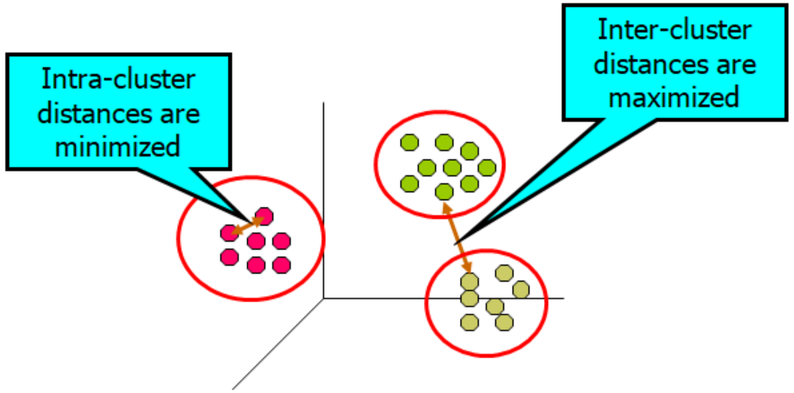
# CLUSTERING


cc-by-sa/2.0 - Clustering round by Alan Murray-Rust - geograph.org.uk/p/1127279

# What is Cluster Analysis?

- Cluster: A collection of data objects
  - similar (or related) to one another within the same group
  - dissimilar (or unrelated) to the objects in other groups
- Cluster analysis (or *clustering*, *data segmentation, …*)
  - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters



# Cluster Analysis

- Unsupervised learning: no predefined classes (i.e., *learning by observations* vs. learning by examples: supervised)
- Clustering is often called an **unsupervised learning** task as no class values denoting an *a priori* grouping of the data instances are given, which is the case in supervised learning.
- Due to historical reasons, clustering is often considered synonymous with unsupervised learning.

# Applications

- Biology: taxonomy of living things: kingdom, phylum, class, order, family, genus and species
- Information retrieval: document clustering, Google search, topic-based news
- Land use: Identification of areas of similar land use in an earth observation database
- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs

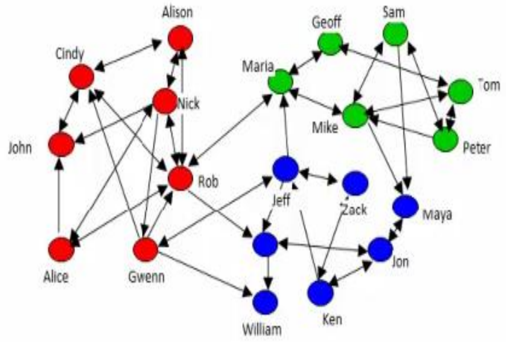
# Applications

- City-planning: Identifying groups of houses according to their house type, value, and geographical location
- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults
- Climate: understanding earth climate, find patterns of atmospheric and ocean

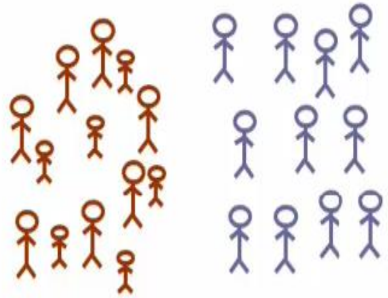
# Applications

- Real Applications: Emerging Applications

Organize computing clusters




Social network analysis



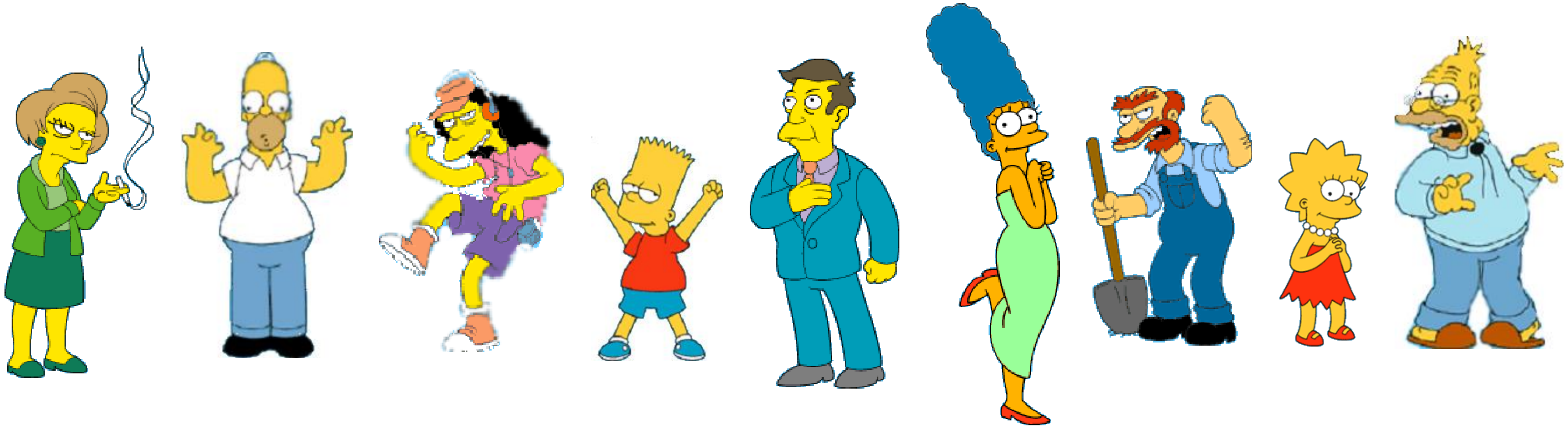
Market segmentation


Astronomical data analysis

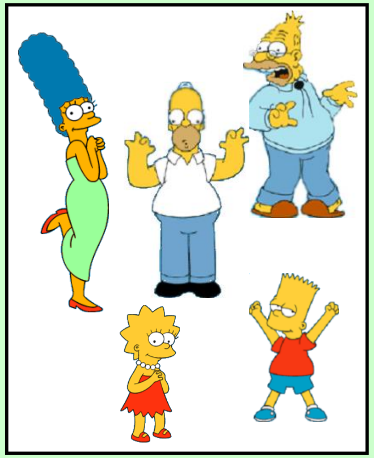
# Considerations for Cluster Analysis

- Partitioning criteria
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)
- Separation of clusters
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- Similarity measure
  - Distance-based (e.g., Euclidian, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- What are the issues?
  - Irrelevant and redundant features
  - Different scales (need normalization)

