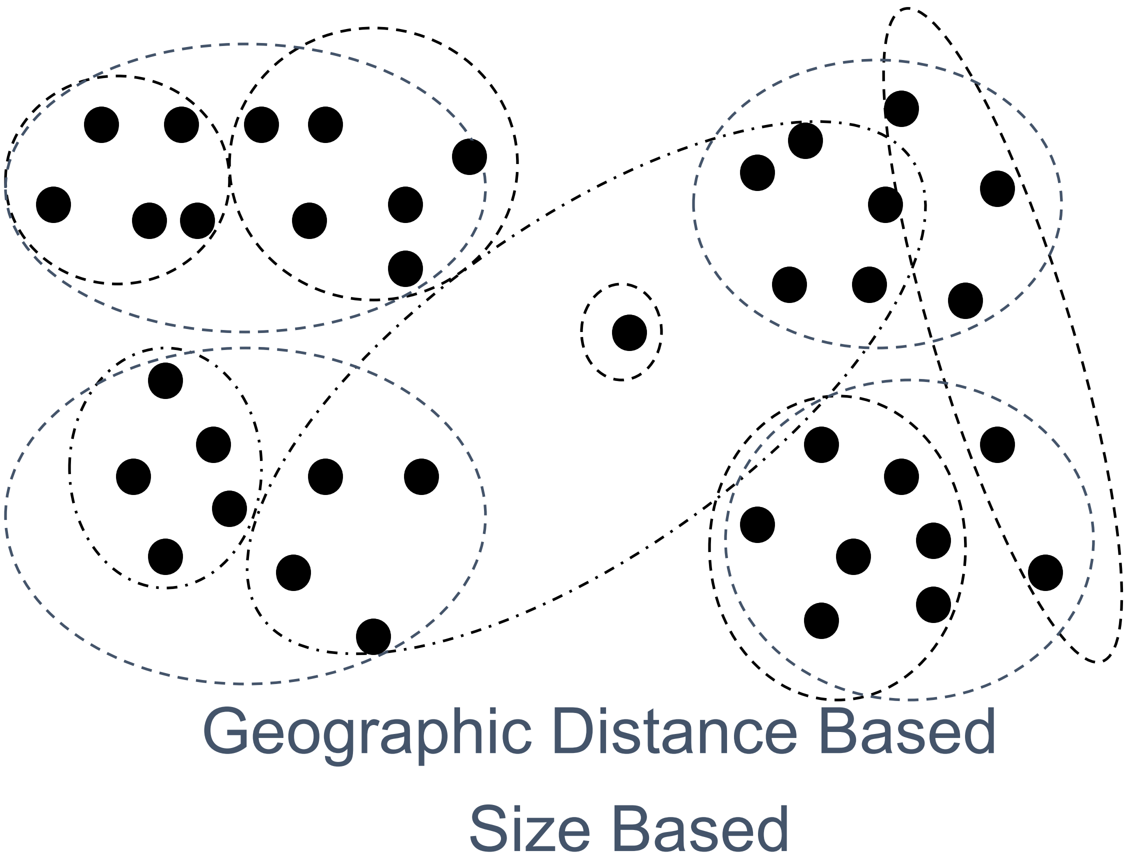
Typical Applications of Clustering Analysis

- Pattern Recognition
- Business: market segmentation
  - discover distinct group of customers
  - characterize customer groups
- Biology:
  - derive plant & animal taxonomies
  - categorizes genes with similar functionality
  - gain insight into structures inherent in populations
- Geography:
  - identification of area of similar land use
- Insurance:
  - identification of groups of insurance holders with high claim cost
- City-planning: identification of house group
- Document management: classify documents of WWW

Requirements of Clustering

- Scalability
- Dealing with different types of attributes (not only numerical data)
- Discovery of clusters with arbitrary shape (not only sphere)
- Minimal requirements for domain knowledge to input design parameters
- Ability to deal with noisy data
- Insensitivity to order of input records
- High dimensionality
- Constraint-based clustering
- Interpretability and usability

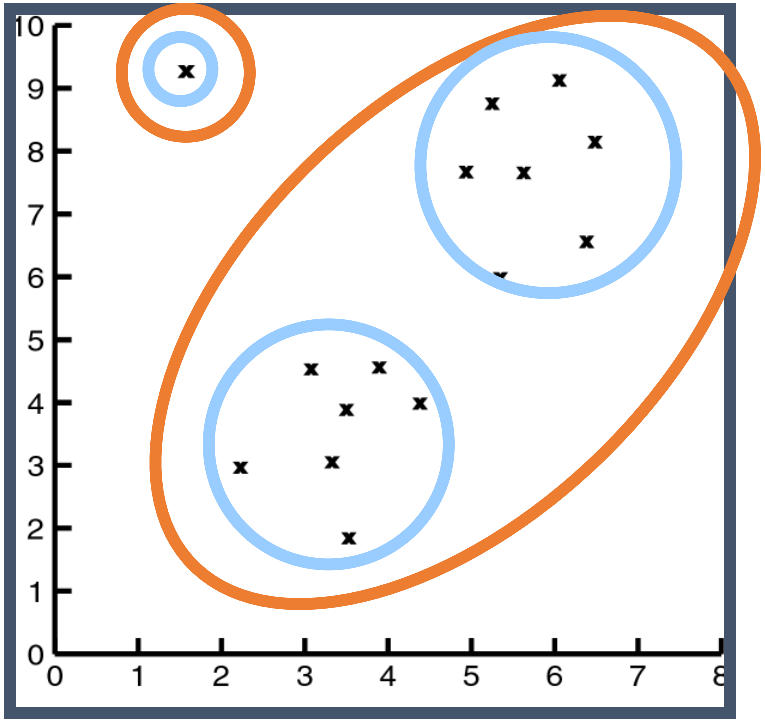
Clustering Houses



Clustering Issues

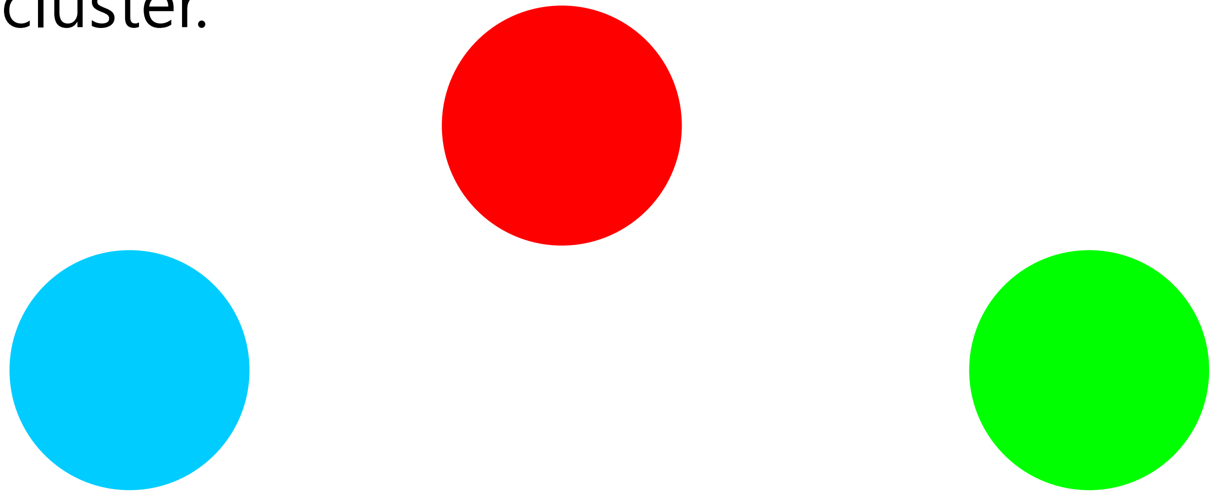
- Outlier handling
- Dynamic data
- Interpreting results (centroid meaning)
- Number of clusters (magic *k*)
- Data to be used
- Scalability

Impact of Outliers on Clustering



Types of Clusters: Well-Separated

- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

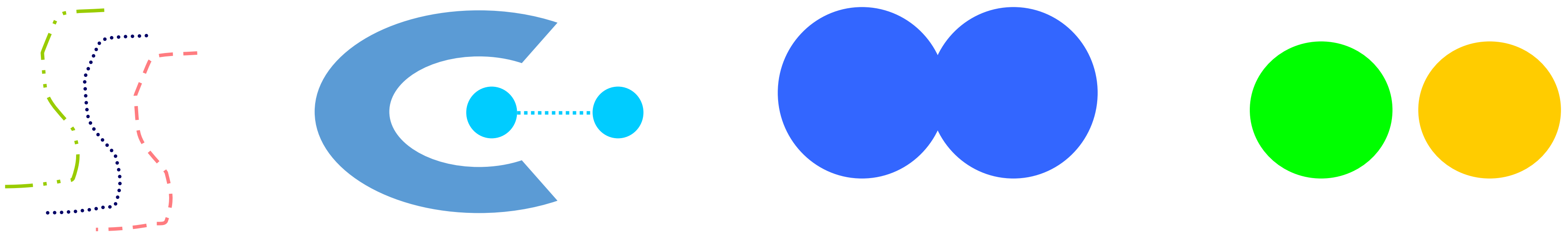
Types of Clusters: Center-Based

- Center-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster

4 center-based clusters

Types of Clusters: Contiguity-Based

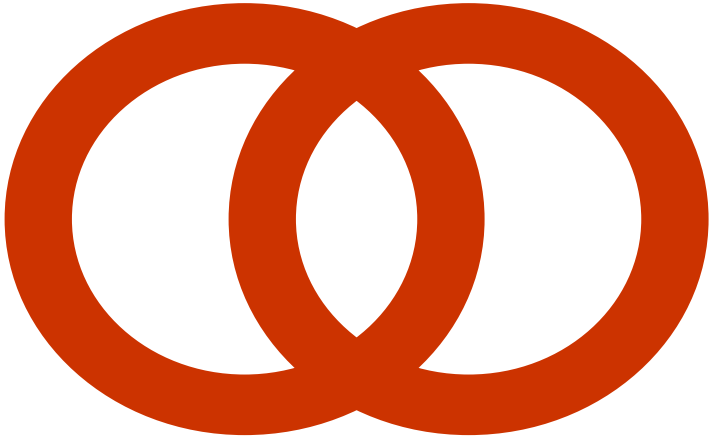
- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



