# *Clustering*

Courtesy: Vipin Kumar and Liaquat Majeed

- Grades
- Active/ Inactive

- Subjective
- Unsupervised

- Distance based
- Center Point (Random, Mean)

- How to handle Outliers/ Abnormal Data?

# Problem Formulation

**Given a set of records,**

**organize the records into clusters (classes)**

***Unsupervised***

- Clustering:

  The process of grouping physical or abstract objects into classes based on their similarity

# What is a cluster?

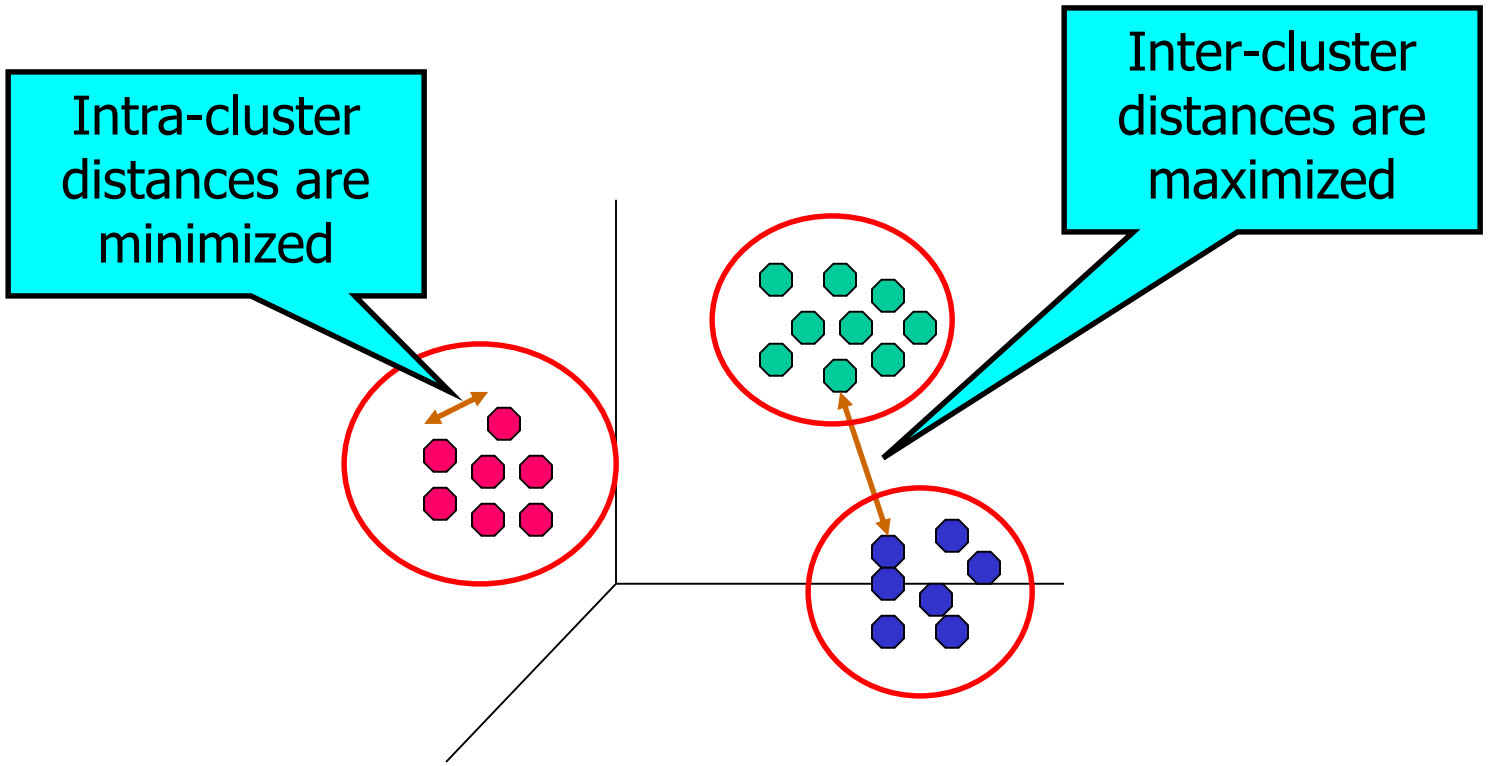
1. **A cluster is a subset of records which are "similar"**

2. **A subset of records such that the distance between any two records in the cluster is less than the distance between any record in other cluster and any record not in it (*Distance Based*).**

3. **A connected region of a multidimensional space containing a relatively high density of records (*Density based*).**

# What is Cluster Analysis?

- ***Fundamental Principal***: Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



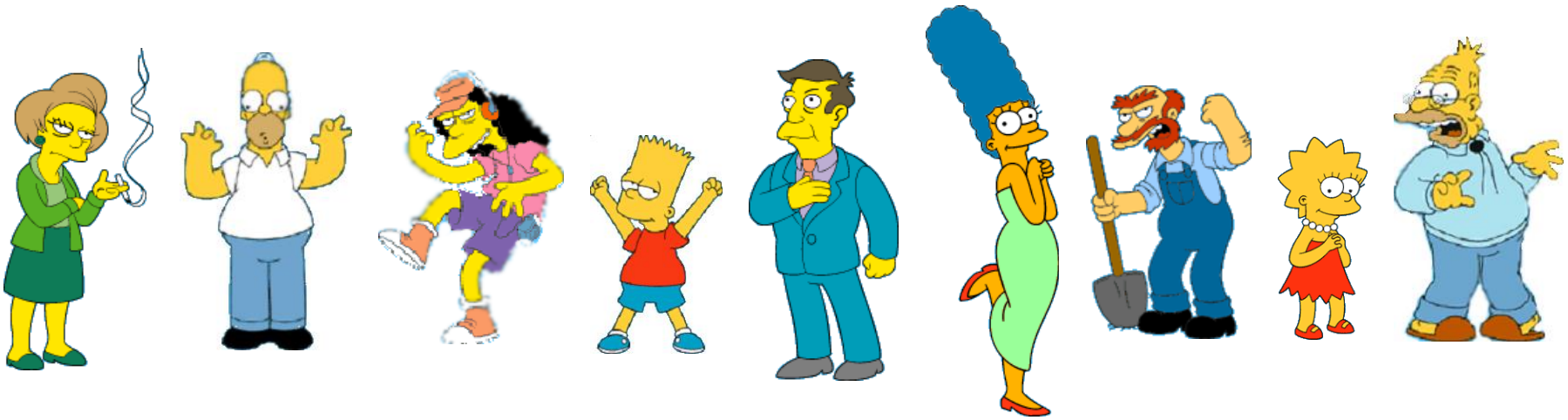
# Density Based

Given points in some multidimensional space – group the points into small number of clusters

