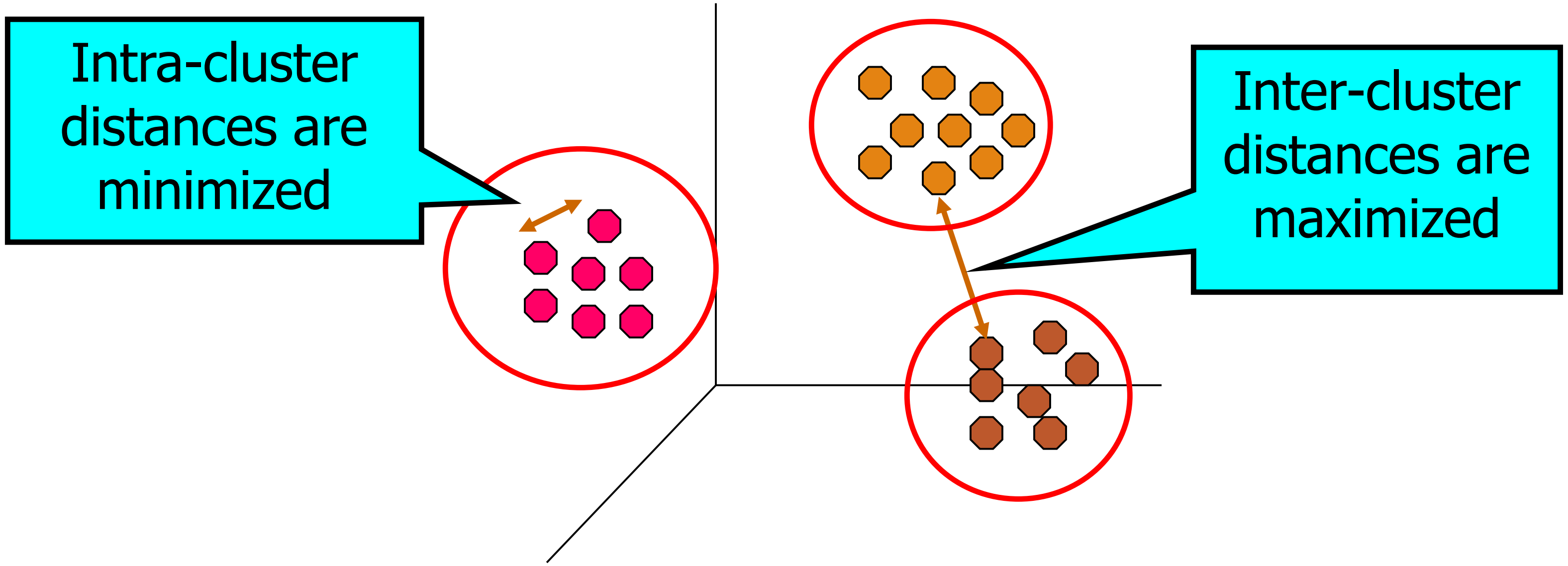
Clustering

What is Clustering ?

Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects.

- Objects in a class will be
  - Similar (or related) to one another
  - Different from (or unrelated to) the objects in other groups
- It is also called unsupervised learning.
- It is a common and important task that finds many applications.



Applications of Cluster Analysis

Applications in Search engines:

- Understanding
  - Group related documents that are similar
- Summarization
  - Reduces the size of large data sets
- Structuring search results
- Suggesting related pages
- Automatic directory construction/update
- Finding near identical/duplicate pages

	Discovered Clusters (content-based)	Group
1	Bank ; River bank	Geography
2	The Bank of the River	Fiction
3	The Left Bank at River Oak Rentals	Apartments

Text Clustering in Search

Clustering can be done at:

- Indexing time
- At query time
  - Applied to documents
  - Applied to snippets

Clustering can be based on:

URL source

Put pages from the same server together

Text Content

-Polysemy (“bat”, “banks”)

-Multiple aspects of a single topic

Links

-Look at the connected components in the link graph (A/H analysis can do it)

-look at co-citation similarity (e.g. as in collaborative filtering

What is not Cluster Analysis?

- Supervised classification
  - Have class label information
- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Graph partitioning
  - Some mutual relevance and synergy, but areas are not identical