8 contiguous clusters

Types of Clusters: Conceptual Clusters

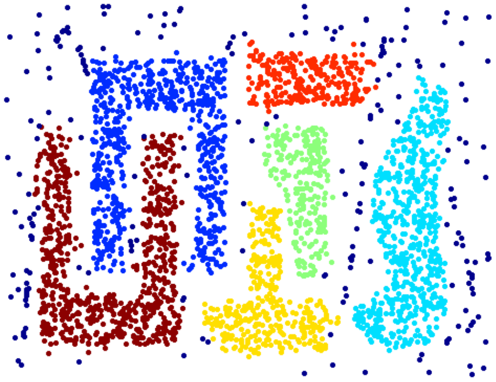
- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.

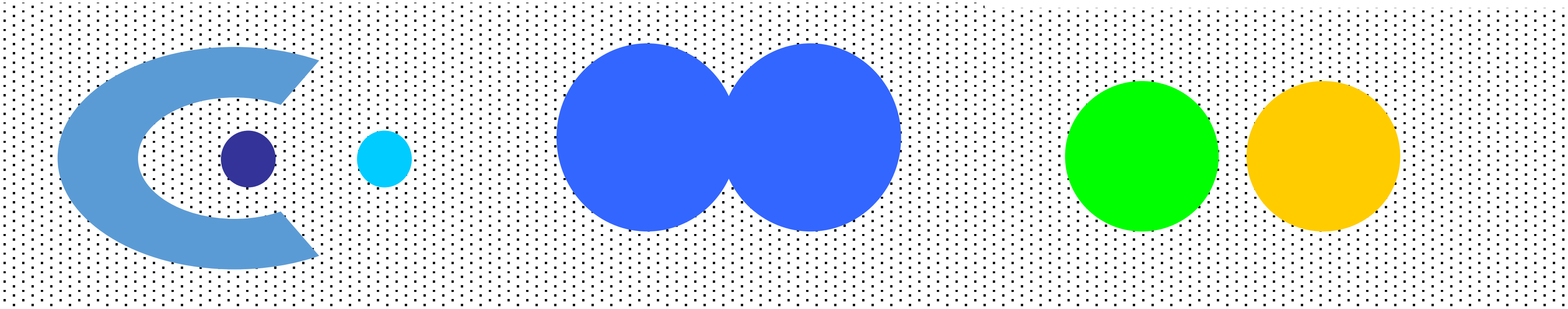


2 Overlapping Circles

Types of Clusters: Density-Based

- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.





6 density-based clusters

Approaches of Clustering Algorithms

Five Categories of Clustering Methods

- Partitioning algorithms
  - Construct various partitions and then evaluate them by some criterion.
- Hierarchy algorithms
  - Create a hierarchical decomposition of the set of data (or objects) using some criterion.
- Density-based
  - based on connectivity and density functions
- Grid-based
  - based on a multiple-level granularity structure
- Model-based
  - A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other.

Partition-based Clustering

Partitioning Algorithms: Basic Concept

- Partitioning method: Construct a partition of a database *D* of *n* objects into a set of *k* clusters
- Given a *k*, find a partition of *k* clusters that optimizes the chosen partitioning criterion.
  - Global optimal: exhaustively enumerate all partitions. (NP-Hard!)
  - Heuristic methods.
    - k-means: each cluster is represented by the center of the cluster
    - k-medoids or PAM (Partition Around Medoids) : each cluster is represented by one of the objects in the cluster.

The *K*-Means Clustering Method

- Given *k*, the *k-means* algorithm:
  1. Partition objects into *k* nonempty subsets
  2. Compute mean as the centroids of the clusters of the current partition
  3. Relocate each object to the nearest cluster
  4. Go back to Step 2, stop when no more new relocation

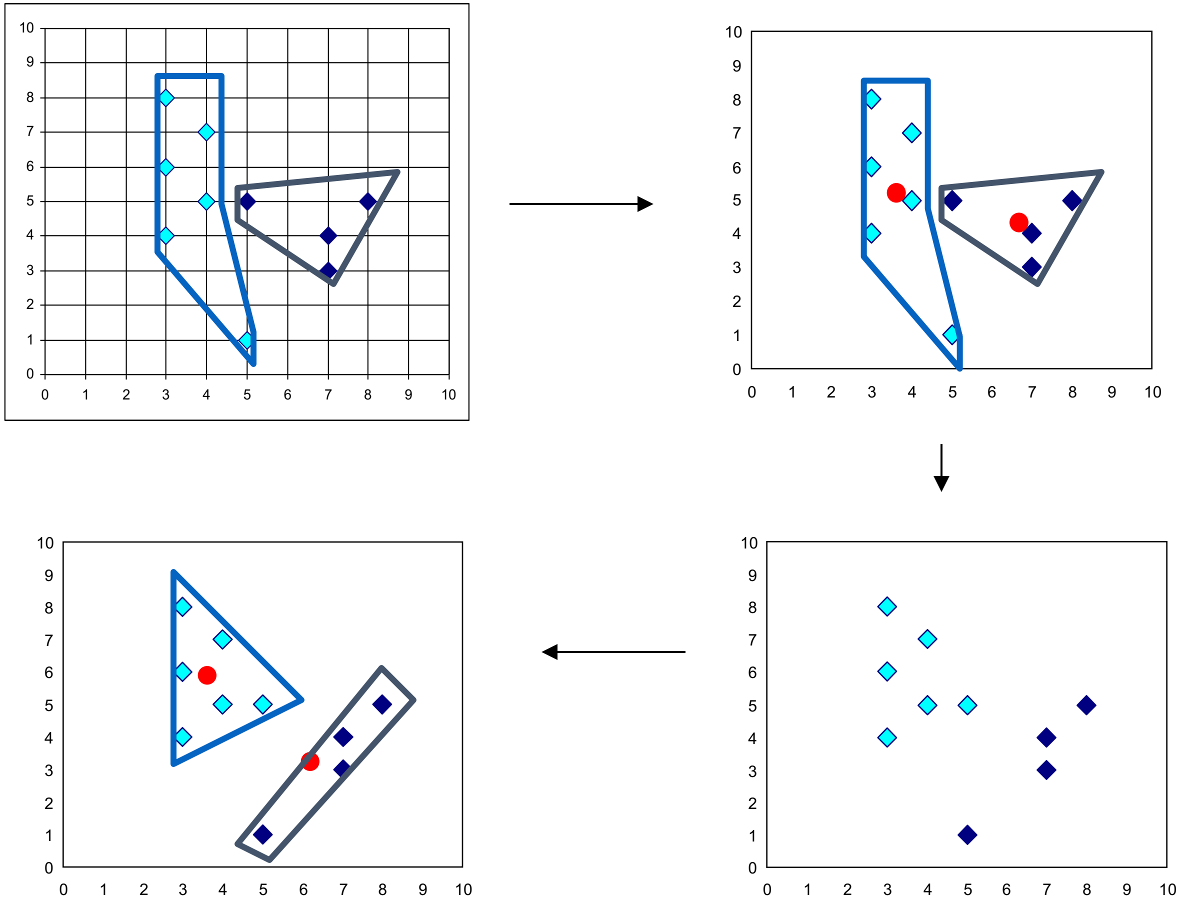
K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, k, must be specified
- The basic algorithm is very simple

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-Means Clustering Method

- Example



K-Means Example

- Given: {2,4,10,12,3,20,30,11,25}, k=2

C1	C2	M1	M2
{2,3}	{4,10,12,20,30,11,25}	2.5	16
{2,3,4}	{10,12,20,30,11,25}	3	18
{2,3,4,10}	{12,20,30,11,25}	4.75	19.6
{2,3,4,10,11,12}	{20,30,25}	7	25
{2,3,4,10,11,12}	{20,30,25}	7	25

K-means Clustering – Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- 'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.
- K-means will converge for common similarity measures mentioned above.
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to 'Until relatively few points change clusters'