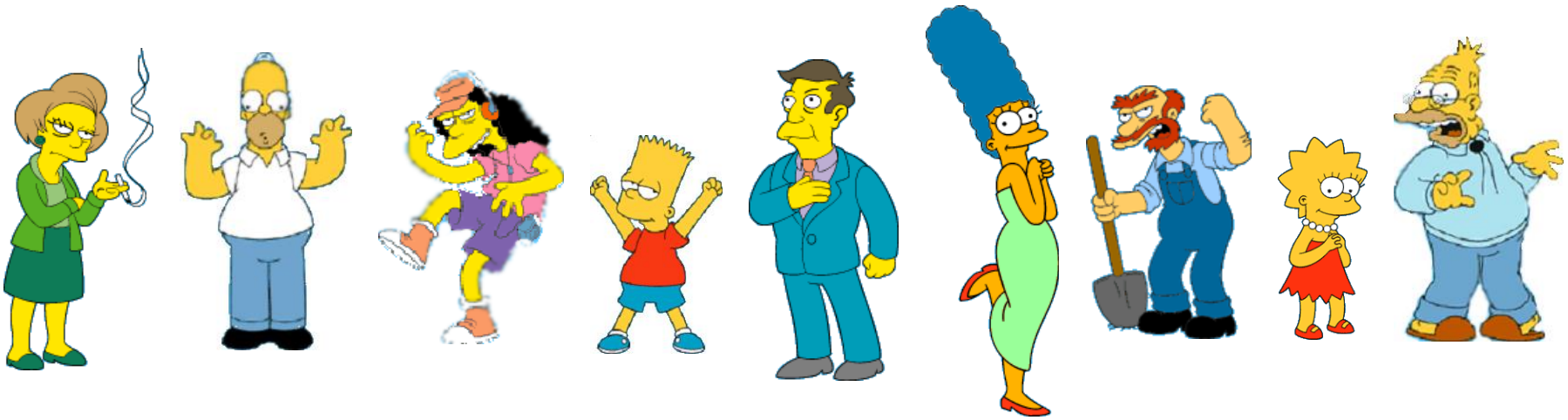
- Skycat project clustered $2 \bullet 10^9$ sky objects into stars, galaxies, quasars, etc. Each object was a point in the 7-dimensions space, with each dimension representing radiation in one band of the spectrum.
- Documents may be thought of as points in a high-dimension space, where each dimension corresponds to one possible word. The position of a document in a dimension is the number of times the word occurs in the document. Clusters of documents correspond to groups of documents on the same topic.

# Clustering is subjective

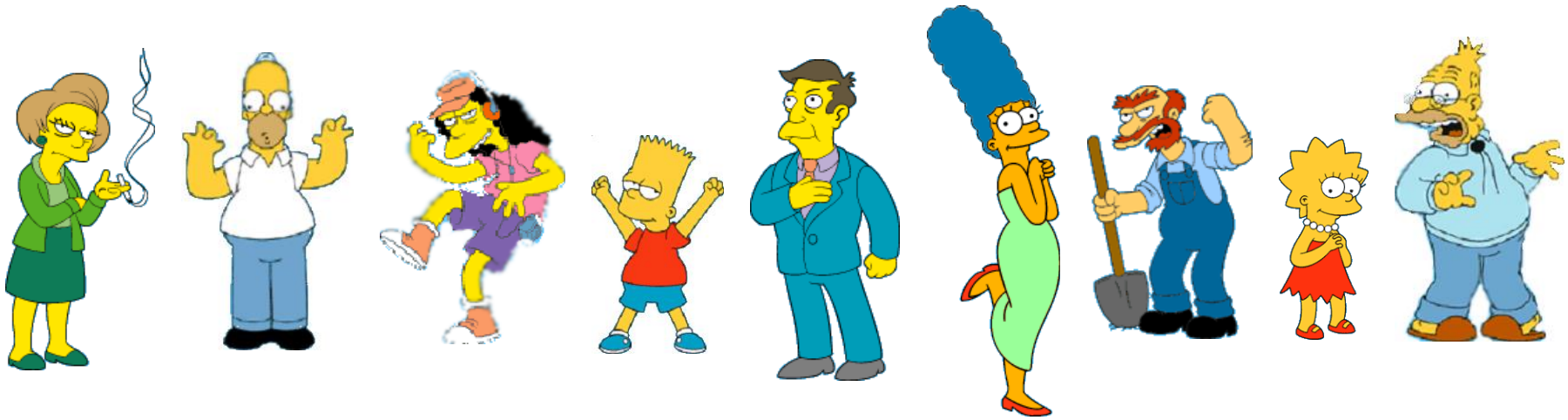
# What is a *natural grouping* among these objects?



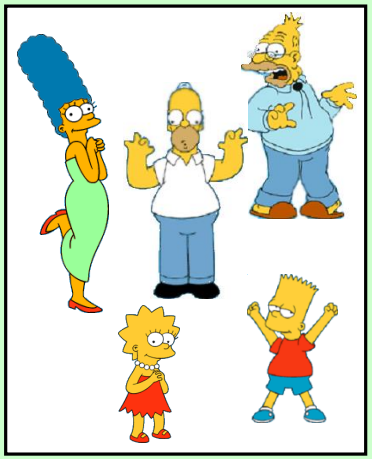
# What is a *natural grouping* among these objects?