# What is a natural grouping among these objects?

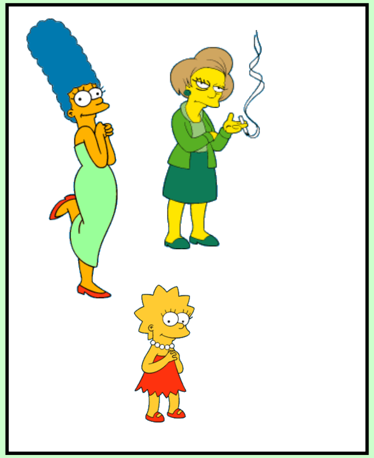


Clustering is subjective



Simpson's Family

School Employees



Females

Males

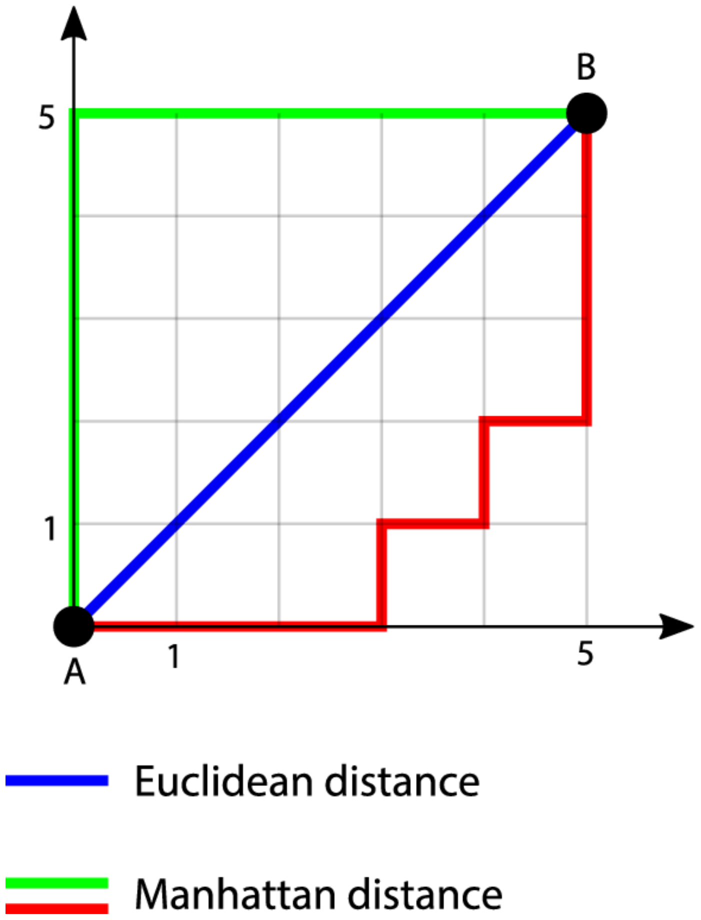
# Requirements and Challenges

- Scalability
  - Clustering all the data instead of only on samples
- Ability to deal with different types of attributes
  - Numerical, binary, categorical, ordinal, linked, and mixture of these
- Constraint-based clustering
  - User may give inputs on constraints
  - Use domain knowledge to determine input parameters
- Others
  - Ability to deal with noisy data
  - Incremental clustering and insensitivity to input order
  - High dimensionality

# Distances

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

$$\mathrm{d}(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2 + (p_2-q_2)^2}.$$




Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

$$|p_1-q_1| + |p_2-q_2|.$$

# Major Clustering Approach

- Partitioning approach
  - Construct various partitions and then evaluate them by some criterion
  - Typical methods:
    - k-means,
    - k-medoids,
    - Squared Error Clustering Algorithm
    - Nearest neighbor algorithm

# Major Clustering Approach(Conti…)

- Hierarchical approach
  - Hierarchical methods obtain a nested partition of the objects resulting in a tree of clusters.
  - Typical methods:
    - BIRCH(Balanced Iterative Reducing and Clustering Using Hierarchies),
    - ROCK(A Hierarchical Clustering Algorithm for Categorical Attributes).
    - Chameleon(A Hierarchical Clustering Algorithm Using Dynamic Modeling).

# Major Clustering Approach(Conti…)

- Density-based approach
  - Based on connectivity and density functions
  - Typical methods:
    - Density based methods include DBSCAN(A Density-Based Clustering Method on Connected Regions with Sufficiently High Density),
    - OPTICS( Ordering Points to Identify the Clustering Structure), DENCLUE(Clustering Based on Density Distribution Functions)

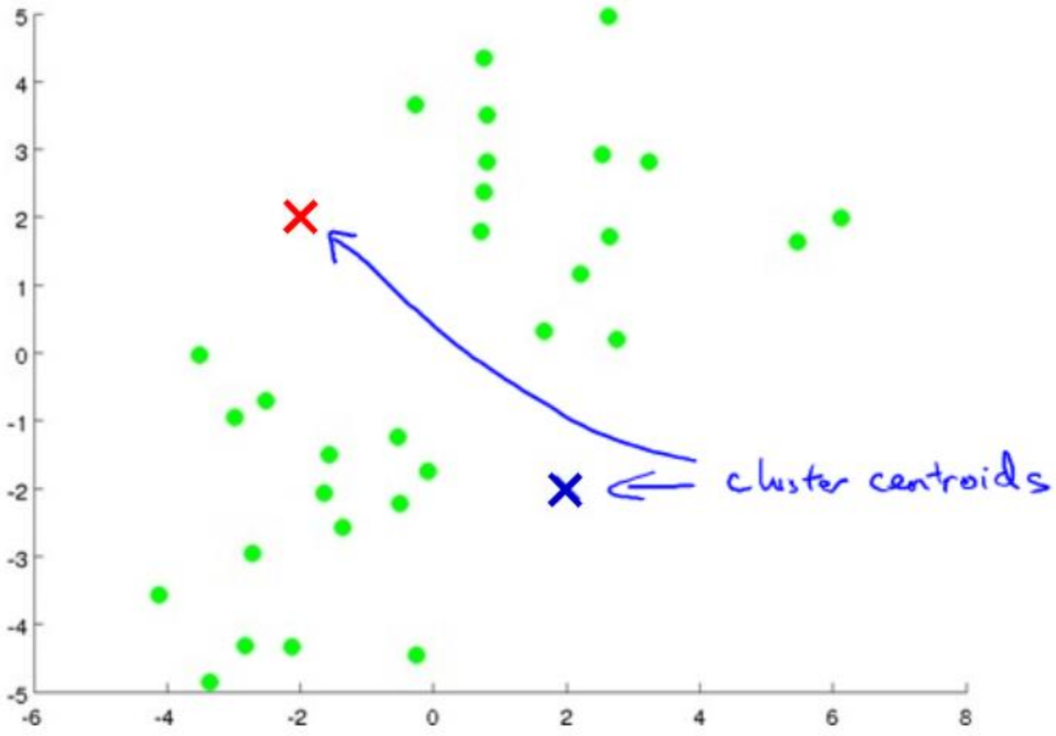
# Other Characteristics

- Exclusive versus non-exclusive
  - In non-exclusive clustering 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)
- Others…

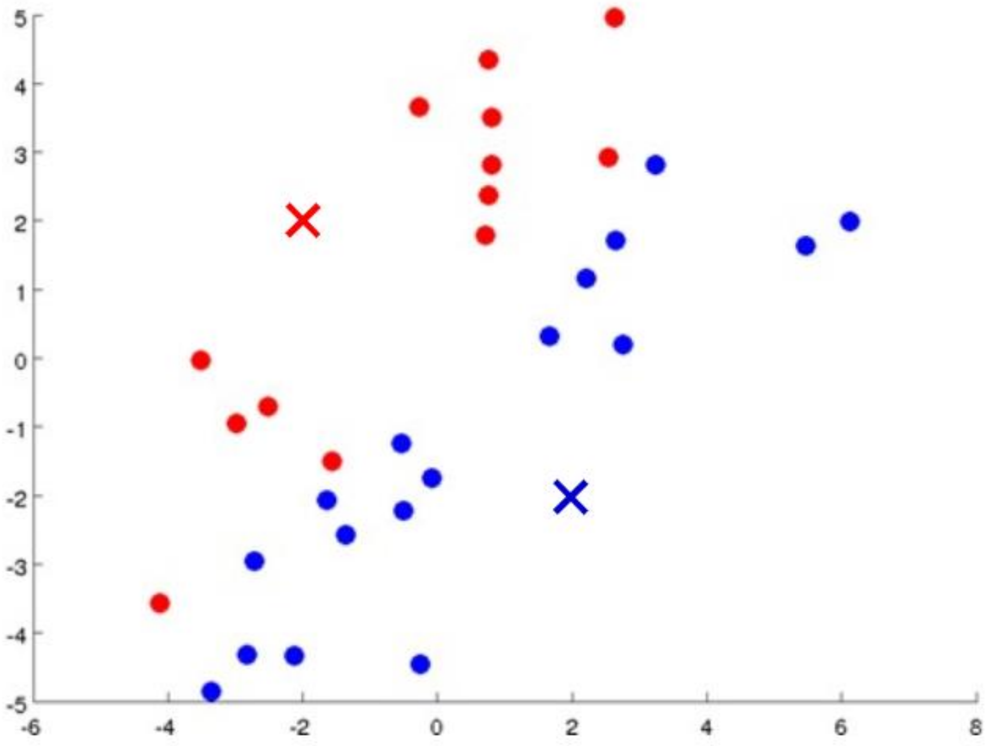
# Clustering Algorithm(K-means)

- K-means Algorithm: The K-means algorithm may be described as follows
  1. Select the number of clusters. Let this number be K.
  2. Pick K seeds as centroids of the k clusters. The seeds may be picked randomly unless the user has some insight into the data.
  3. Compute the Euclidean distance of each object in the dataset from each of the centroids.
  4. Allocate each object to the cluster it is nearest to base on the distances computer in the previous step.
  5. Compute the centroids of the clusters by computing the means of the attribute values of the objects in each cluster.
  6. Cheek if the stopping criterion has been met(e.g. the cluster membership is unchanged) if yes go to step 7. If not, go to step 3.
  7. [optional] One may decide to stop at this stage or to split a cluster or combine two clusters heuristically until a stopping criterion is met.