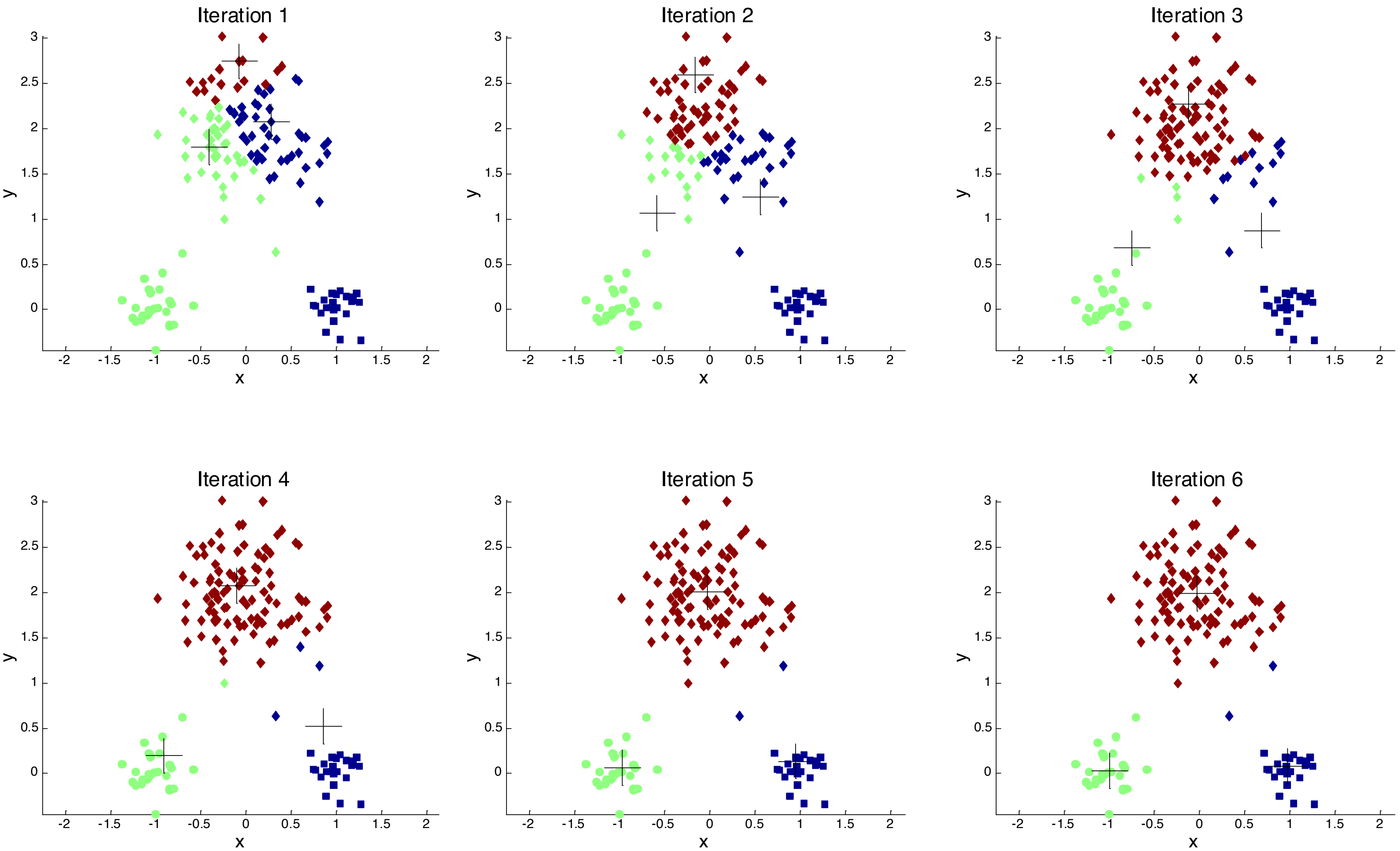
Comments on the *K-Means* Method

- Strength
  - Efficient, Complexity is O(n * K * I * d)
    - n = number of points, K = number of clusters, I = number of iterations, d = number of attributes
  - Often terminates at a *local optimum*
- Weakness
  - Applicable only when *mean* is defined (categorical data?)
  - Need to specify *k,* the *number* of clusters, in advance
  - Unable to handle noisy data and *outliers*

Variations of the *K-Means* Method

- Variants of the *k*-means
  - Selection of the initial *k* means
  - Dissimilarity calculations
  - Strategies to calculate cluster means
- Handling categorical data: *k-modes*
  - Replacing means of clusters with modes (distance=0 or 1)
  - *k-prototype*: a mixture of categorical and numerical data

Importance of Choosing Initial Centroids

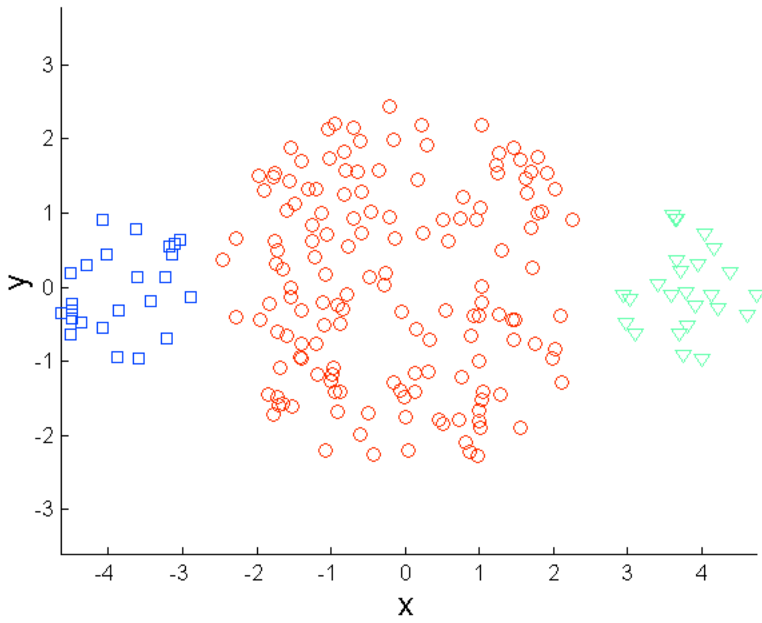


Limitations of K-means

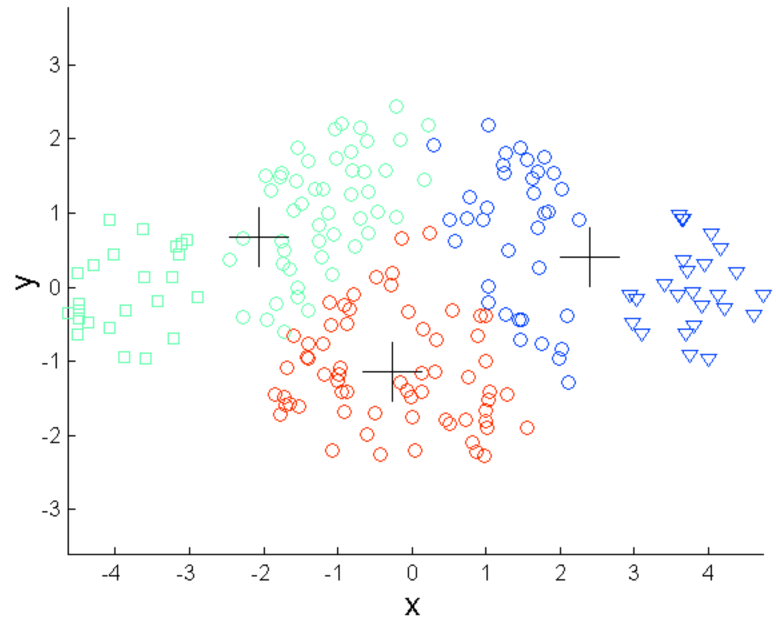
- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes

- K-means has problems when the data contains outliers.