# What is a *natural grouping* among these objects?

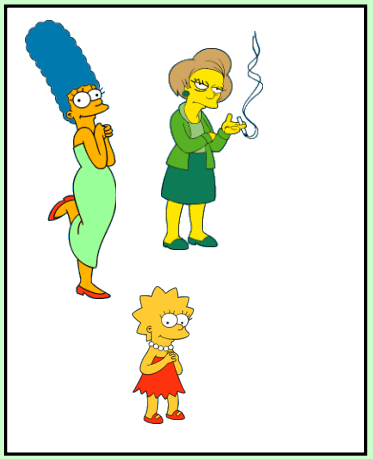


## Clustering is subjective



Simpson's Family

School Employees



Females

Males

# Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

Four Clusters

# What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

**Webster's Dictionary**

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

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

# What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

**Webster's Dictionary**

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

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

# What is Similarity?

The quality or state of being similar; likeness; resemblance; as, a similarity of features.

**Webster's Dictionary**

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

The real meaning of similarity is a philosophical question. We will take a more pragmatic approach.

# Distance Measures

To discuss whether a set of points is close enough to be considered a cluster, we need a distance measure - D(x, y)

# Distance Measures

Assume a k-dimensional Euclidean space, the distance between two points, x=[x1, x2, ..., xk] and y=[y1, y2, ..., yk] may be defined using one of the measures:

- Euclidean distance: ("$L_2$ norm") $\sqrt{\sum_{i=1}^{k}(xi-yi)^2}$
- Manhattan distance: ("$L_1$ norm") $\sum_{i=1}^{k}|xi-yi|$
- Max of dimensions: ("$L_\infty$ norm") $\max_{i=1}^{k}|xi-yi|$

# Distance Measures

- Minkowski distance: $$\left(\sum_{i=1}^{k}\left(\left|x_i - y_i\right|\right)^q\right)^{1/q}$$

When there is no Euclidean space in which to place the points, clustering becomes more difficult:

Examples: ***Web page accesses, DNA sequences, customer sequences, categorical attributes, documents***, etc.

- Mean  (Total Data Set)
- Mean of each group
- Distance of each data point with the calculated mean of the group (mean)
  - Minimum Distance from each mean (we will categorize it in that specific cluster) Same will be followed for all others
- Centroid
- Iterative

# Web Pages: Distance Measures

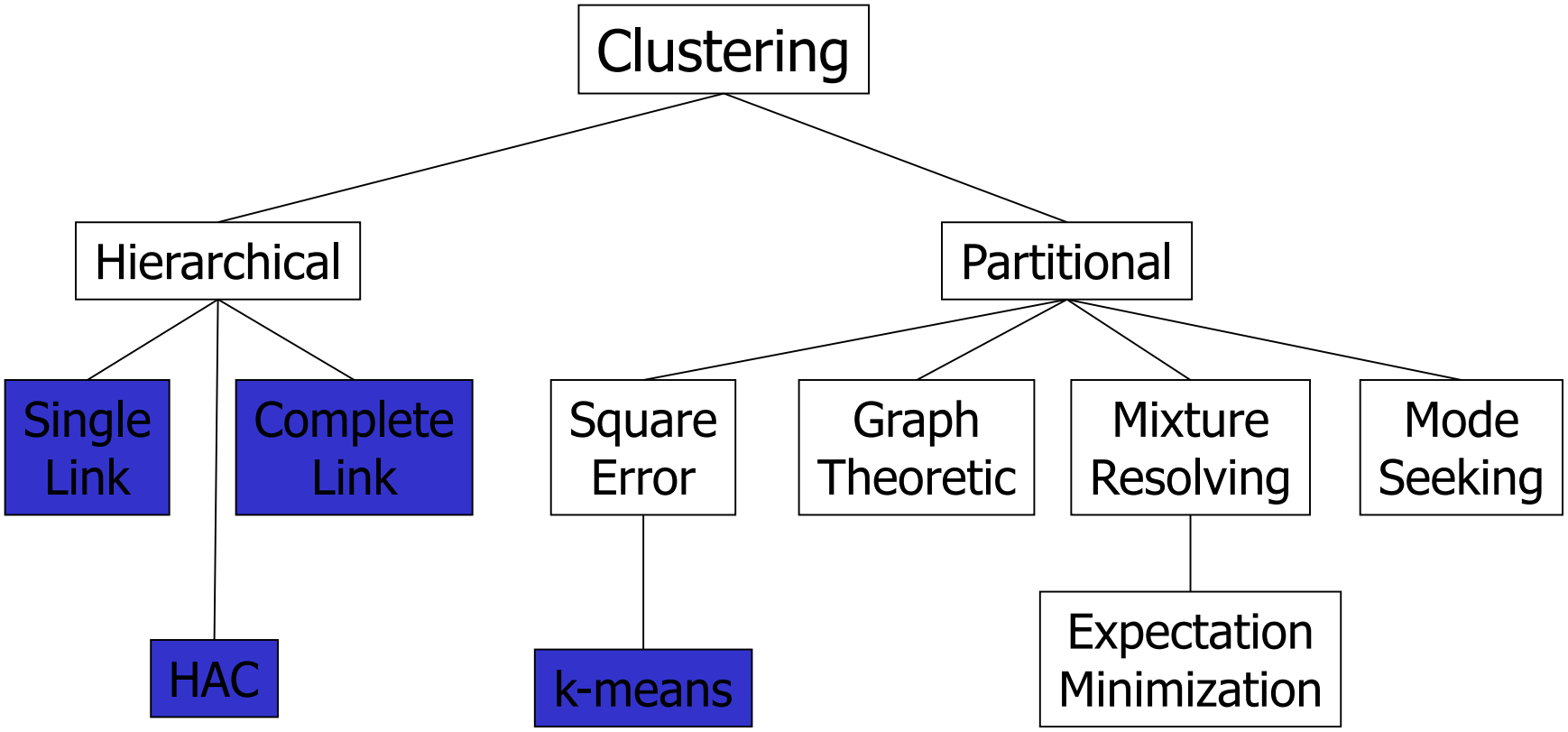
Web pages: points in a multidimensional space where each dimension corresponds to one word.