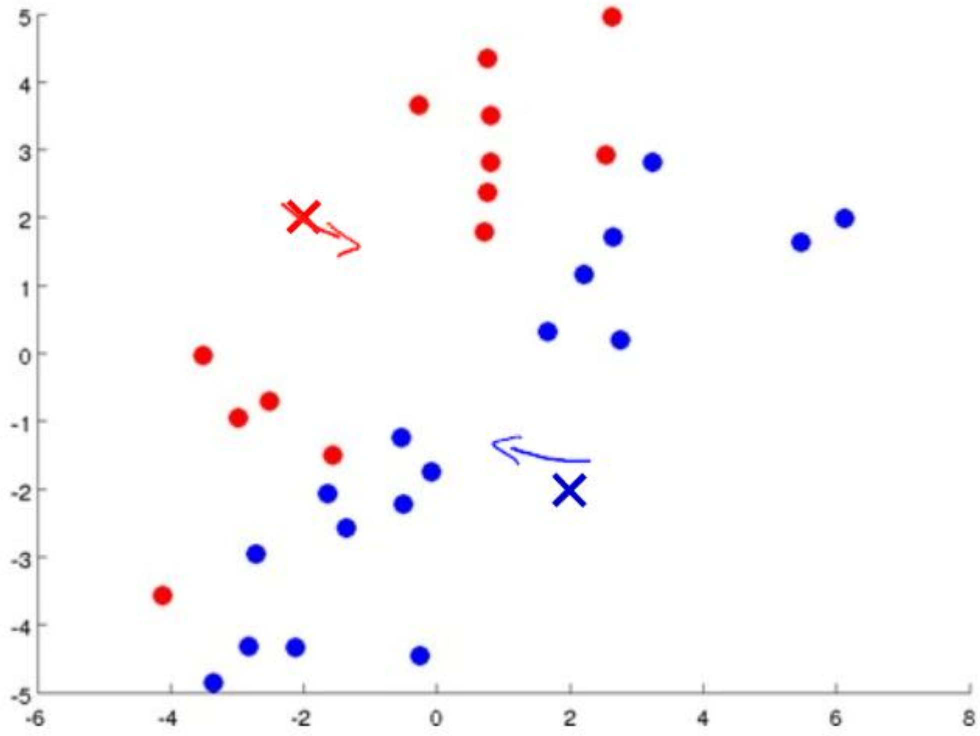
# Clustering Algorithm(K-means)



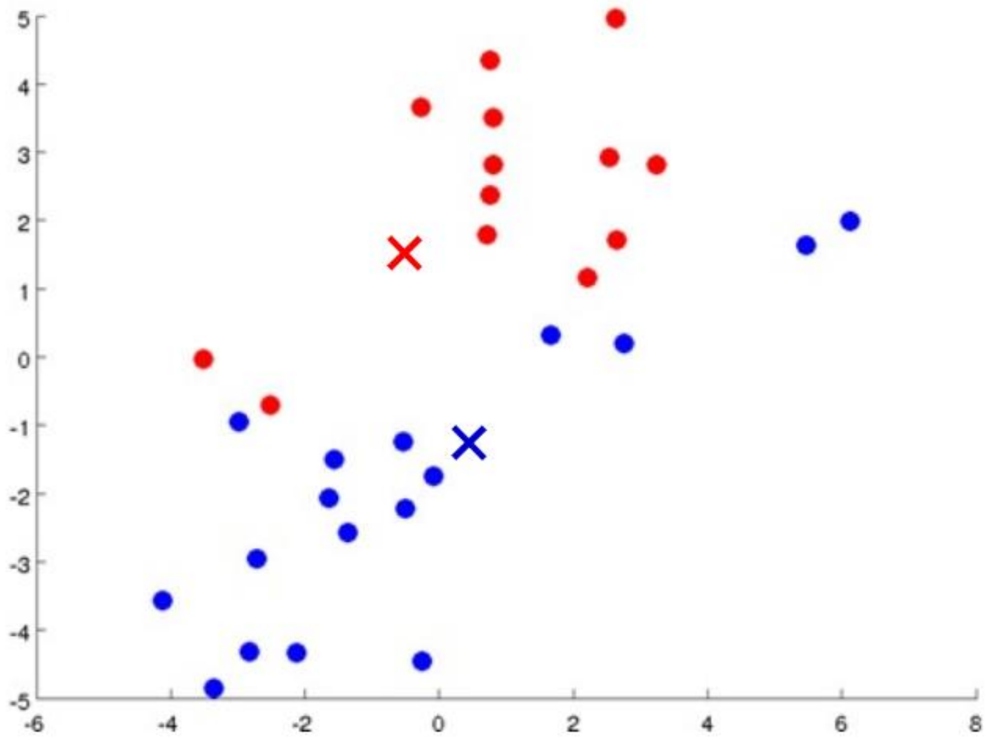
# Clustering Algorithm(K-means)



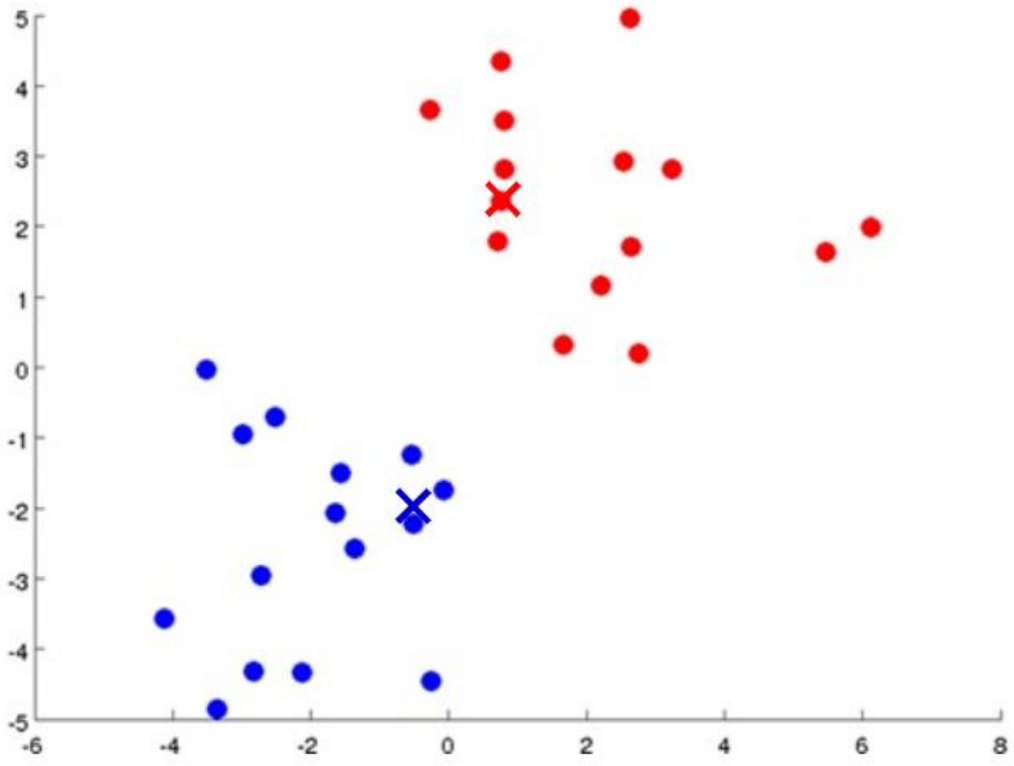
# Clustering Algorithm(K-means)