Limitations of K-means: Differing Sizes

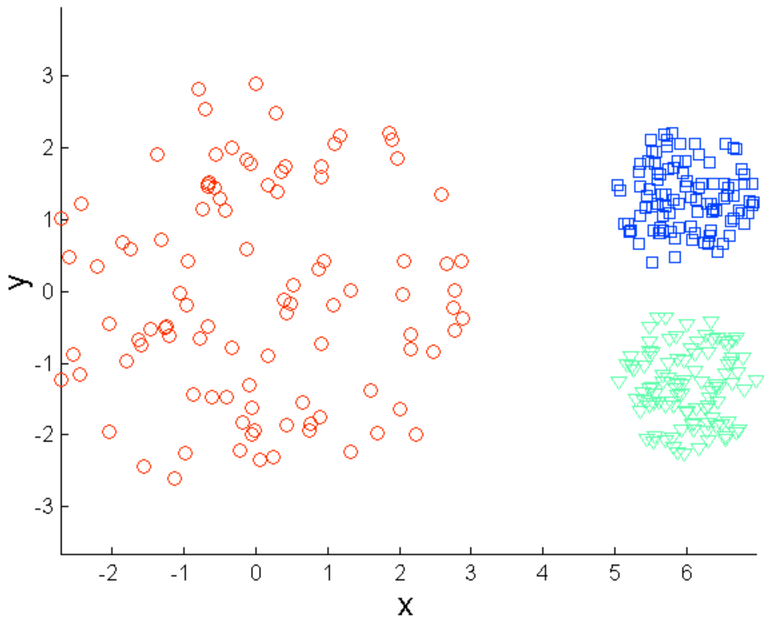


Original Points

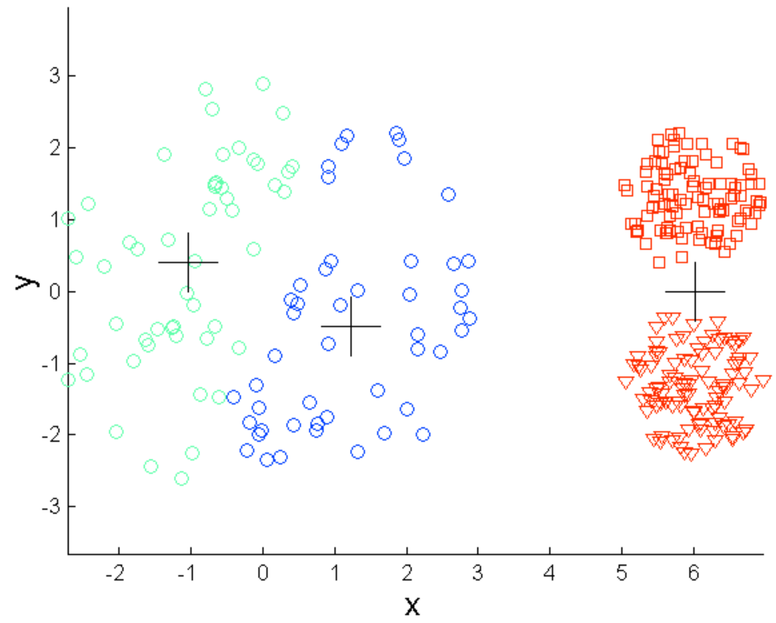


K-means (3 Clusters)

Limitations of K-means: Differing Density

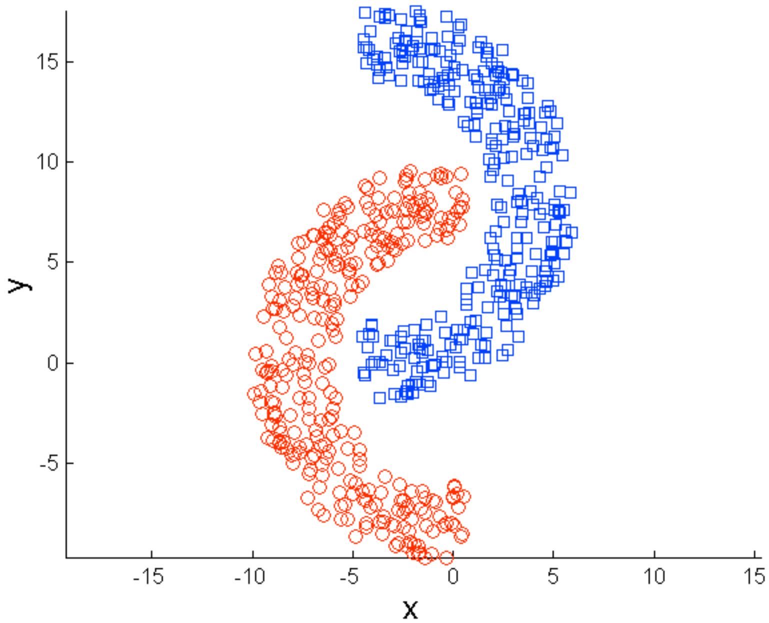


Original Points

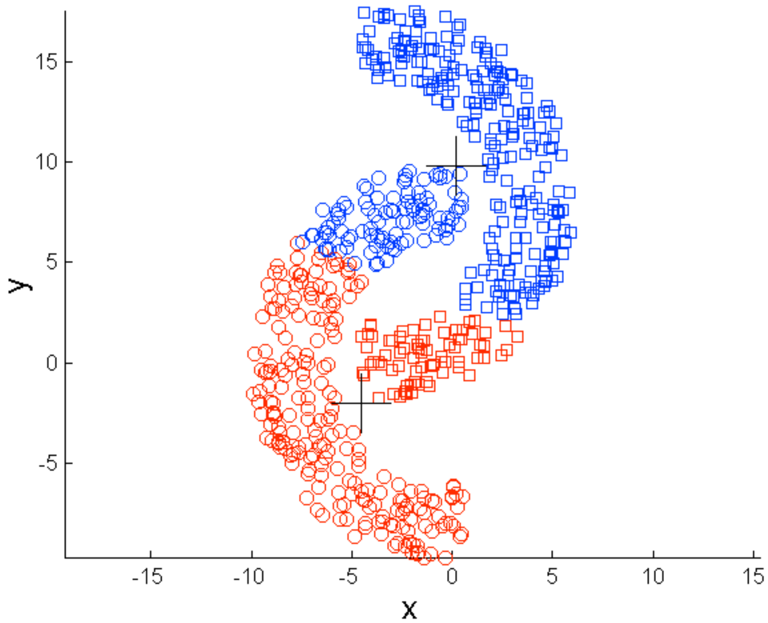


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes



Original Points

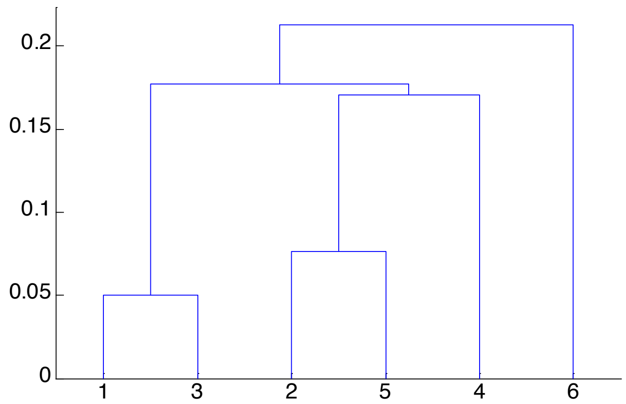


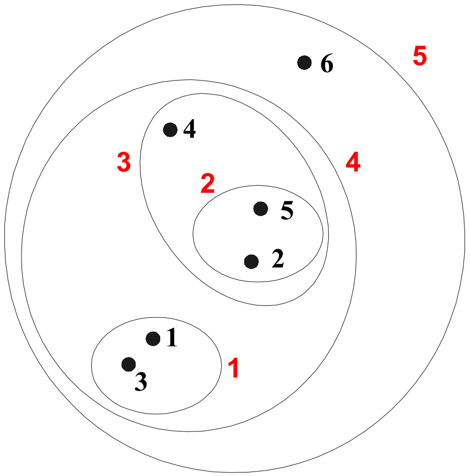
K-means (2 Clusters)

Hierarchical Clustering

Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits

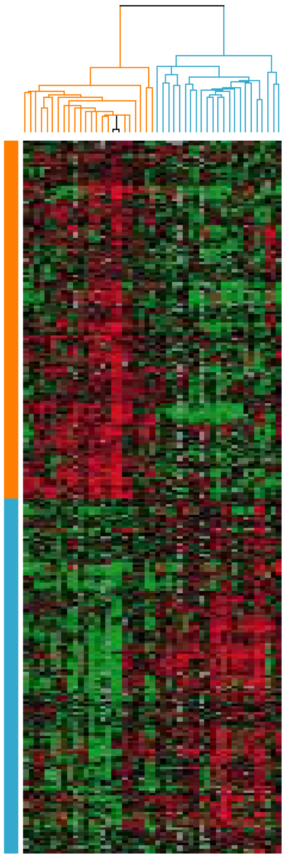


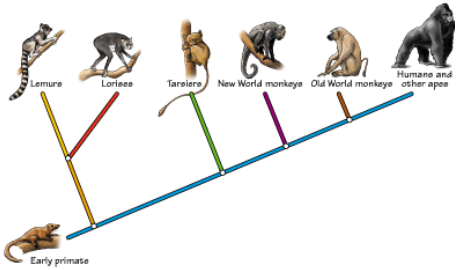


Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by ‘cutting’ the dendogram at the proper level

- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, …)