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

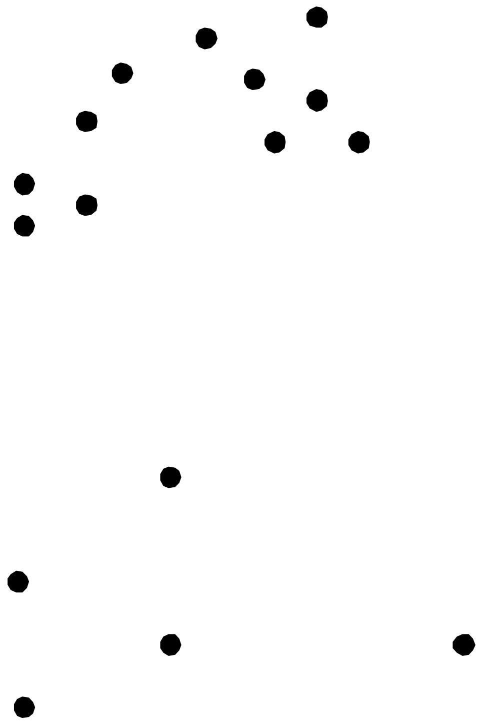
Two Clusters

Four Clusters

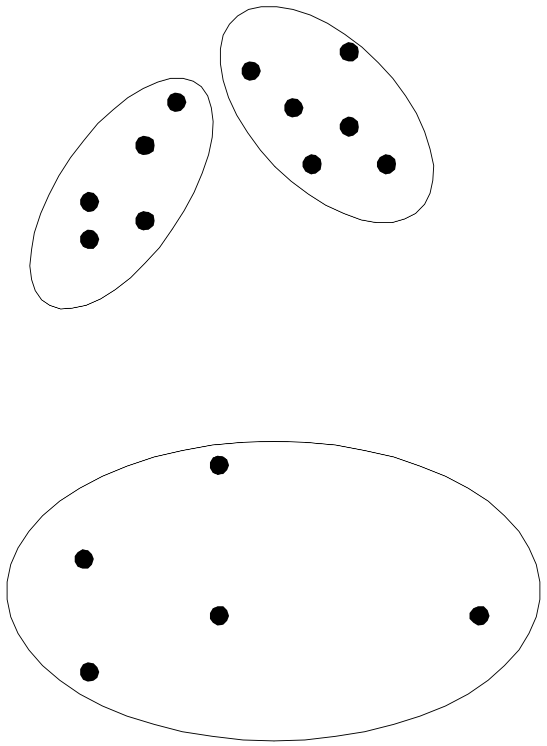
Types of Clustering

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

Partitional Clustering

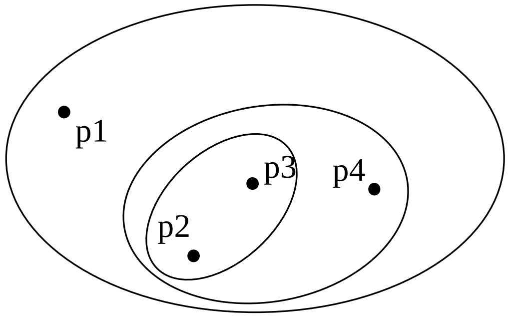


Original Points

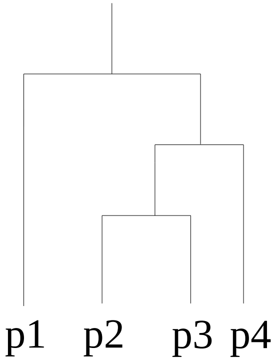


A Partitional Clustering

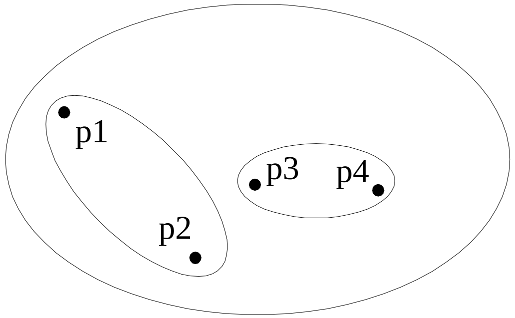
Hierarchical Clustering



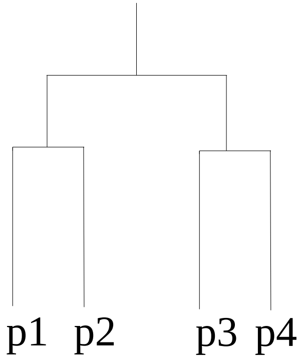
Traditional Hierarchical Clustering



Traditional Dendrogram

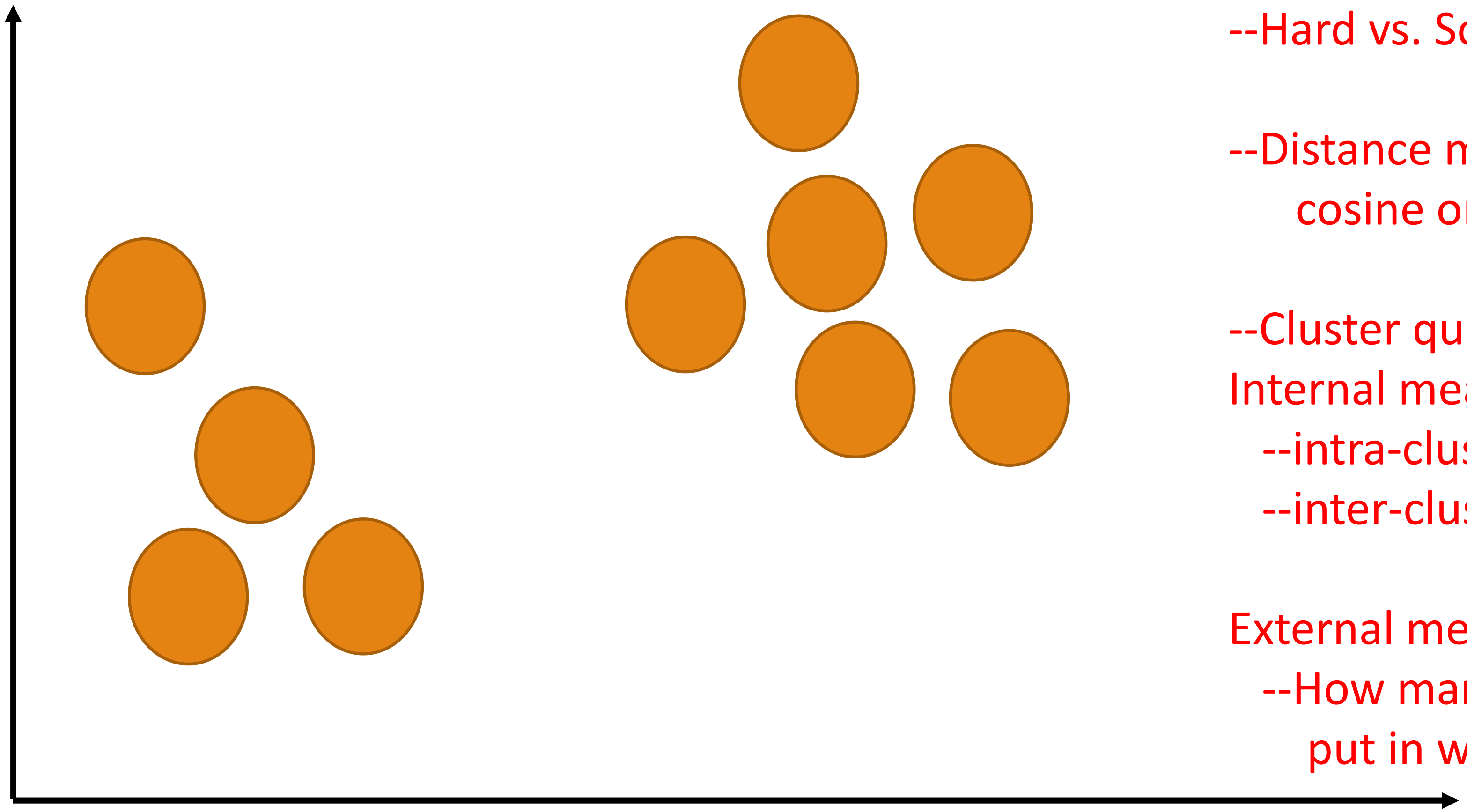


Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

Clustering issues



--Hard vs. Soft clusters

--Distance measures
cosine or Jaccard or..

--Cluster quality:
Internal measures
--intra-cluster tightness
--inter-cluster separation

External measures
--How many points are
put in wrong clusters.

Cluster Evaluation

- Clusters can be evaluated with “internal” as well as “external” measures
  - Internal measures are related to the inter/intra cluster distance
    - A good clustering is one where