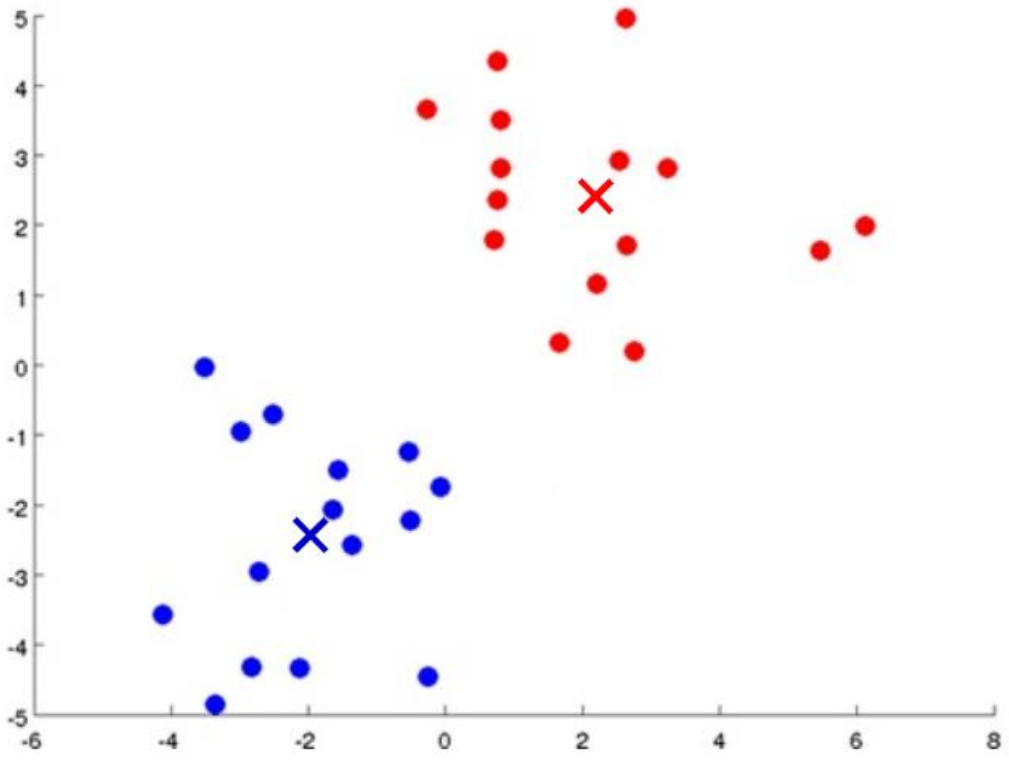
# Clustering Algorithm(K-means)



# Clustering Algorithm(K-means)



# Clustering Algorithm(K-means)



# K-means Clustering – Details

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

# K-means Clustering – Example

Cluster the following eight points (with (x, y) representing locations) into three clusters:
A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)

Notes:

- Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2).
- The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-
  - $P(a, b) = |x2 - x1| + |y2 - y1|$
- Use K-Means Algorithm to find the three cluster centers after the second iteration.

Source: https://www.gatevidyalay.com/k-means-clustering-algorithm-example/

# K-means Clustering – Example

**Calculating Distance Between A1(2, 10) and C1(2, 10)-**

P(A1, C1)
= |x2 – x1| + |y2 – y1|
= |2 – 2| + |10 – 10|
= 0

**Calculating Distance Between A1(2, 10) and C2(5, 8)-**

P(A1, C2)
= |x2 – x1| + |y2 – y1|
= |5 – 2| + |8 – 10|
= 3 + 2
= 5

**Calculating Distance Between A1(2, 10) and C3(1, 2)-**

P(A1, C3)
= |x2 – x1| + |y2 – y1|
= |1 – 2| + |2 – 10|
= 1 + 8
= 9

**In the similar manner, we calculate the distance of other points from each of the center of the three clusters**

Source: https://www.gatevidyalay.com/k-means-clustering-algorithm-example/

# K-means Clustering – Example

| Given Points | Distance from center (2, 10) of Cluster-01 | Distance from center (5, 8) of Cluster-02 | Distance from center (1, 2) of Cluster-03 | Point belongs to Cluster |
|---|---|---|---|---|
| A1(2, 10) | 0 | 5 | 9 | C1 |
| A2(2, 5) | 5 | 6 | 4 | C3 |
| A3(8, 4) | 12 | 7 | 9 | C2 |
| A4(5, 8) | 5 | 0 | 10 | C2 |
| A5(7, 5) | 10 | 5 | 9 | C2 |
| A6(6, 4) | 10 | 5 | 7 | C2 |
| A7(1, 2) | 9 | 10 | 0 | C3 |
| A8(4, 9) | 3 | 2 | 10 | C2 |

- We draw a table showing all the results.
- Using the table, we decide which point belongs to which cluster.
- The given point belongs to that cluster whose center is nearest to it.

Source: https://www.gatevidyalay.com/k-means-clustering-algorithm-example/

# K-means Clustering – Example

| Given Points | Distance from center (2, 10) of Cluster-01 | Distance from center (5, 8) of Cluster-02 | Distance from center (1, 2) of Cluster-03 | Point belongs to Cluster |
|---|---|---|---|---|
| A1(2, 10) | 0 | 5 | 9 | C1 |
| A2(2, 5) | 5 | 6 | 4 | C3 |
| A3(8, 4) | 12 | 7 | 9 | C2 |
| A4(5, 8) | 5 | 0 | 10 | C2 |
| A5(7, 5) | 10 | 5 | 9 | C2 |
| A6(6, 4) | 10 | 5 | 7 | C2 |
| A7(1, 2) | 9 | 10 | 0 | C3 |
| A8(4, 9) | 3 | 2 | 10 | C2 |

From here, New clusters are-

**Cluster-01:**

First cluster contains points-
- A1(2, 10)

**Cluster-02:**

Second cluster contains points-
- A3(8, 4)
- A4(5, 8)
- A5(7, 5)
- A6(6, 4)
- A8(4, 9)

**Cluster-03:**

Third cluster contains points-
- A2(2, 5)
- A7(1, 2)

Now, we recompute the new cluster by taking mean of all the points contained in that cluster.

**For Cluster-01:**

We have only one point So, cluster center remains the same.

**For Cluster-02:**

Center of Cluster-02
= ((8 + 5 + 7 + 6 + 4)/5, (4 + 8 + 5 + 4 + 9)/5) = (6, 6)

**For Cluster-03:**

Center of Cluster-03
= ((2 + 1)/2, (5 + 2)/2)
= (1.5, 3.5)

This is completion of Iteration-01.

Source: https://www.gatevidyalay.com/k-means-clustering-algorithm-example/

# K-means Clustering – Example

| Given Points | Distance from center (2, 10) of Cluster-01 | Distance from center (6, 6) of Cluster-02 | Distance from center (1.5, 3.5) of Cluster-03 | Point belongs to Cluster |
|---|---|---|---|---|
| A1(2, 10) | 0 | 8 | 7 | C1 |
| A2(2, 5) | 5 | 5 | 2 | C3 |
| A3(8, 4) | 12 | 4 | 7 | C2 |
| A4(5, 8) | 5 | 3 | 8 | C2 |
| A5(7, 5) | 10 | 2 | 7 | C2 |
| A6(6, 4) | 10 | 2 | 5 | C2 |
| A7(1, 2) | 9 | 9 | 2 | C3 |
| A8(4, 9) | 3 | 5 | 8 | C1 |

From here, New clusters are-

**Cluster-01:**

First cluster contains points-
- A1(2, 10)
- A8(4, 9)

**Cluster-02:**

Second cluster contains points-
- A3(8, 4)
- A4(5, 8)
- A5(7, 5)
- A6(6, 4)

**Cluster-03:**

Third cluster contains points-
- A2(2, 5)
- A7(1, 2)

**Re Computation:**

**For Cluster-01:**

Center of Cluster-01
= ((2 + 4)/2, (10 + 9)/2) = (3, 9.5)

**For Cluster-02:**

Center of Cluster-02
= ((8 + 5 + 7 + 6)/4, (4 + 8 + 5 + 4)/4) = (6.5, 5.25)