Idea: Distance D(x, y) is based on the:

- Dot product of vectors corresponding to x and y.

- The length of the vector is equal to the square root of the sum of the squares of the numbers of occurrences of each word.

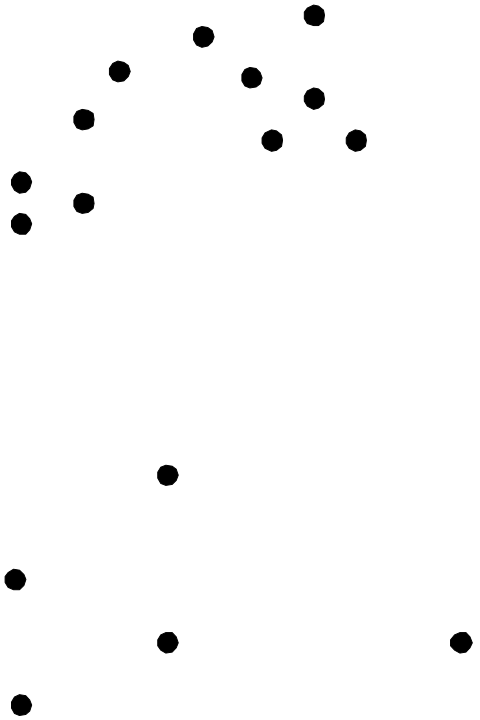
# Taxonomy of Clustering



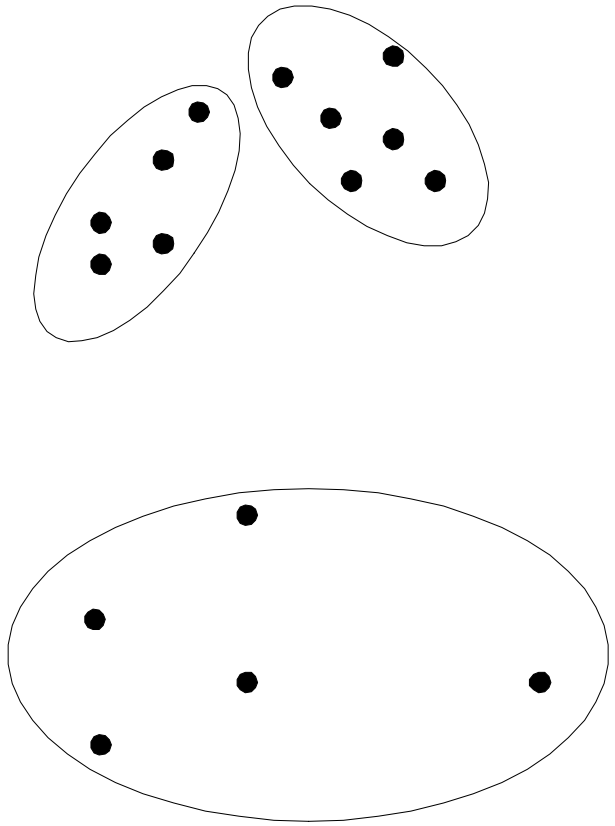
# Types of Clusterings

- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
- Partitional Clustering
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree

# Partitional Clustering

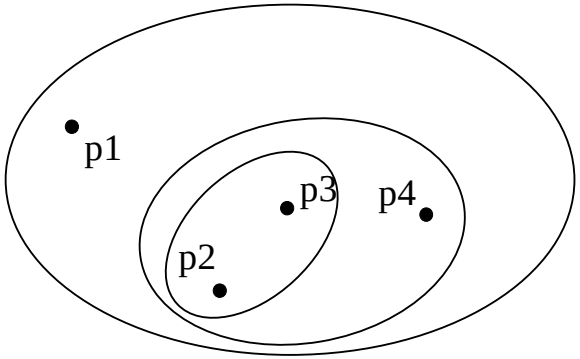


Original Points

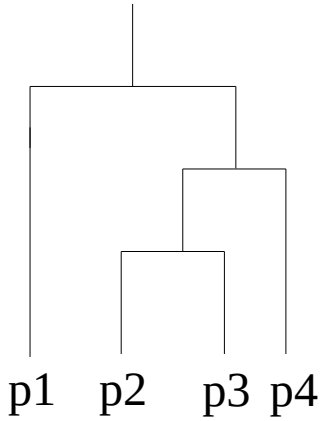


A Partitional Clustering

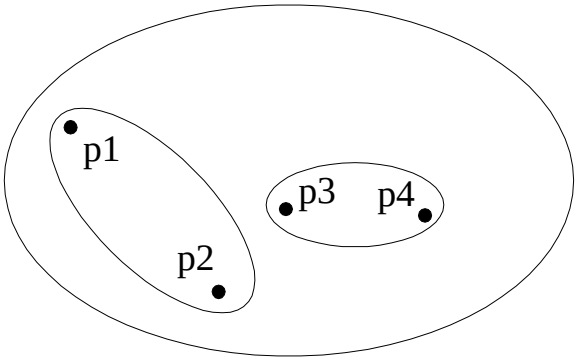
# Hierarchical Clustering



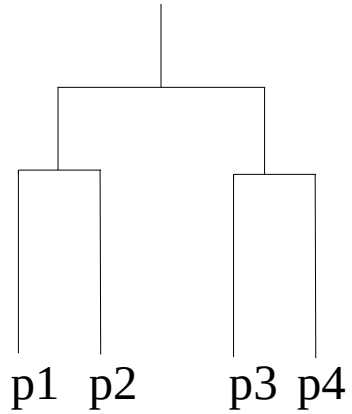

Traditional Hierarchical Clustering




Traditional Dendrogram




Non-traditional Hierarchical Clustering




Non-traditional Dendrogram

# Other Distinctions Between Sets of Clusters

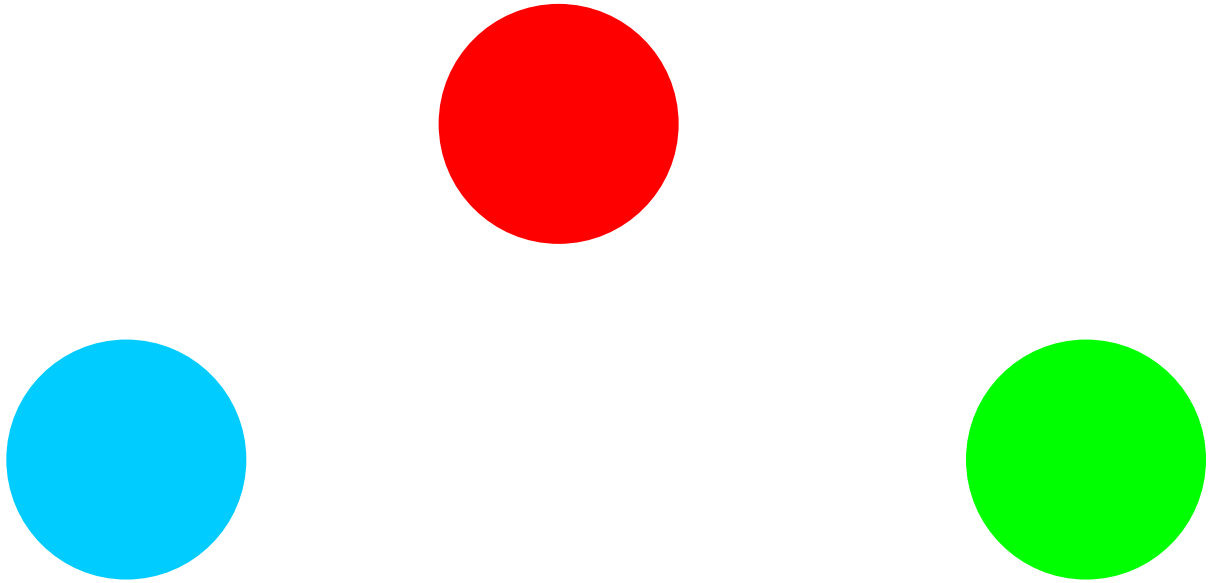
- Exclusive versus non-exclusive
  - In non-exclusive clusterings, points may belong to multiple clusters.
  - Can represent multiple classes or ‘border’ points
- Fuzzy versus non-fuzzy
  - In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
  - Weights must sum to 1
  - Probabilistic clustering has similar characteristics
- Partial versus complete
  - In some cases, we only want to cluster some of the data
- Heterogeneous versus homogeneous
  - Cluster of widely different sizes, shapes, and densities

# Types of Clusters

- Well-separated clusters
- Center-based clusters
- Contiguous clusters
- Density-based clusters
- Property or Conceptual
- Described by an Objective Function

# Types of Clusters: Well-Separated

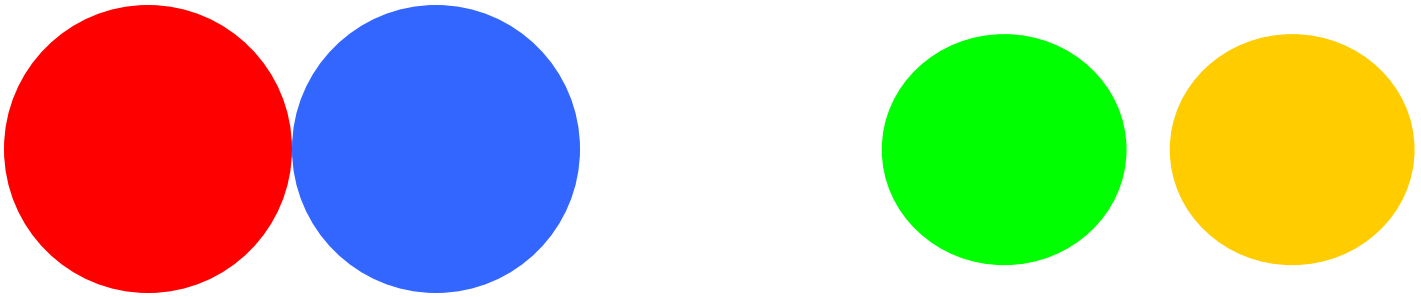
- Well-Separated Clusters:
  - A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



**3 well-separated clusters**

# Types of Clusters: Center-Based

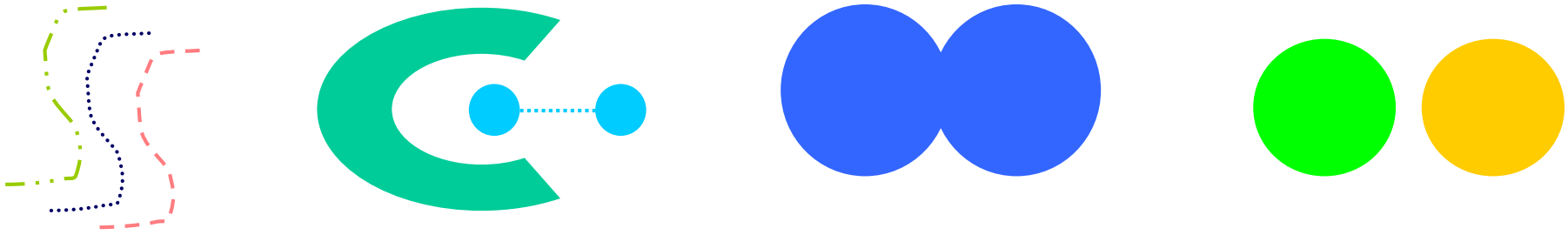
- Center-based
  - A cluster is a set of objects such that an object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster
  - The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most “representative”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