      - (Intra-cluster distance) the sum of distances between objects in the same cluster are minimized,
      - (Inter-cluster distance) while the distances between different clusters are maximized
      - Objective to minimize: F(Intra, Inter)
  - External measures are related to how representative are the current clusters to “true” classes. Measured in terms of purity, entropy or F-measure
    - Note that in real world, you often *don’t know* what the true classes are. (This is why clustering is called unsupervised learning)

Inter and Intra Cluster Distances

Intra-cluster distance/tightness

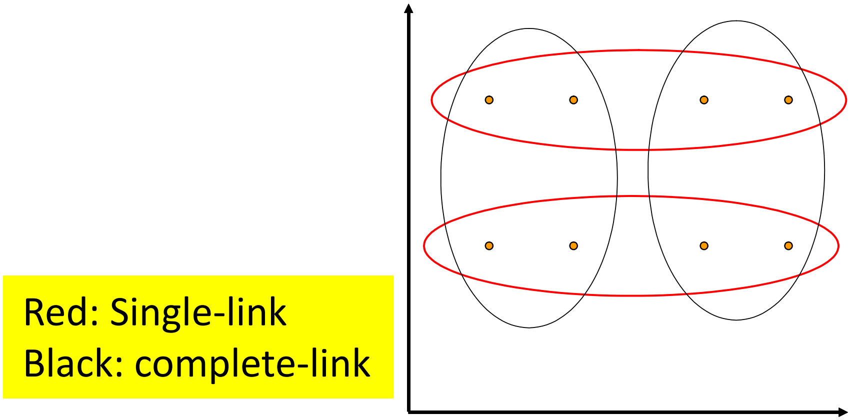
(Sum/Min/Max/Avg) the (absolute/squared) distance between

- All pairs of points in the cluster OR
- "diameter"—two farthest points
- Between the centroid /medoid and all points in the cluster

Inter-cluster distance

Sum the (squared) distance between all pairs of clusters
Where distance between two clusters is defined as:

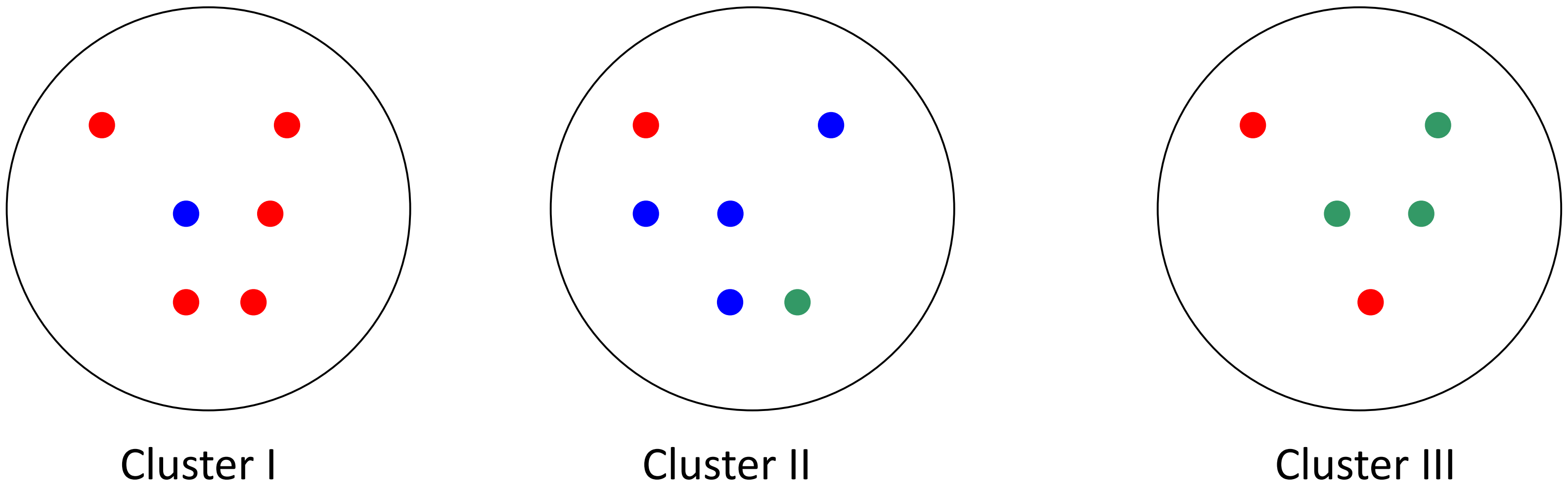
- distance between their centroids/medoids
- Distance between farthest pair of points (complete link)
- Distance between the closest pair of points belonging to the clusters (single link)