**For Cluster-03:**

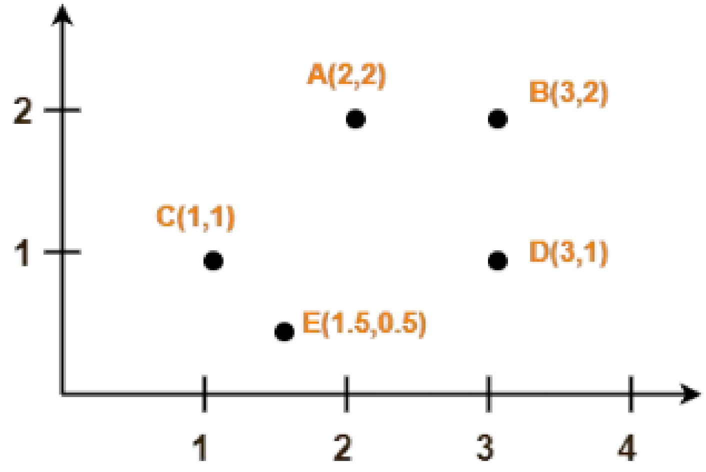
Center of Cluster-03
= ((2 + 1)/2, (5 + 2)/2) = (1.5, 3.5)

**After second iteration, the center of the three clusters are-**
- **C1(3, 9.5)**
- **C2(6.5, 5.25)**
- **C3(1.5, 3.5)**

Source: https://www.gatevidyalay.com/k-means-clustering-algorithm-example/

# K-means Clustering – Exercise

Use K-Means Algorithm to create two clusters for the points given in the figure.




Notes:

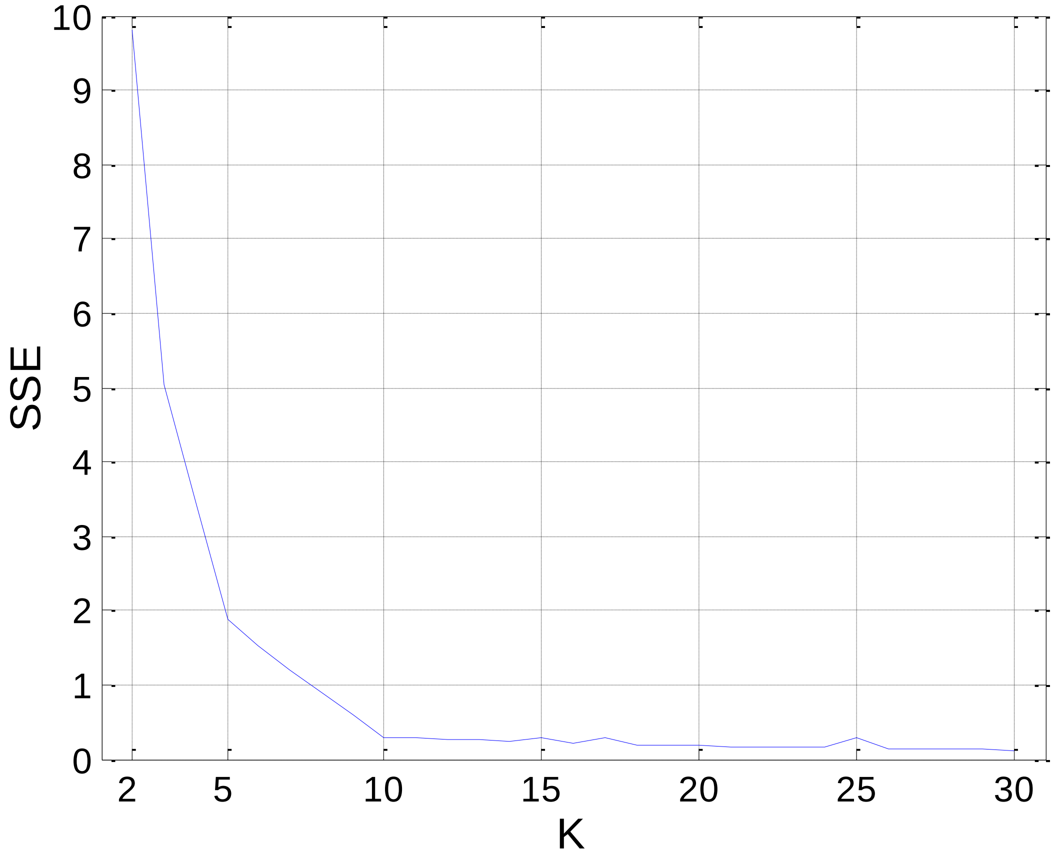
- Initial cluster centers are: A(2, 2) and C(1, 1)
- The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as-
    - P(a, b) = SQRT[$(x2 – x1)^2 + (y2 – y1)^2$]
- Use K-Means Algorithm to find the two cluster centers after the second iteration.

Source: https://www.gatevidyalay.com/k-means-clustering-algorithm-example/

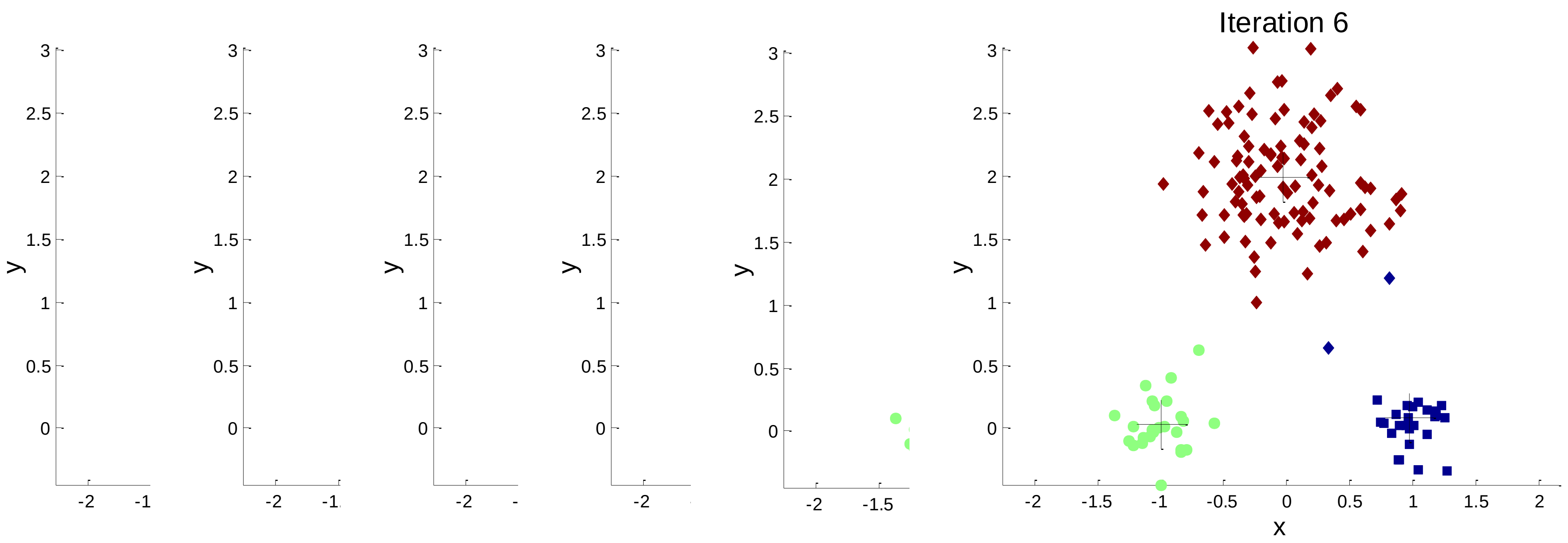
# Evaluating K-means Clusters

- How do you objectively evaluate the quality of a clustering (so can choose best)?
  - Most common measure: Sum of Squared Error (SSE)
  - Error for each point is distance to nearest cluster center
- What happens to SSE if you increase K, the number of clusters?
  - SSE would go down (Can use relationship between K and SSE to find proper K)
- A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