**4 center-based clusters**

# Types of Clusters: Contiguity-Based

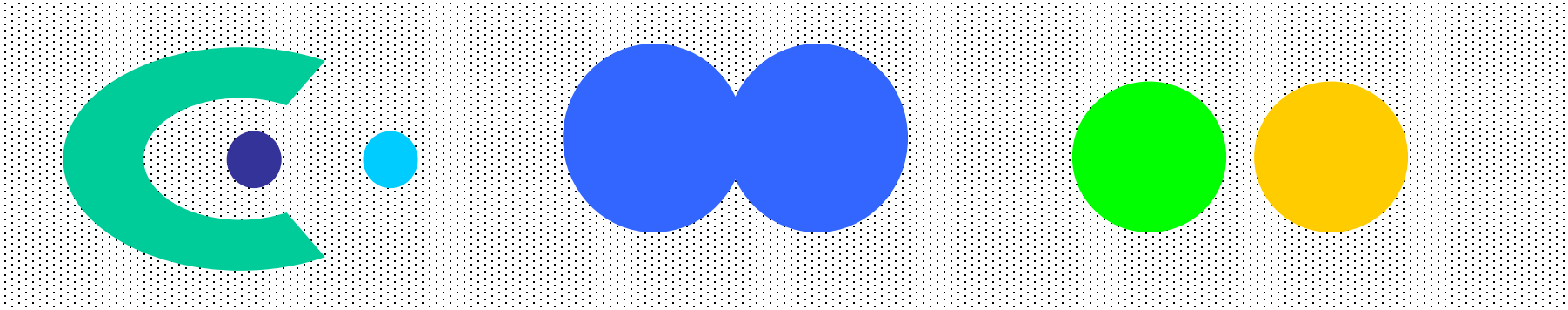
- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.



**8 contiguous clusters**

# Types of Clusters: 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.

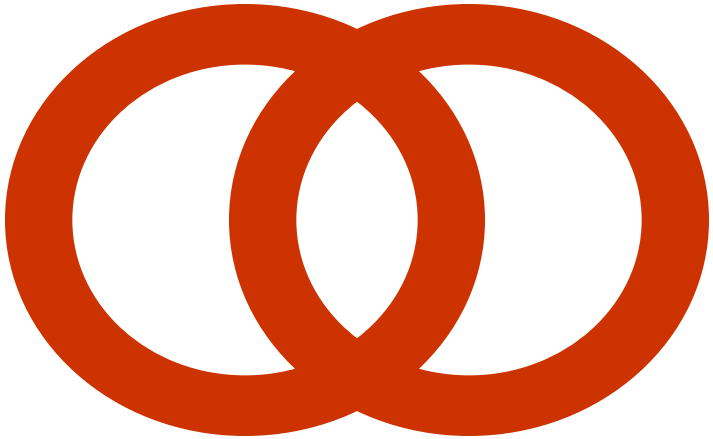


**6 density-based clusters**

# 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**

# Clustering Algorithms

- K-means and its variants
- Hierarchical clustering
- Density-based clustering

# 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

---

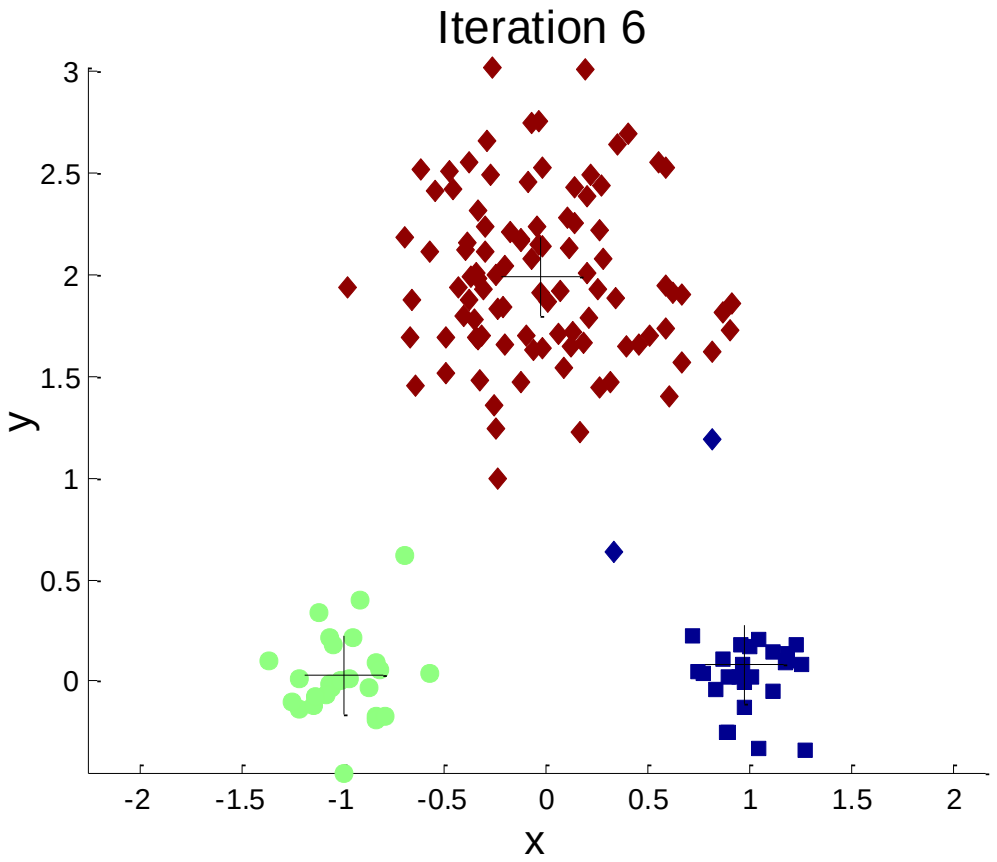
1: Select $K$ points as the initial centroids.
2: **repeat**
3: Form $K$ clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