Cluster Evaluations

- Clusters can be evaluated with “internal” as well as “external” measures
  - Internal measures are related to the inter/intra cluster distance
    - A good clustering is one where
      - (Intra-cluster distance) the sum of distances between objects in the same cluster are minimized,
      - (Inter-cluster distance) while the distances between different clusters are maximized
      - Objective to minimize: F(Intra, Inter)
  - External measures are related to how representative are the current clusters to “true” classes. Measured in terms of purity, entropy or F-measure

Cluster Purity (Given gold standard classes)



Pure size of a cluster = # elements from the majority class

Purity of clustering: $\dfrac{\text{Sum of pure sizes of clusters}}{\text{Total number of elements across clusters}}$

= (5 + 4 + 3)/ (6 + 6 + 5) = 12/17 = 0.71

Will it work if you allow # of clusters to increase?

Rand Index Example

Compare to Standard Precision & Recall

The following table classifies all pairs of entities (of which there are n choose 2) into one of four classes

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	TP	FN
Different classes in ground truth	FP	TN

Is the cluster putting non-class items in?

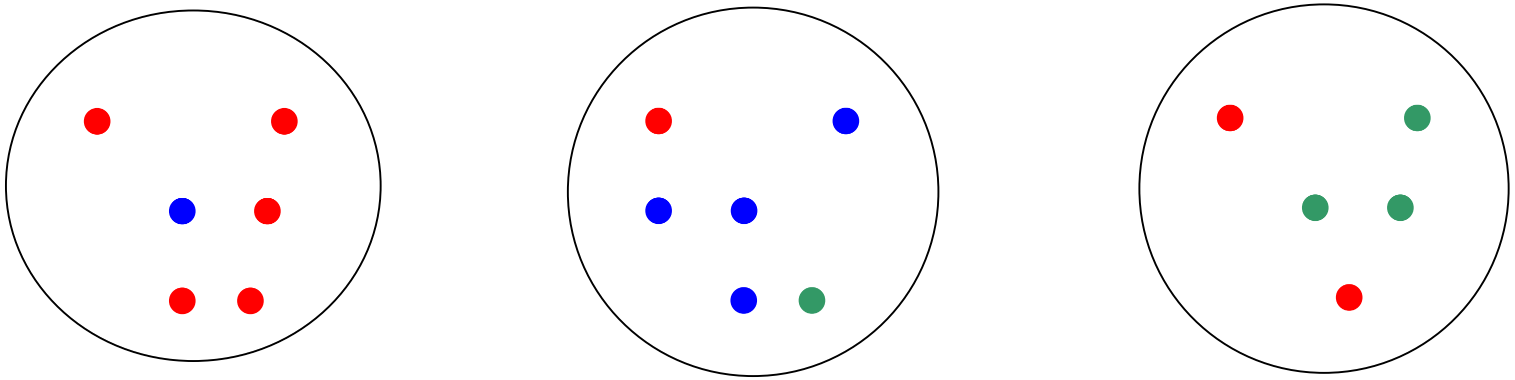
$$P = \frac{TP}{TP + FP}$$

Is the cluster missing any in-class items?

$$R = \frac{TP}{TP + FN}$$

$$RI = \frac{TP + TN}{TP + TN + FP + FN}$$

Rand Index Example



Cluster I

Cluster II

Cluster III

RI= 20+72/(20+20+24+72) =0.68

Number of points	Same Cluster in clustering	Different Clusters in clustering
Same class in ground truth	20	24
Different classes in ground truth	20	72

Elementary combinatorics

TP+FP (total pairs in the same clusters)

= 6 C 2 + 6 C 2 + 5 C 2 = 40

To get TP

= 5 C 2 + 4 C 2 + 3 C 2 + 2 C 2 = 20

You can compute FN/TN similarly

Unsupervised?

Clustering is normally seen as an instance of unsupervised learning algorithm

- So how can you have external measures of cluster validity?
- The truth is that you have a continuum between unsupervised vs. supervised
  - Answer: Think of “no teacher being there” vs. “lazy teacher” who checks your work once in a while.
  - Examples:
    - Fully unsupervised (no teacher)
    - Teacher tells you how many clusters are there
    - Teacher tells you that certain pairs of points will fall or will not fill in the same cluster
    - Teacher may occasionally evaluate the goodness of your clusters (external measures of validity)

How hard is clustering?