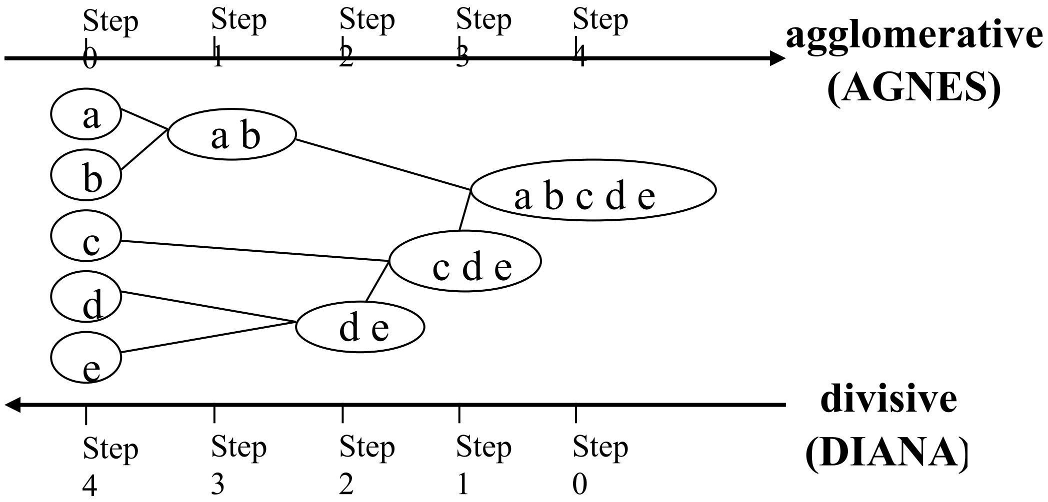


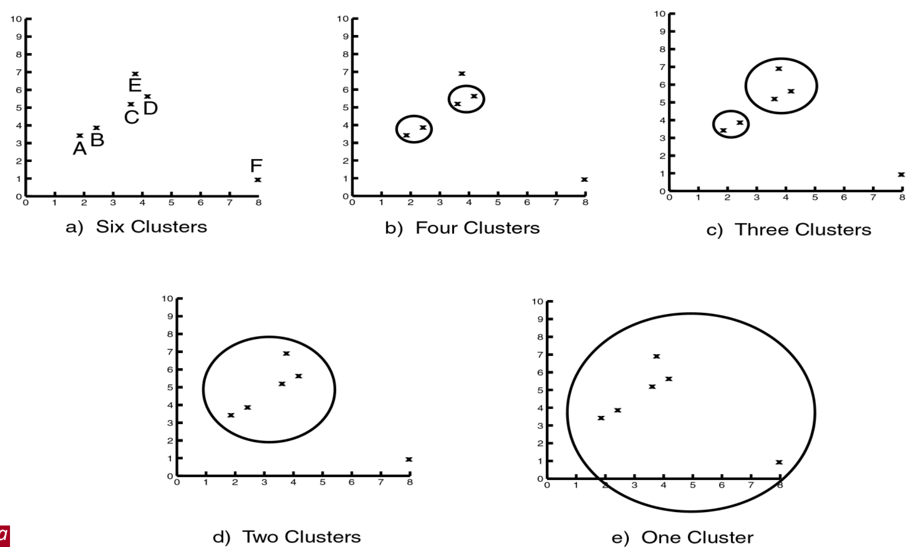


Hierarchical Clustering

- Two main types of hierarchical clustering
  - Agglomerative:
    - Start with the points as individual clusters
    - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left
  - Divisive:
    - Start with one, all-inclusive cluster
    - At each step, split a cluster until each cluster contains a point (or there are k clusters)
- Traditional hierarchical algorithms use a similarity or distance matrix
  - Merge or split one cluster at a time

Hierarchical Clustering



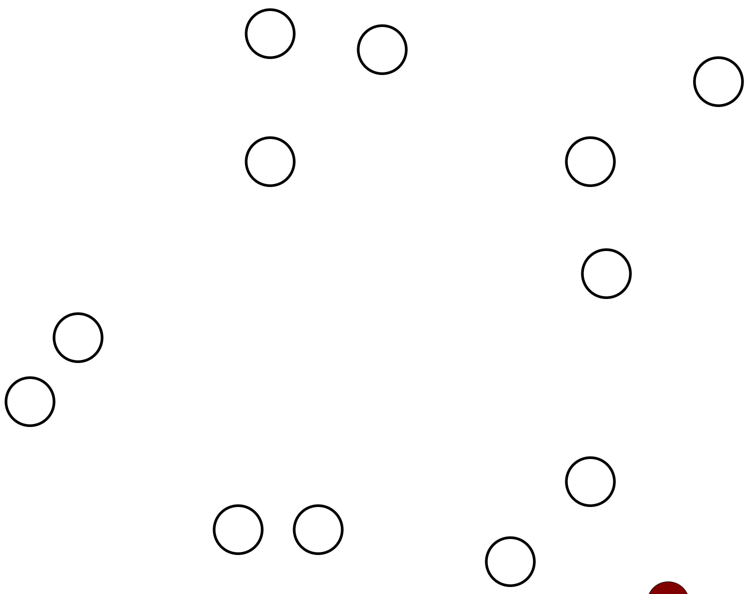


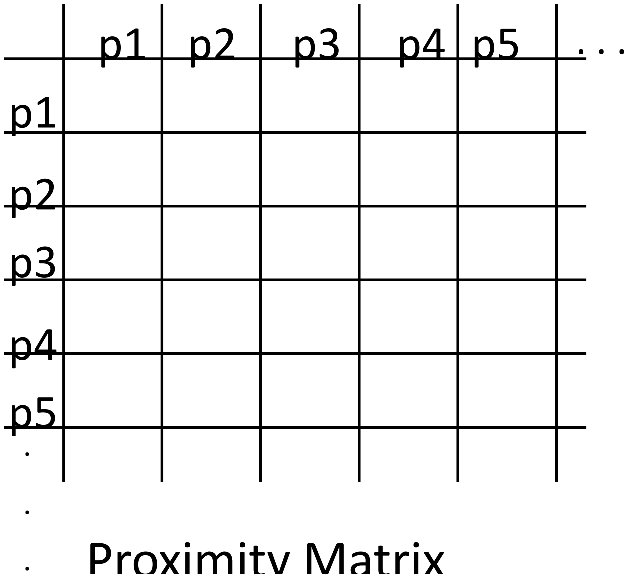
Agglomerative Clustering Algorithm

- More popular hierarchical clustering technique
- Basic algorithm is straightforward
  1. Compute the proximity matrix
  2. Let each data point be a cluster
  3. **Repeat**
  4. Merge the two closest clusters
  5. Update the proximity matrix
  6. **Until** only a single cluster remains
- Key operation is the computation of the proximity of two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms

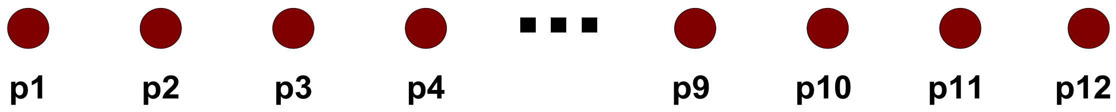
Starting Situation

- Start with clusters of individual points and a proximity matrix



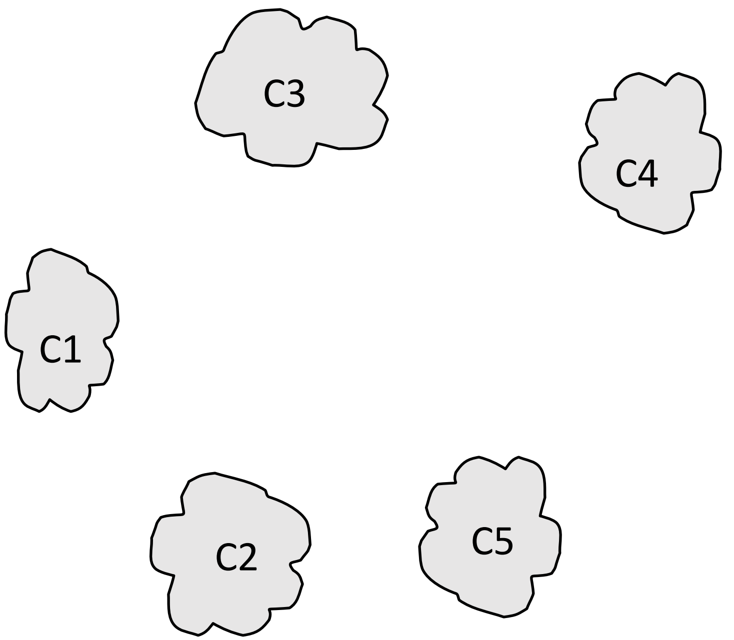


Proximity Matrix



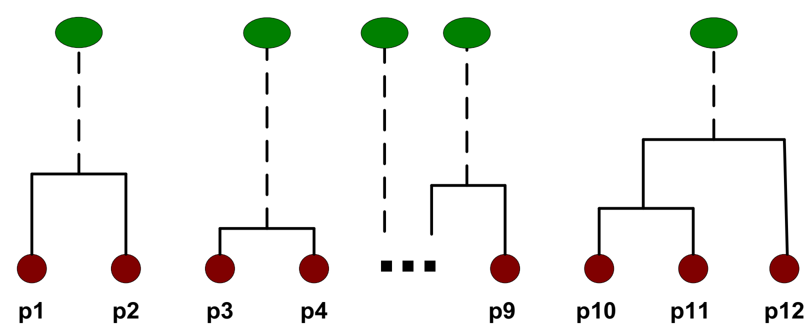
Intermediate Situation

- After some merging steps, we have some clusters



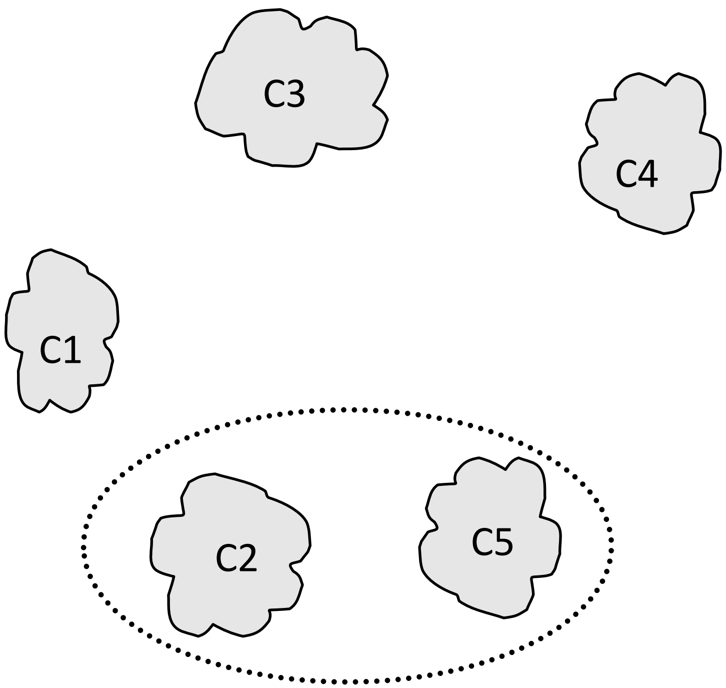
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