---

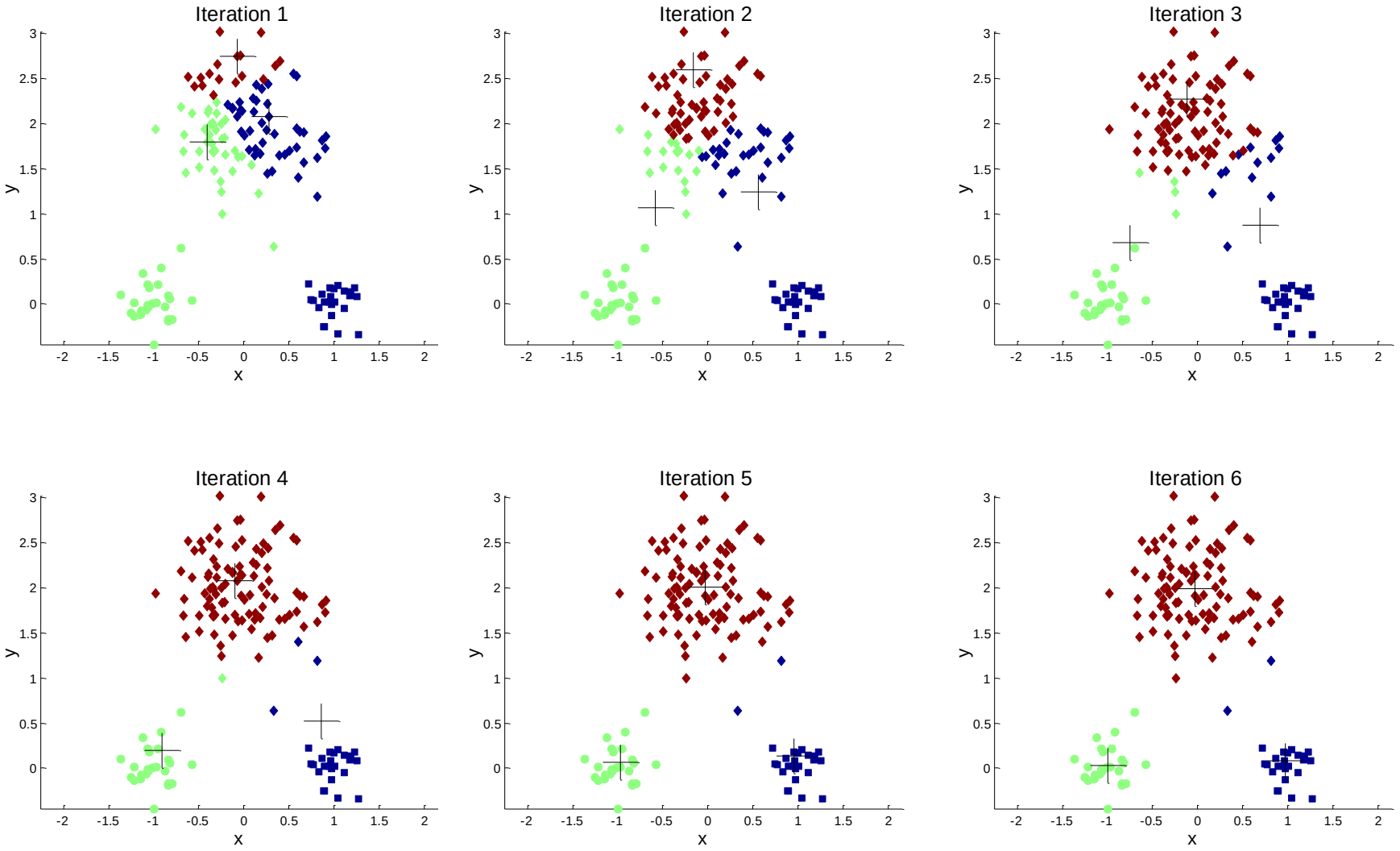
# K-means Clustering – Details

- 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’

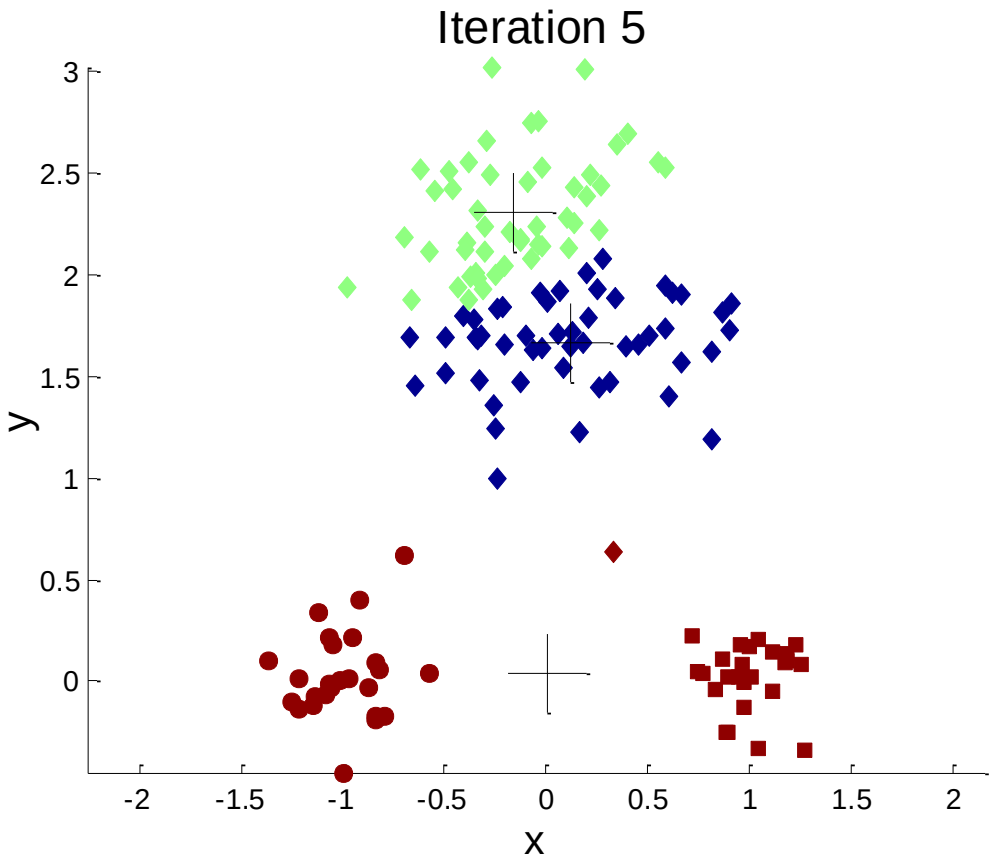
# Importance of Choosing Initial Centroids