One idea is to consider all possible clusterings, and pick the one that has best inter and intra cluster distance properties

Suppose we are given n points, and would like to cluster them into k-clusters

- How many possible clusterings?

$$\sum_{k=1}^{n} \frac{k^n}{k!}$$

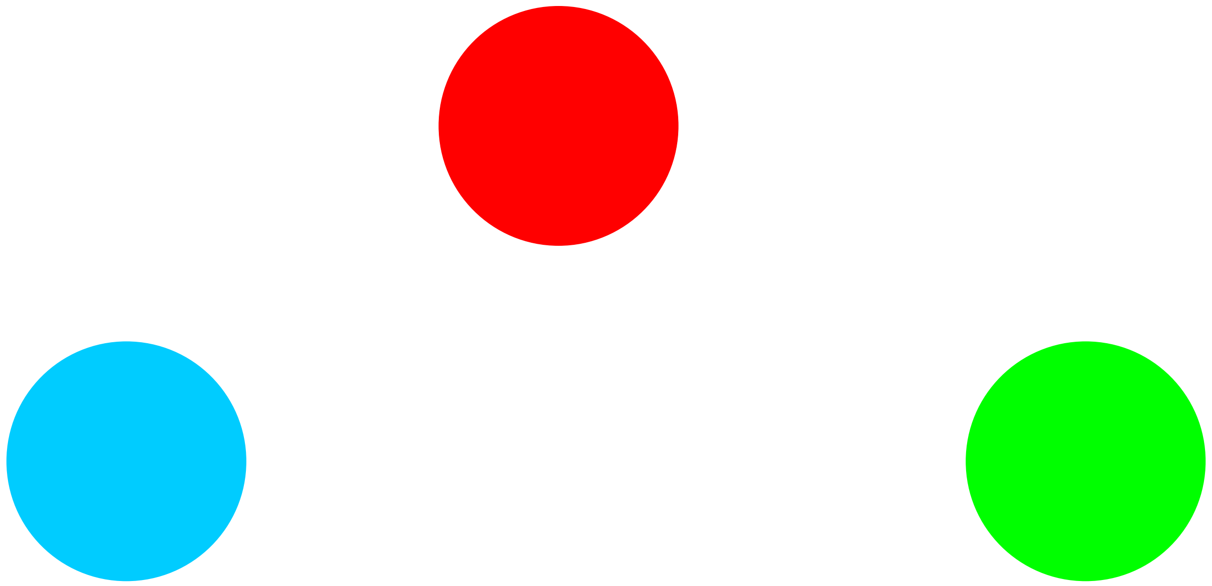
- Too hard to do it brute force or optimally
- Solution: Iterative optimization algorithms
  - Start with a clustering, iteratively improve it (eg. K-means)

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

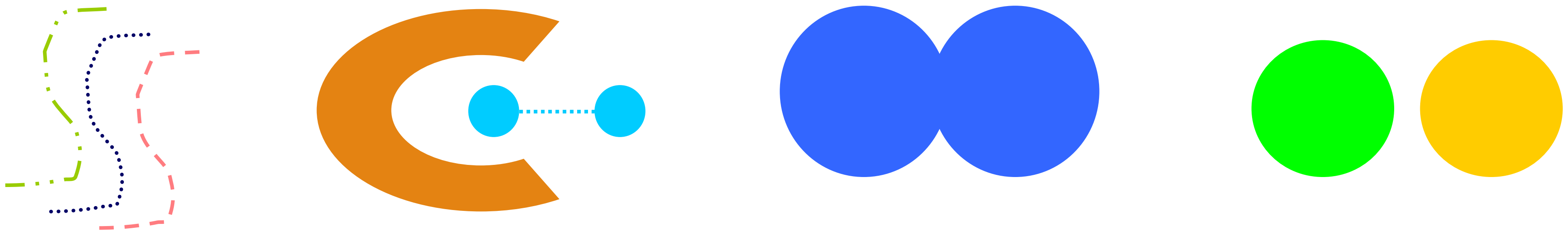
Types of Clusters: Center-Based

- Center-based
  - A cluster is a set of objects such that an object in a cluster is close (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

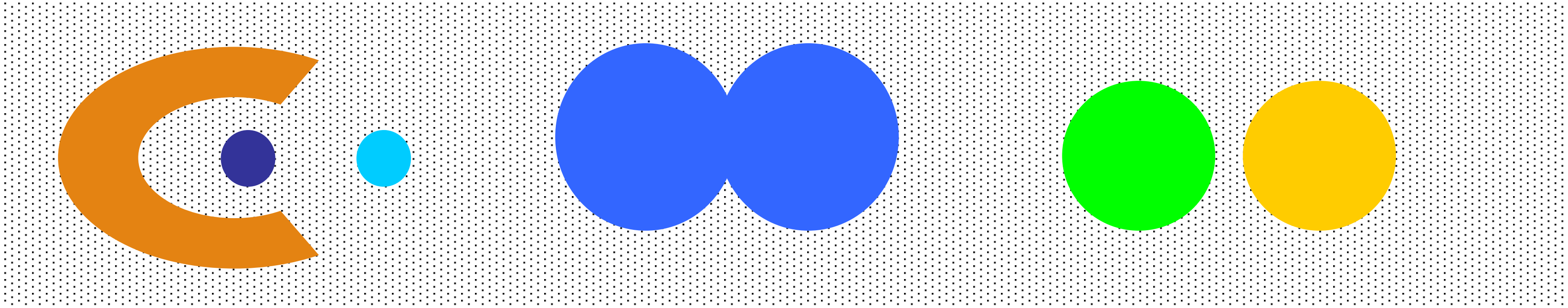
- Contiguous Cluster (Nearest Neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is close (or more similar) to one or more other points in the cluster than to any point not in the cluster