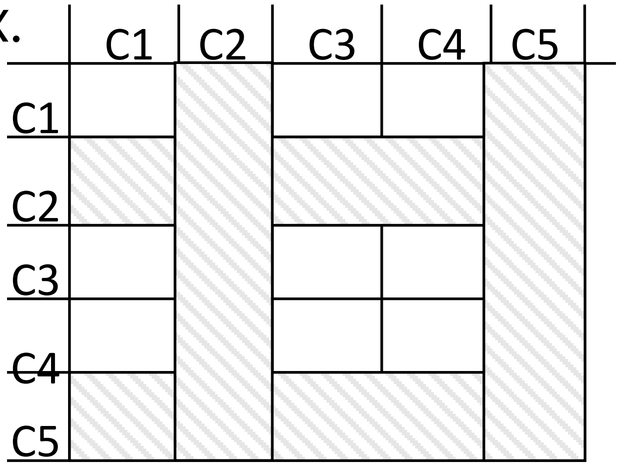
Proximity Matrix



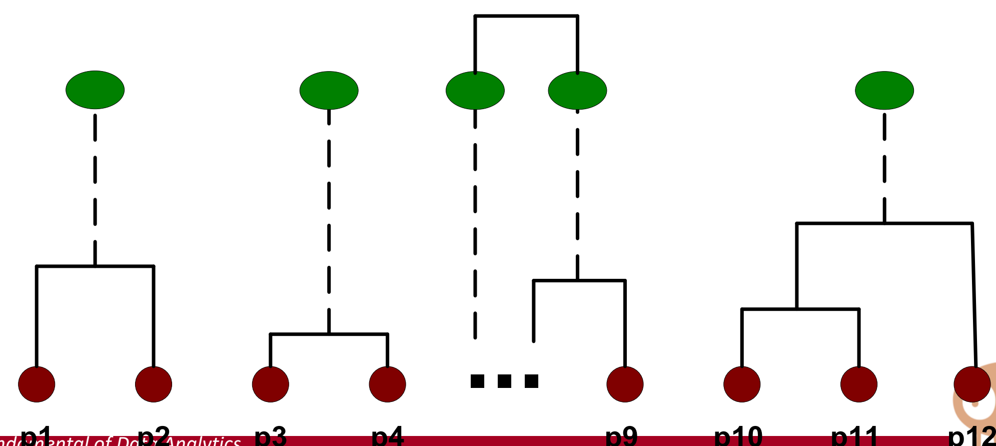
Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



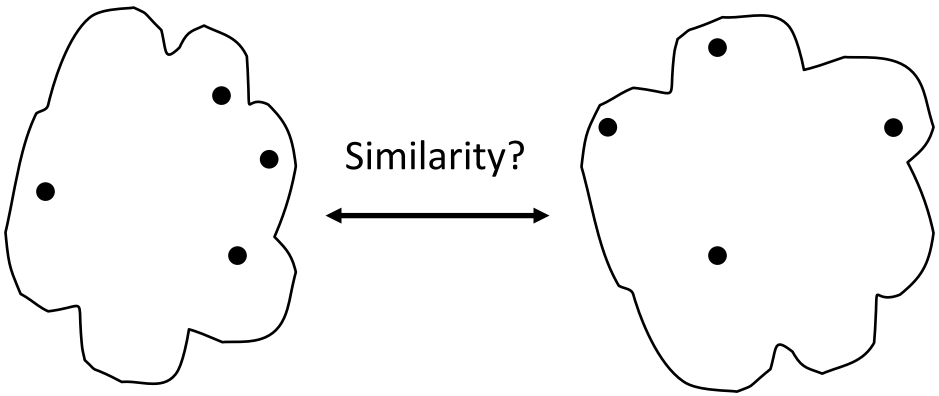


Proximity Matrix



The question is "How do we update the proximity matrix?"

How to Define Inter-Cluster Similarity

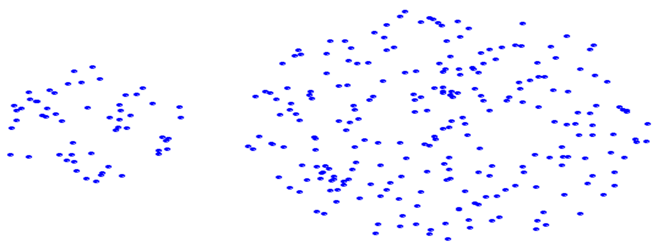


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

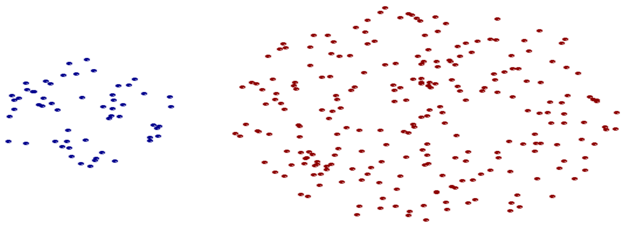
Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Strength/Limitation of MIN