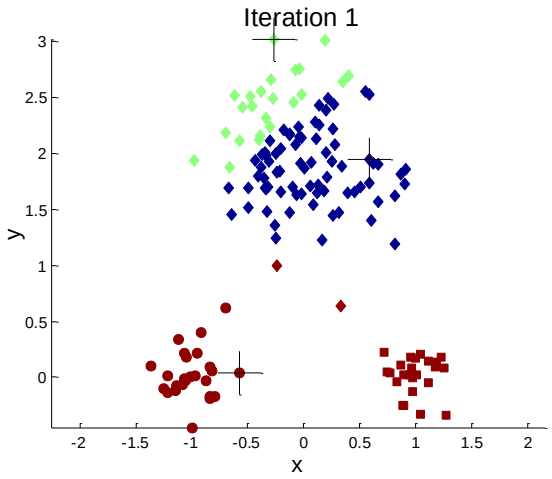
# Importance of Choosing Initial Centroids

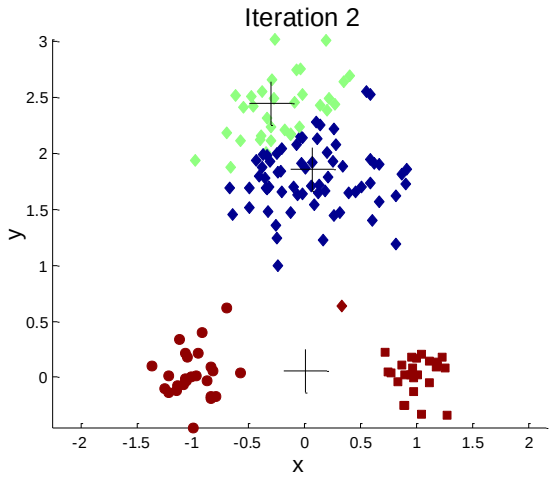


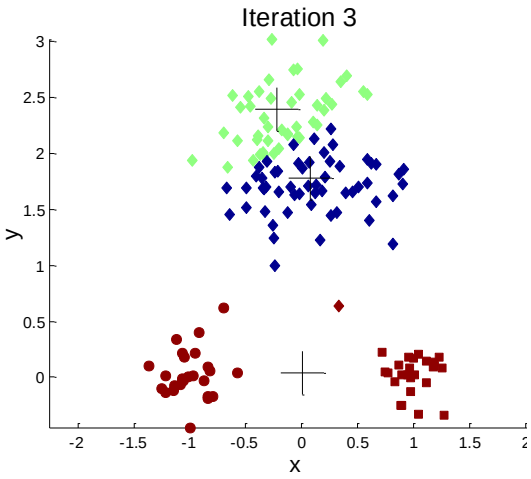
# Importance of Choosing Initial Centroids …

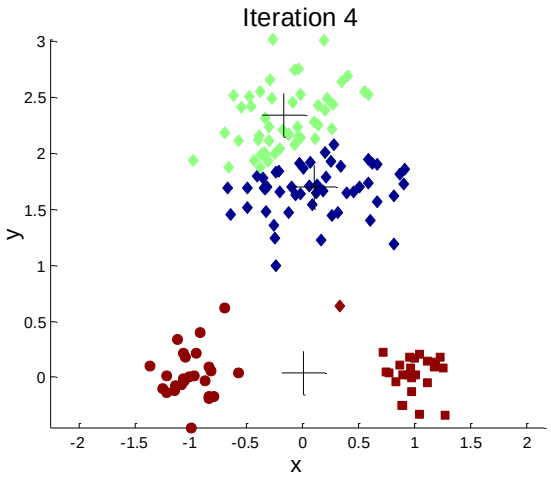


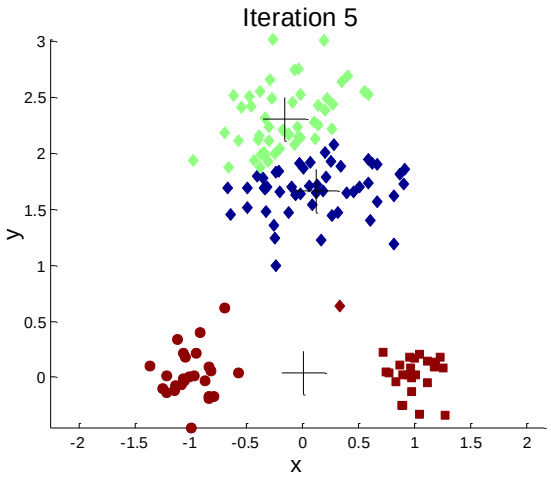
# Importance of Choosing Initial Centroids …



# Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, but probability is not on your side
- Sample and use hierarchical clustering to determine initial centroids
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Postprocessing

# Pre-processing and Post-processing

- Pre-processing
  - Normalize the data
  - Eliminate outliers
- Post-processing
  - Eliminate small clusters that may represent outliers
  - Split ‘loose’ clusters, i.e., clusters with relatively high SSE
  - Merge clusters that are ‘close’ and that have relatively low SSE
  - Can use these steps during the clustering process
    - ISODATA

# 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.

- 99 Total : 450 **- (4.5)**
- 1 255
- 700 – Average : **7**
- **4.5 * 99 = 445.5+(255) =700.5**

- 5 clusters
- 4 clusters (normal behavior)
- 5th cluster (Average) More that 7

# Scaling and Weighting

- For clustering to be effective, all attributes should be converted to a similar scale unless we want to give more weight to some attributes that are relatively large in scale.