8 contiguous clusters

Types of Clusters: Density-Based

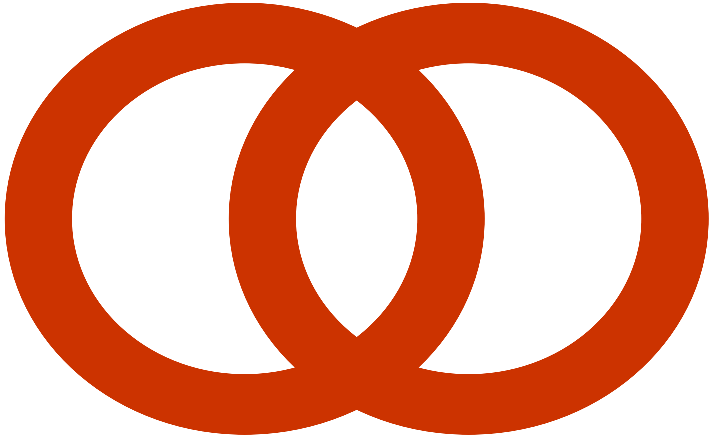
- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density
  - Used when the clusters are irregular or inter-twined, and when noise and outliers are present



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

Types of Clusters: Objective Function

- Clusters defined by an objective function
  - Find clusters that minimize or maximize an objective function
  - Enumerate all possible ways of dividing the points into clusters and evaluate the 'goodness' of each potential set of clusters by using the given objective function (NP Hard)
  - Can have global or local objectives
  - A variation of the global objective function approach is to fit the data to a parameterized model
    - Parameters for the model are determined from the data
    - Mixture models assume that the data is a 'mixture' of a number of statistical distributions
- Map the clustering problem to a different domain and solve a related problem in that domain
  - Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points
  - Clustering is equivalent to breaking the graph into connected components, one for each cluster
  - Want to minimize the edge weight between clusters and maximize the edge weight within clusters

Characteristics of the Input Data are Important

- Type of proximity or density measure
  - This is a derived measure, but central to clustering
- Sparseness
  - Dictates type of similarity
  - Adds to efficiency
- Attribute type
  - Dictates type of similarity
- Type of data
  - Dictates type of similarity
  - Other characteristics e.g., autocorrelation
- Dimensionality
- Noise and outliers
- Type of distribution

Classical clustering methods

Partitioning methods

- k-Means (and EM), k-Medoids

Hierarchical methods

- agglomerative, divisive, BIRCH

Model-based clustering methods

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

K-Means Clustering Algorithm

Works when we know k, the number of clusters we want to find

Algorithm:

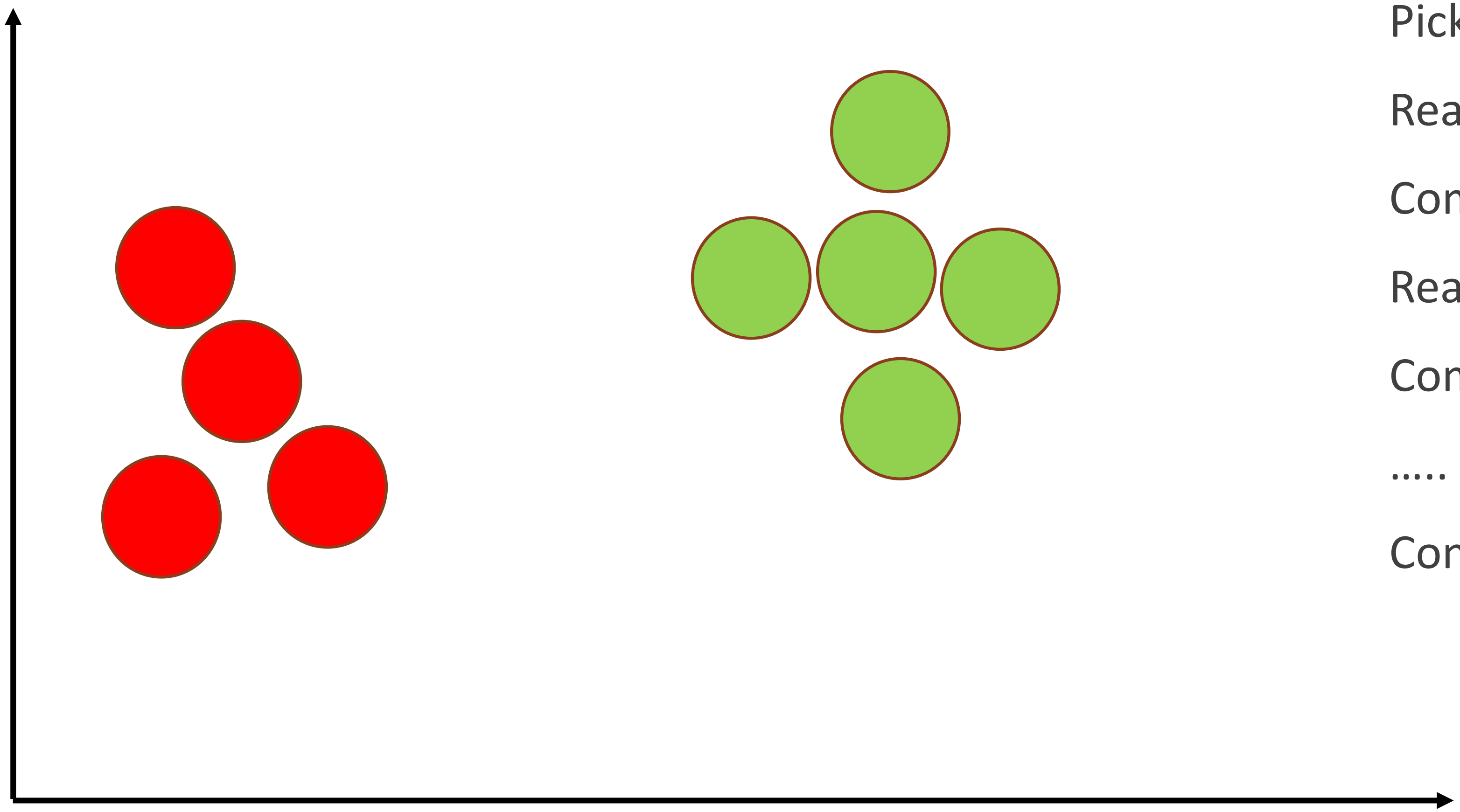
Select K points as the initial centroids.
repeat
  Form K clusters by assigning all points to the closest centroid.
  Recompute the centroid of each cluster.
until The centroids don't change