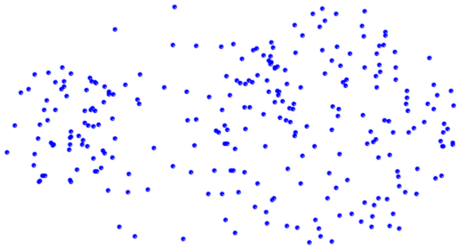


Original Points

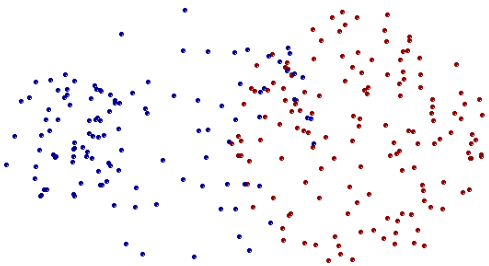


Two Clusters

- Can handle non-elliptical shapes



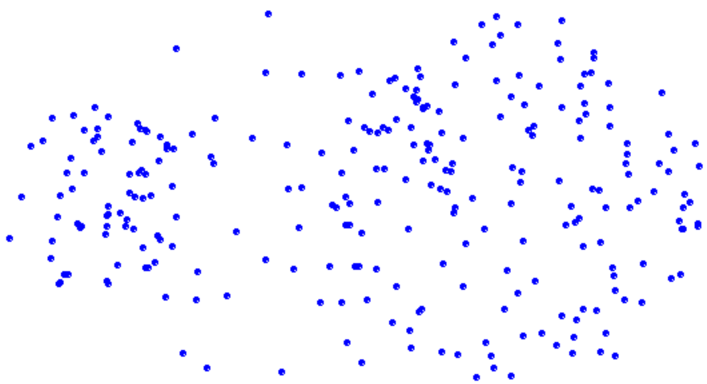
Original Points



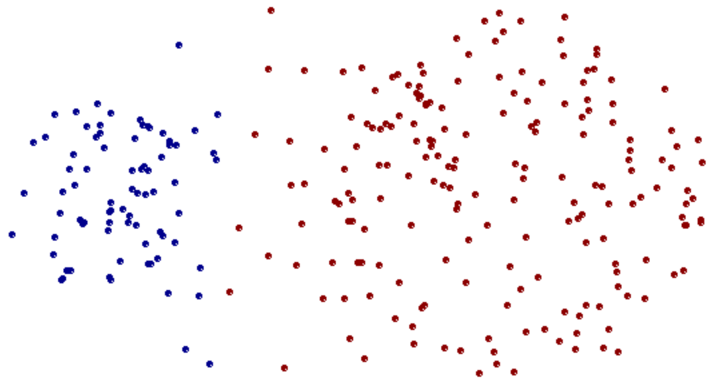
Two Clusters

- Sensitive to noise and outliers

Strength of MAX



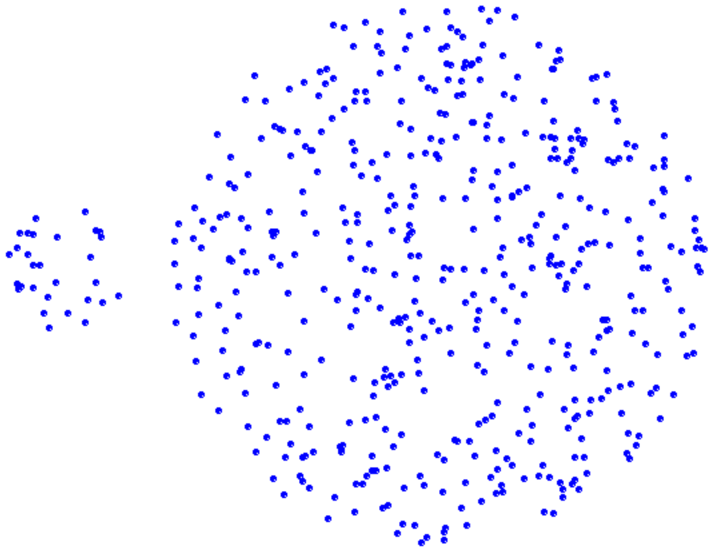
Original Points



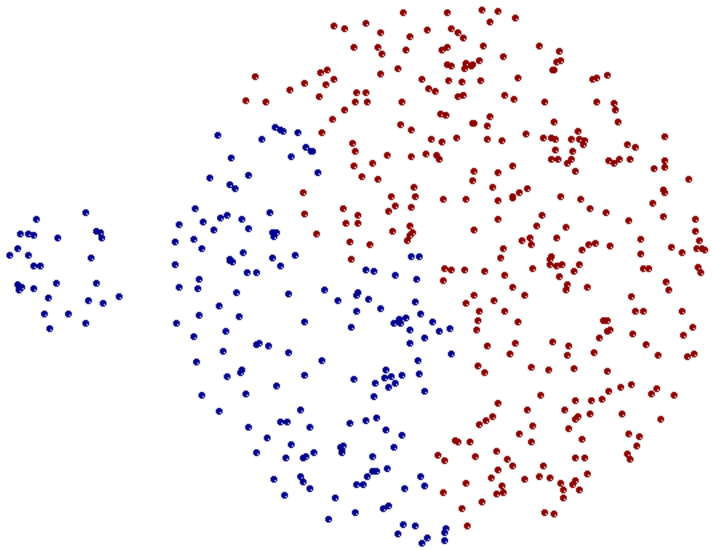
Two Clusters

- Less susceptible to noise and outliers

Limitations of MAX



Original Points



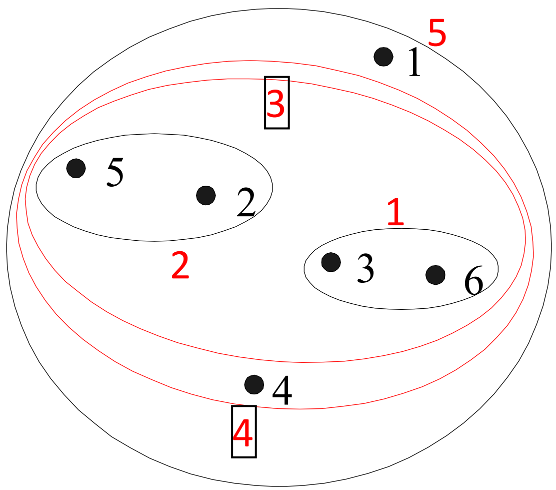
Two Clusters

- Tends to break large clusters
- Biased towards globular clusters

Cluster Similarity: Ward’s Method

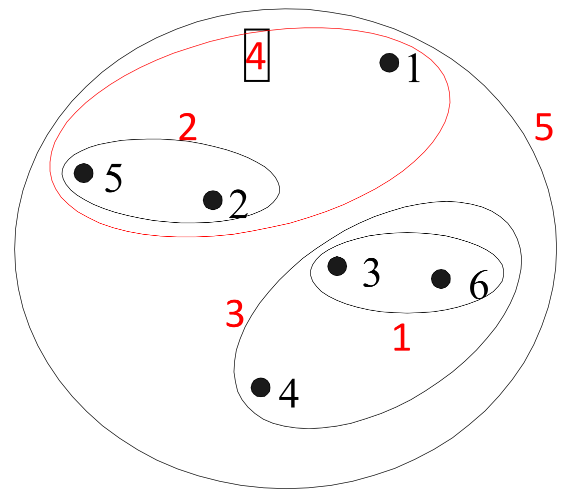
- Similarity of two clusters is based on the increase in squared error when two clusters are merged
  - Similar to group average if distance between points is distance squared
- Less susceptible to noise and outliers
- Biased towards globular clusters
- Hierarchical analogue of K-means
  - Can be used to initialize K-means

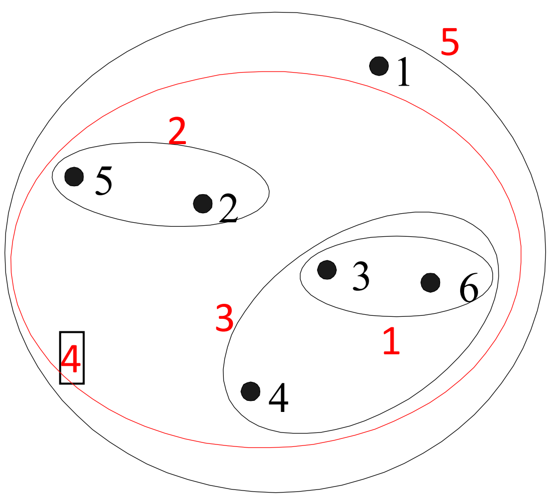
Hierarchical Clustering: Comparison



MIN

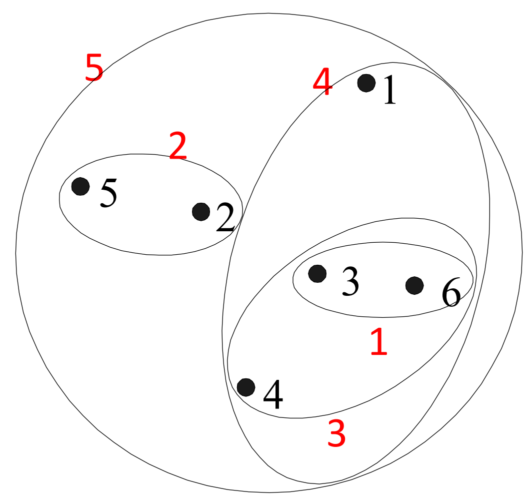
MAX





Group Average

Ward's Method



Hierarchical Clustering: Time and Space requirements

- $O(N^2)$ space since it uses the proximity matrix.
  - N is the number of points.

- $O(N^3)$ time in many cases
  - There are N steps and at each step the size, N^2, proximity matrix must be updated and searched
  - Complexity can be reduced to $O(N^2 \log(N))$ time for some approaches

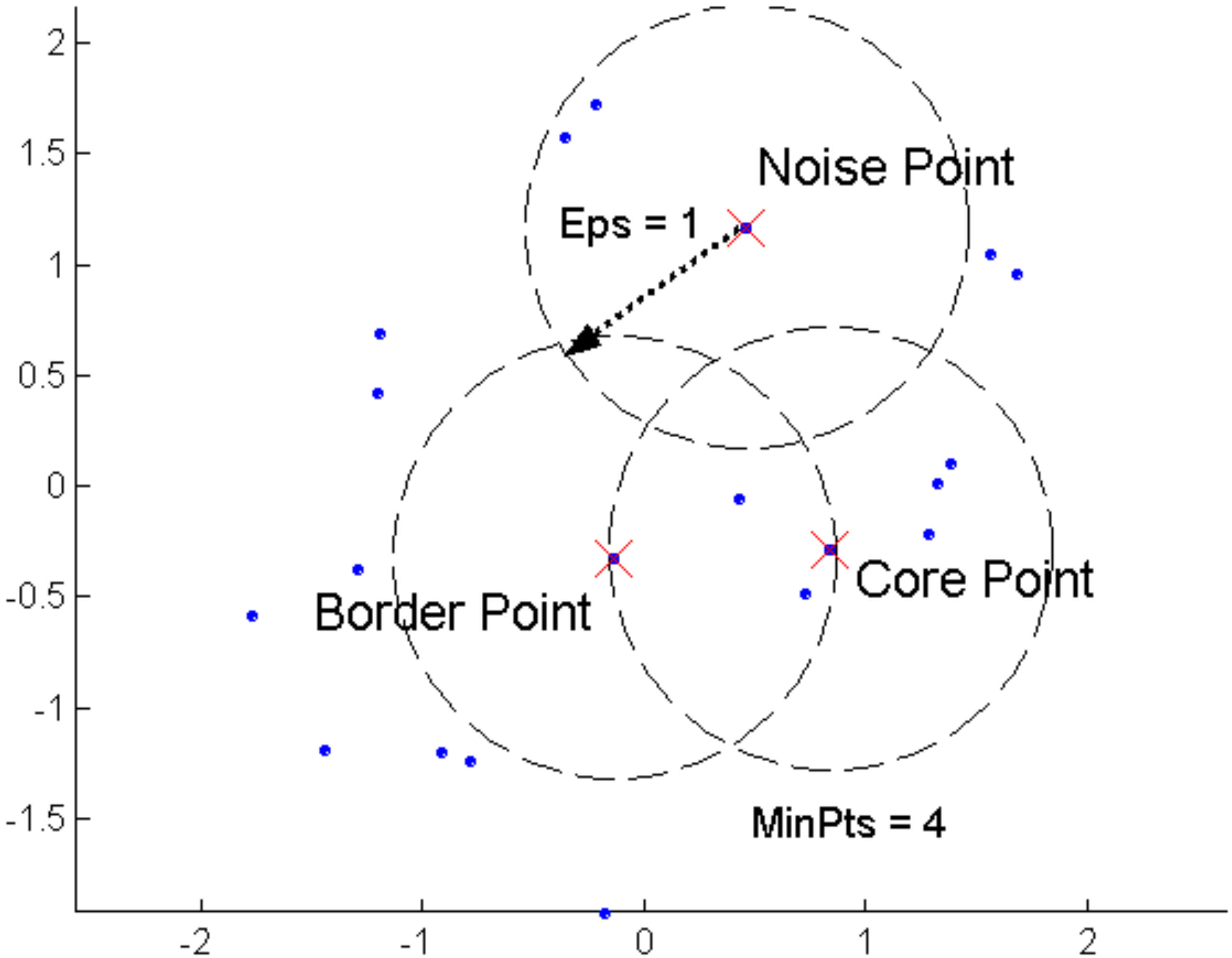
Hierarchical Clustering: Problems and Limitations