# Relationship between K and SSE

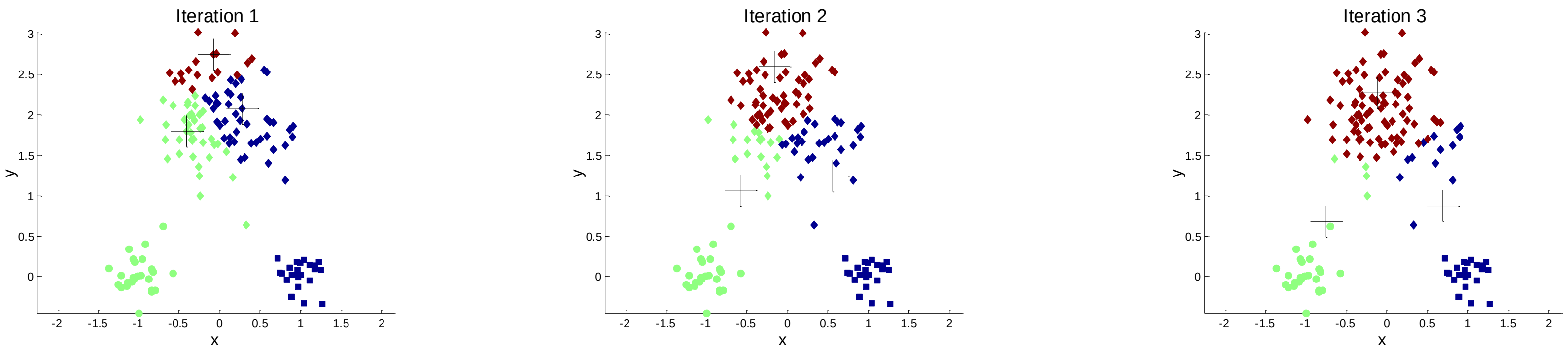


# Importance of Choosing Initial Centroids

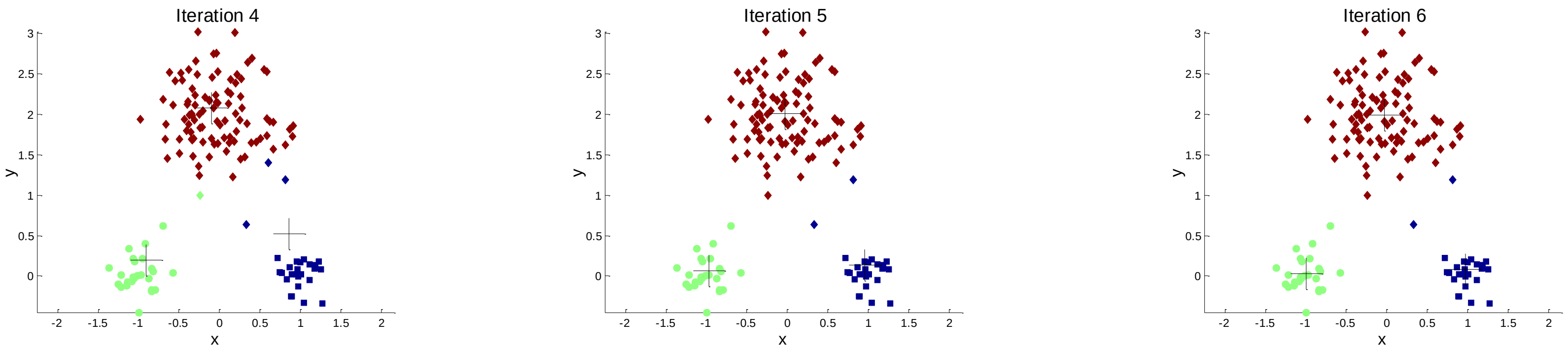



**If you happen to choose good initial centroids, then you will get this after 6 iterations**

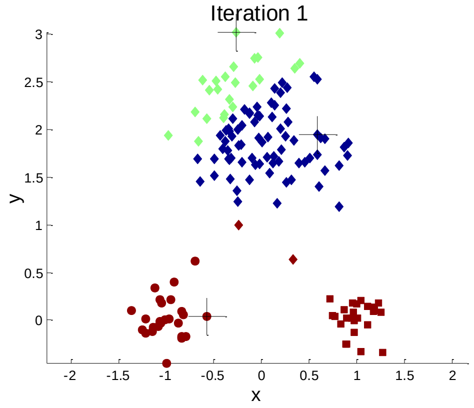
# Importance of Choosing Initial Centroids

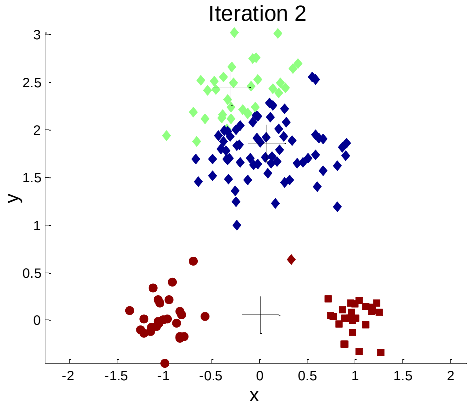



**Good clustering**

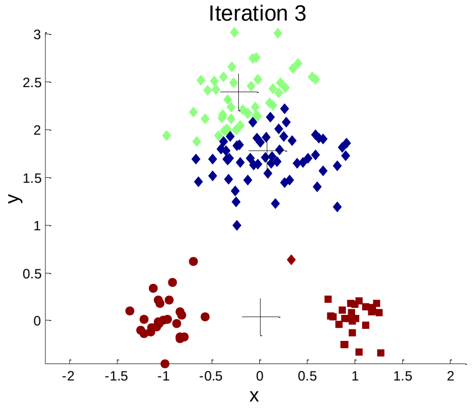


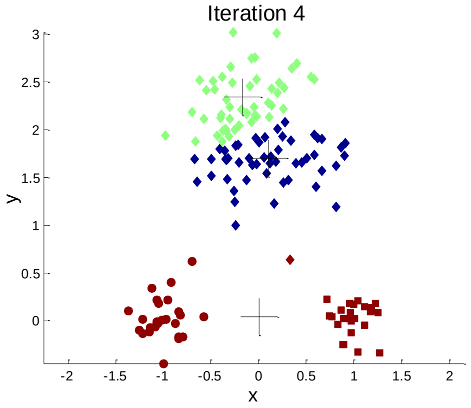
# Importance of Choosing Initial Centroids …

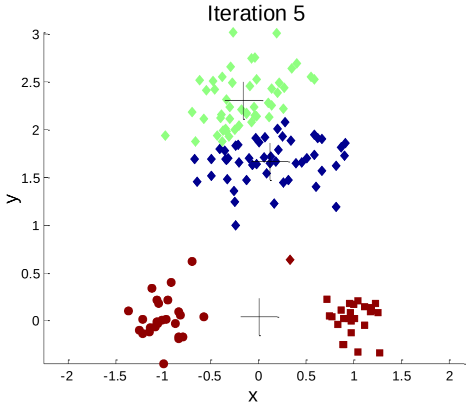



**Bad Clustering**



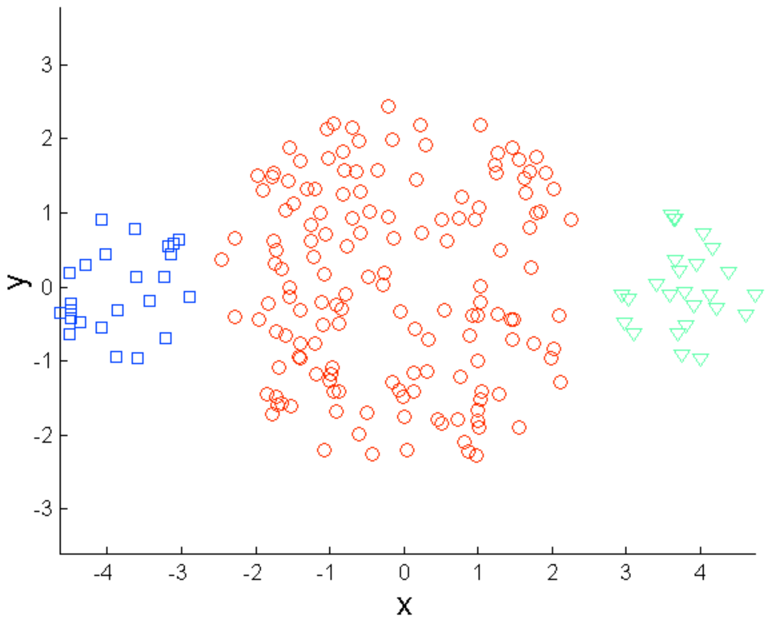
# Pre-processing and Post-processing

- Pre-processing
  - Normalize the data
  - Eliminate outliers
- Post-processing
  - Eliminate small clusters that may represent outliers
  - Split 'loose' clusters (with relatively high SSE)
  - Merge clusters that are 'close' and that have relatively low SSE