Iterative improvement of the objective function:

Sum of the squared distance (or Error -- SSE) from each point to the centroid of its cluster

(Notice that since K is fixed, maximizing tightness also maximizes inter-cluster distance)

K-means Example (k=2)



Pick seeds

Reassign clusters

Compute centroids

Reassign clusters

Compute centroids

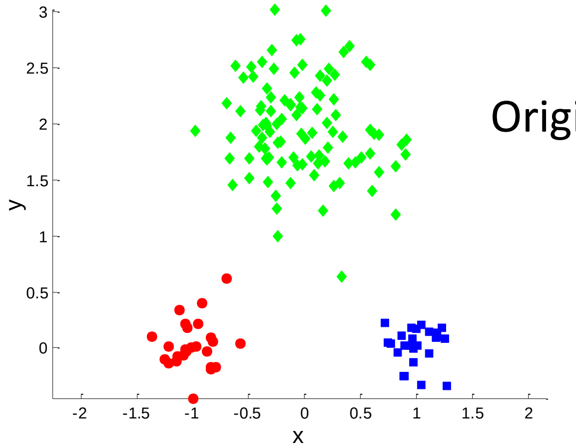
…..

Converged!

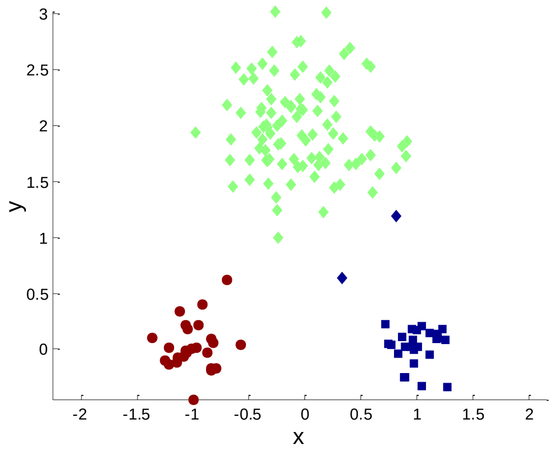
K-means Clustering Algorithm

- Initial centroids are often chosen randomly
- 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’
- Complexity is $O(n * K * I * d)$
  - n = number of data points
  - K = number of clusters
  - I = number of iterations
  - d = number of attributes

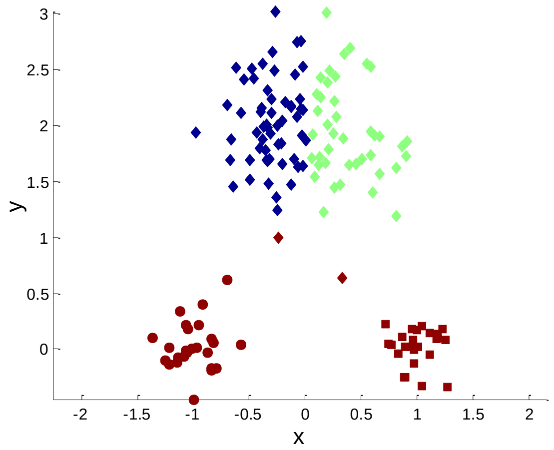
Two different K-means Clusterings



Original Points



Optimal Clustering



Sub-optimal Clustering

Evaluating K-means Clusters

Most common measure is Sum of Squared Error (SSE)

- For each point, the error is the distance to the nearest cluster
- To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- x is a data point in cluster C_i and m_i is the representative point for cluster C_i
  - can show that m_i corresponds to the center (mean) of the cluster
- Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase K, the number of clusters
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

Handling Empty Clusters

- Basic K-means algorithm can yield empty clusters
- Several strategies:
  - Choose the point that contributes most to SSE
  - Choose a point from the cluster with the highest SSE
  - If there are several empty clusters, the above can be repeated several times.

Updating Centers Incrementally

- In the basic K-means algorithm, centroids are updated after all points are assigned to a centroid
- An alternative is to update the centroids after each assignment (incremental approach)
  - Each assignment updates zero or two centroids
  - More expensive
  - Introduces an order dependency
  - Never get an empty cluster
  - Can use “weights” to change the impact