- Once a decision is made to combine two clusters, it cannot be undone (one direction)
- No objective function is directly minimized
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and convex shapes
  - Breaking large clusters

DBSCAN

- DBSCAN is a density-based algorithm.
  - Density = number of points within a specified radius (Eps)
  - A point is a core point if it has more than a specified number of points (MinPts) within Eps
    - These are points that are at the interior of a cluster
  - A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point
  - A noise point is any point that is not a core point or a border point.

DBSCAN: Core, Border, and Noise Points



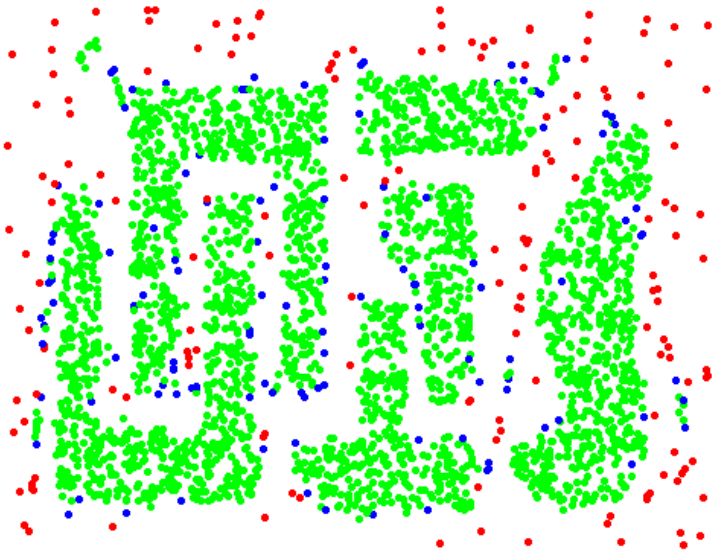
DBSCAN Algorithm

- Eliminate noise points
- F

```
current_cluster_label ← 1
for all core points do
    if the core point has no cluster label then
        current_cluster_label ← current_cluster_label + 1
        Label the current core point with cluster label current_cluster_label
    end if
    for all points in the Eps-neighborhood, except i^th the point itself do
        if the point does not have a cluster label then
            Label the point with cluster label current_cluster_label
        end if
    end for
end for
```

DBSCAN: Core, Border and Noise Points

Original Points

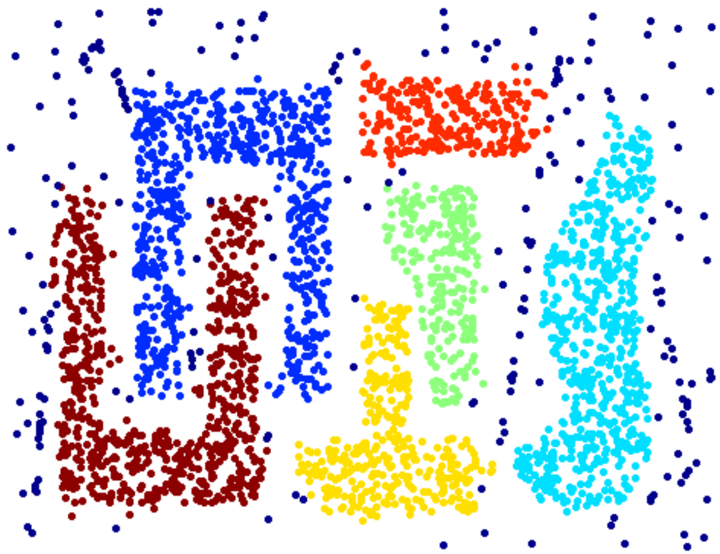


Point types: core, border and noise

Eps = 10, MinPts = 4

When DBSCAN Works Well

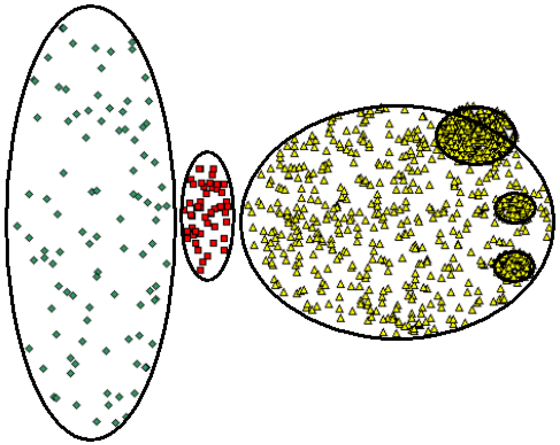
Original Points



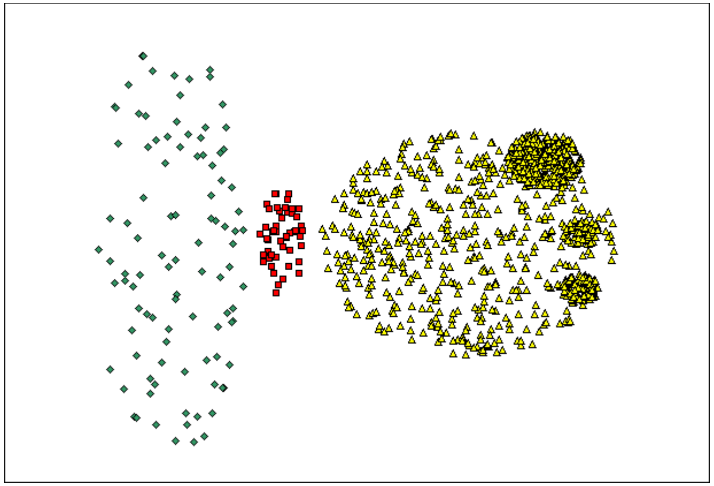
Clusters

- Resistant to Noise
- Can handle clusters of different shapes and sizes

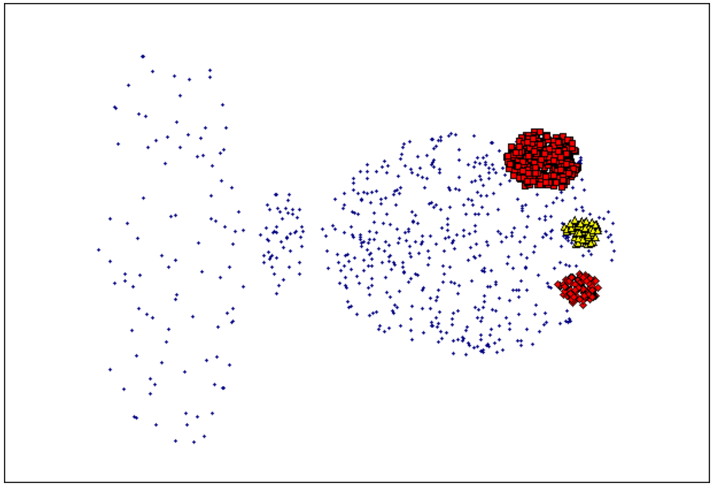
When DBSCAN Does NOT Work Well



Original Points



(MinPts=4, Eps=9.75).

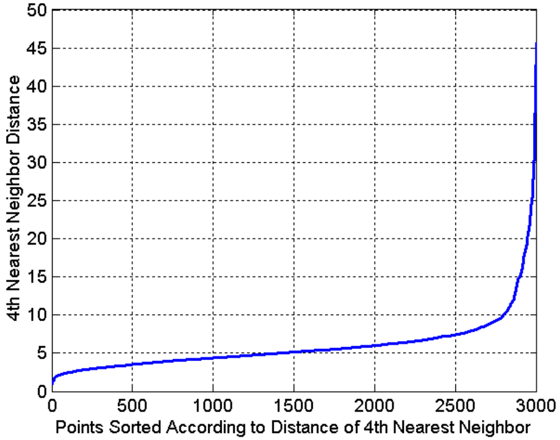


(MinPts=4, Eps=9.92)

- Varying densities
- High-dimensional data

DBSCAN: Determining EPS and MinPts

- Idea is that for points in a cluster, their k^{th} nearest neighbors are at roughly the same distance
- Noise points have the k^{th} nearest neighbor at farther distance
- So, plot sorted distance of every point to its k^{th} nearest neighbor



Clustering is subjective