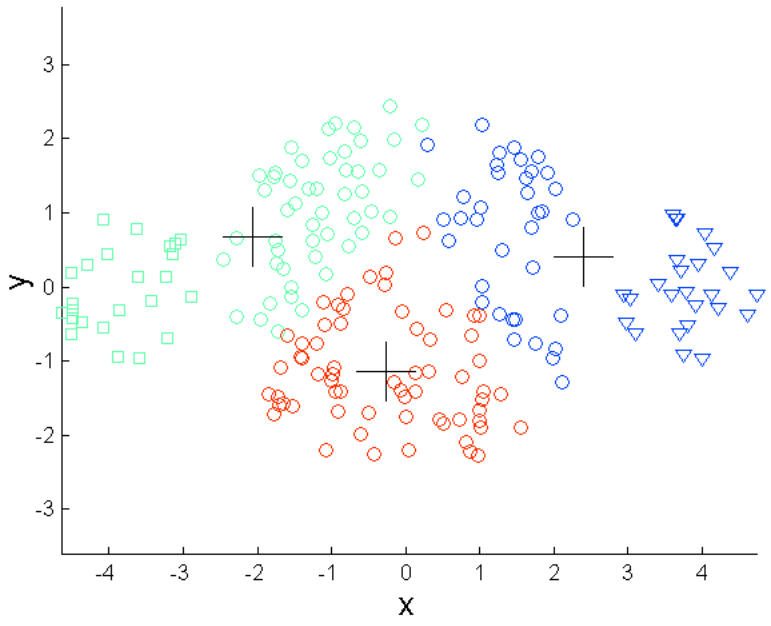
## Limitations of K-means

- K-means has problems when clusters are of differing:
  - Sizes (biased toward the larger clusters)
  - Densities
  - Non-spherical shapes
- K-means has problems with outliers

# Limitations of K-means: Differing Sizes

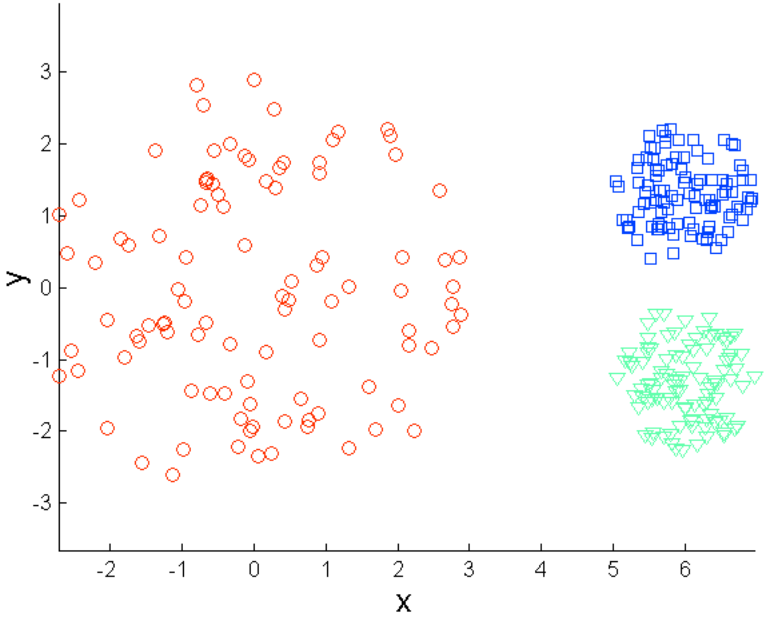



Original Points

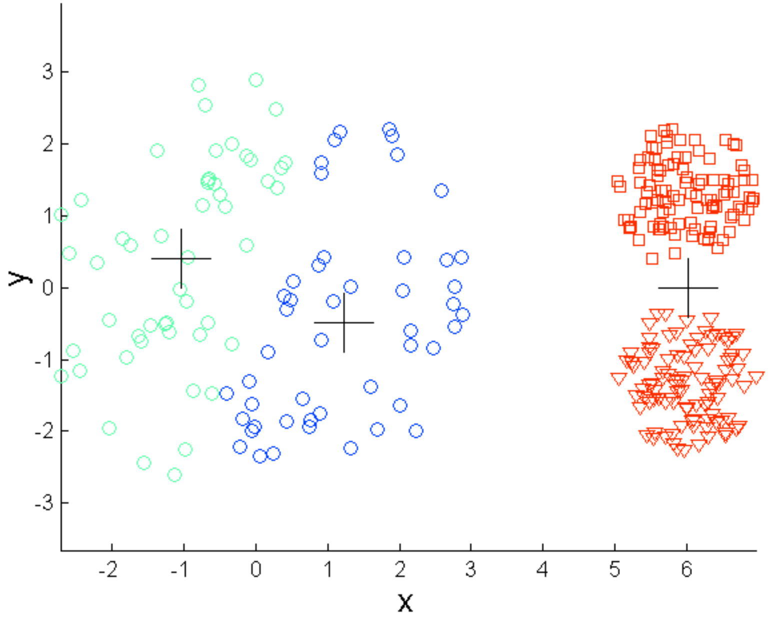



K-means (3 Clusters)

# Limitations of K-means: Differing Density

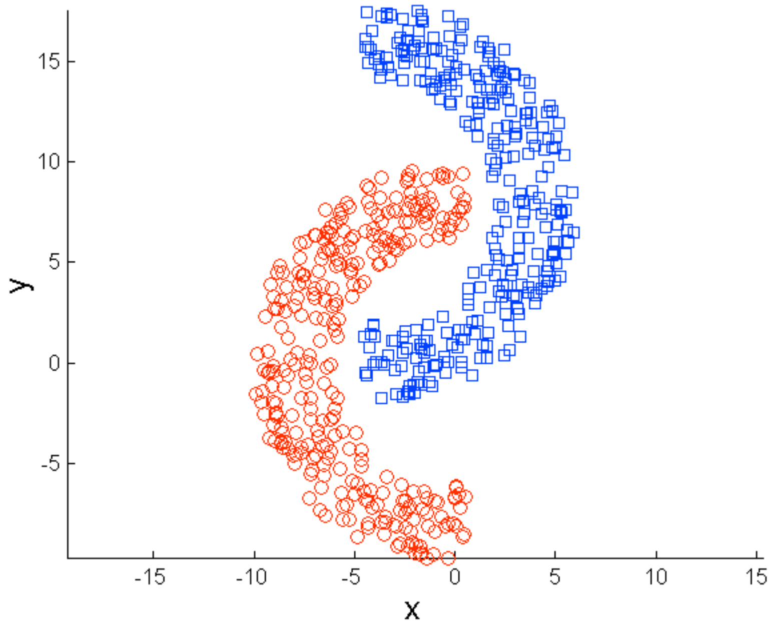



**Original Points**

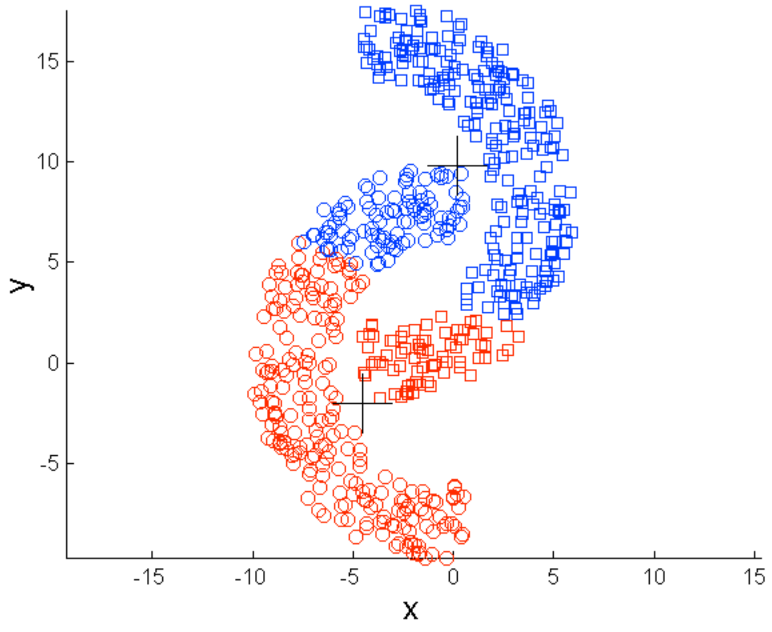



**K-means (3 Clusters)**

# Limitations of K-means: Non-globular Shapes

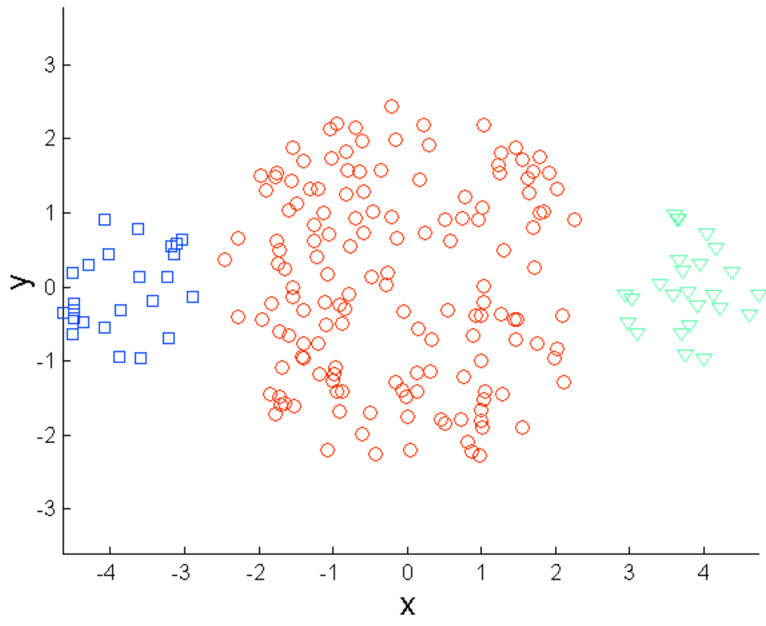



Original Points

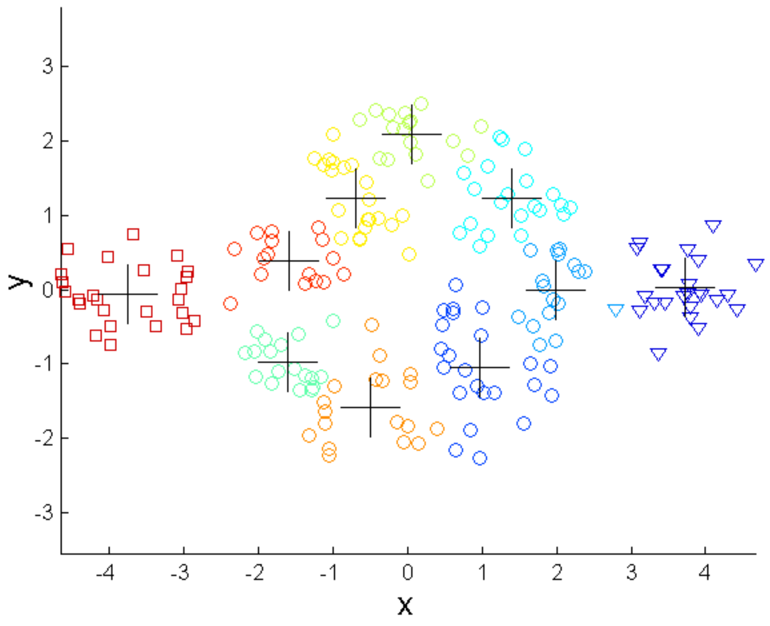



K-means (2 Clusters)

# Overcoming K-means Limitations



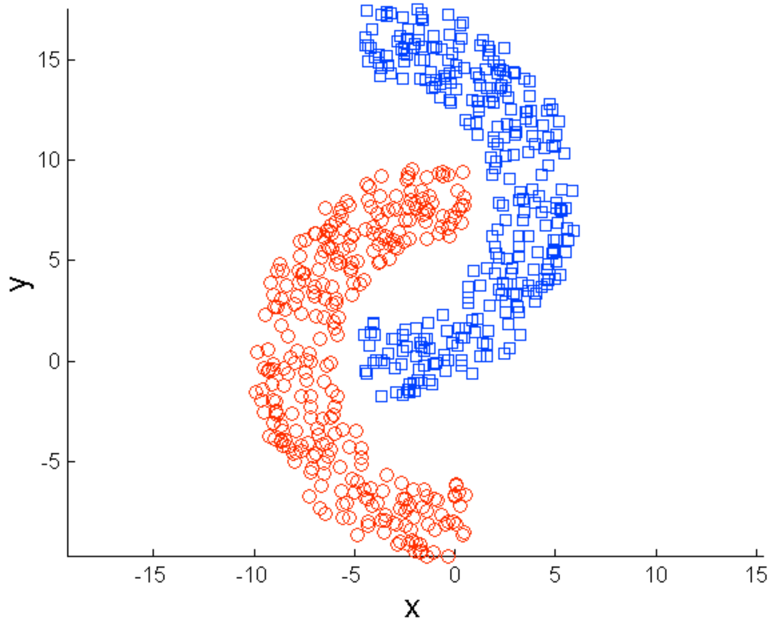

**Original Points**



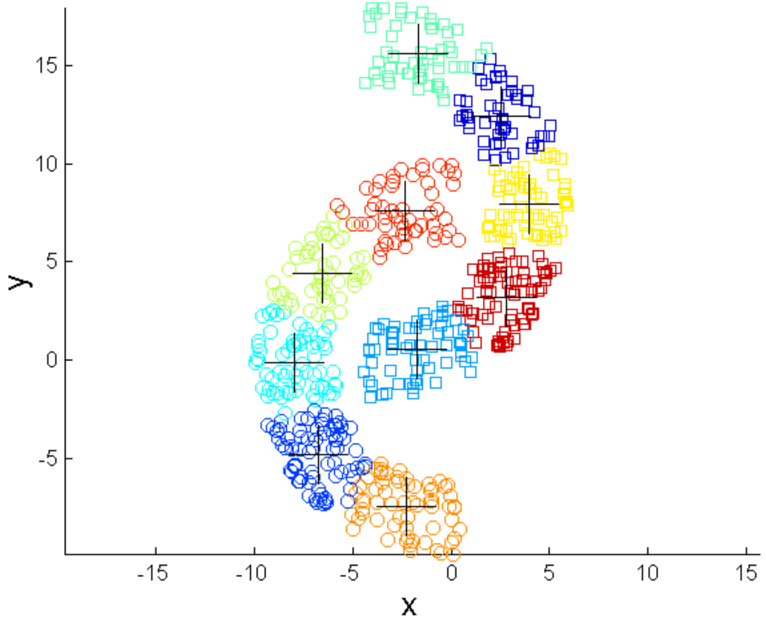

**K-means Clusters**

One solution is to use many clusters. Find parts of clusters, but need to put together.

# Overcoming K-means Limitations




Original Points




K-means Clusters

# Thank You!

Slide Courtesy: **Gary M. Weiss, CIS Dept, Fordham University and miscellaneous**