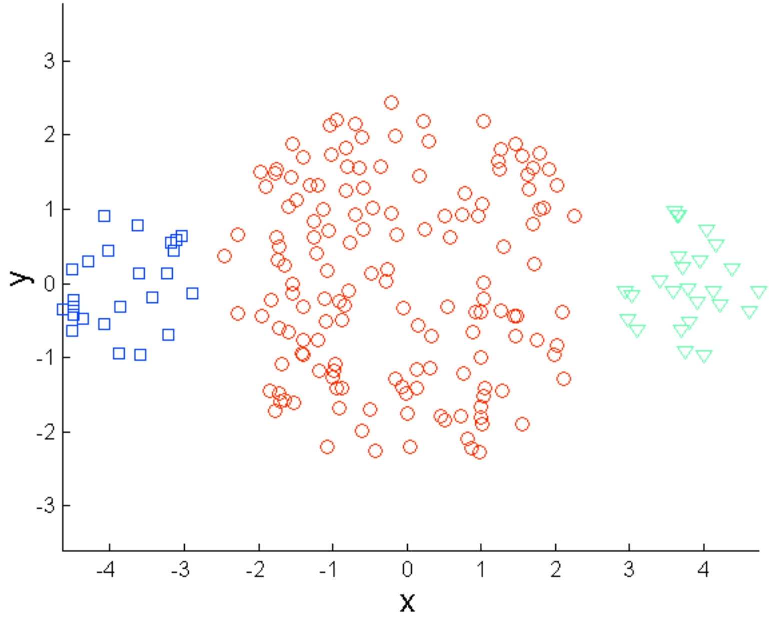
Limitations of K-means

K-means has problems when clusters are of differing

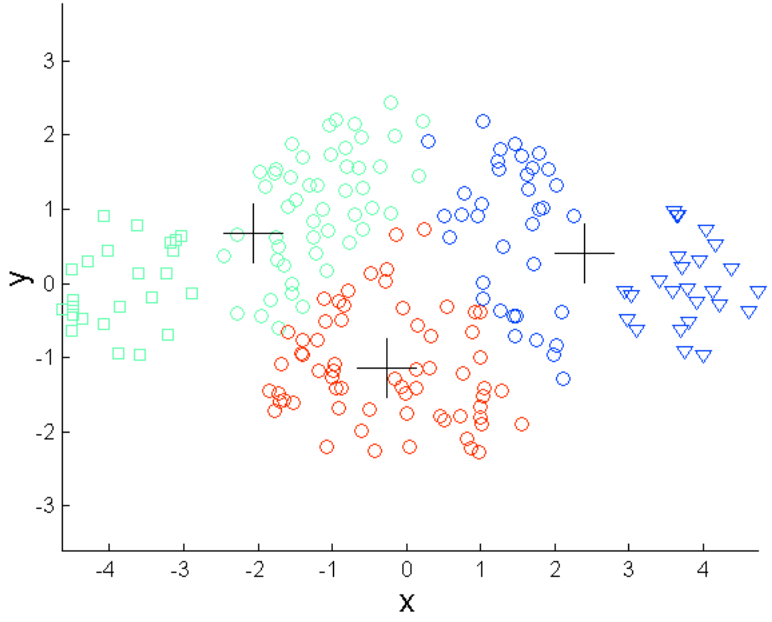
- Sizes
- Densities
- Non-globular shapes

K-means has problems when the data contains outliers.

Limitations of K-means: Differing Sizes

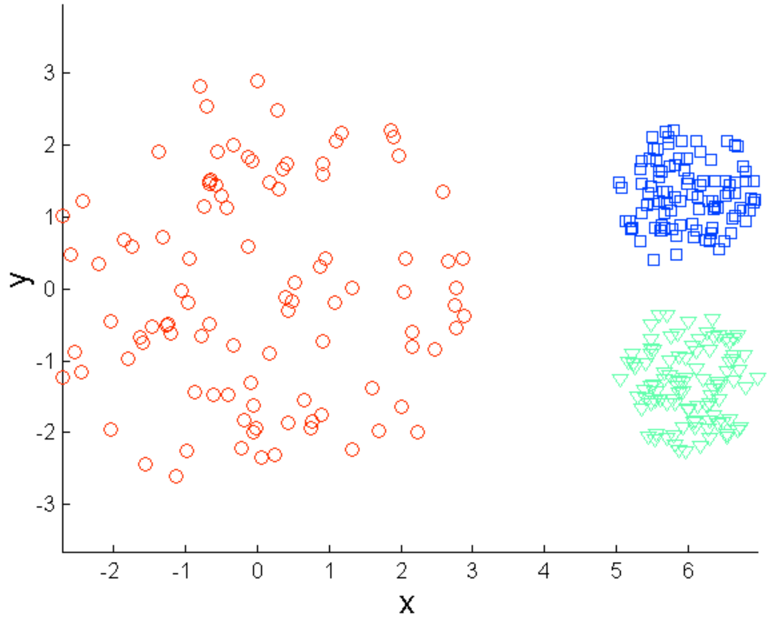


Original Points

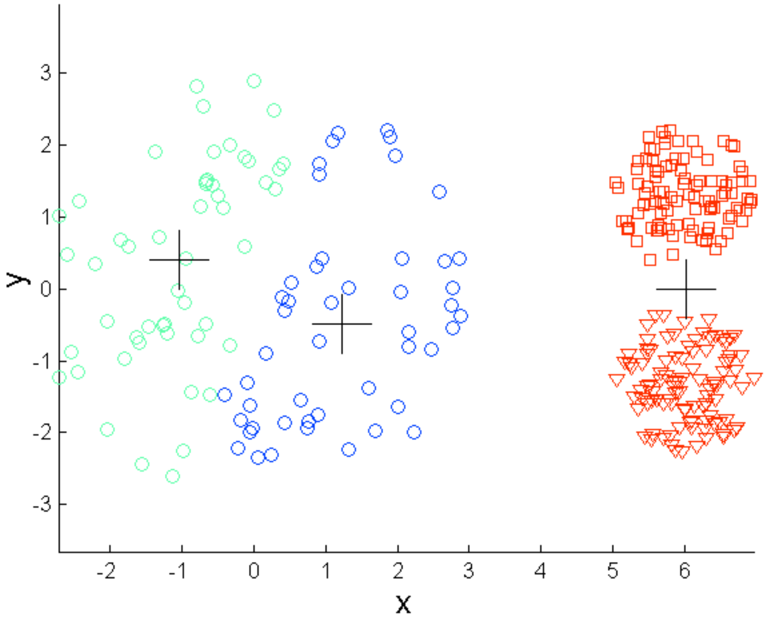


K-means (3 Clusters)

Limitations of K-means: Differing Density

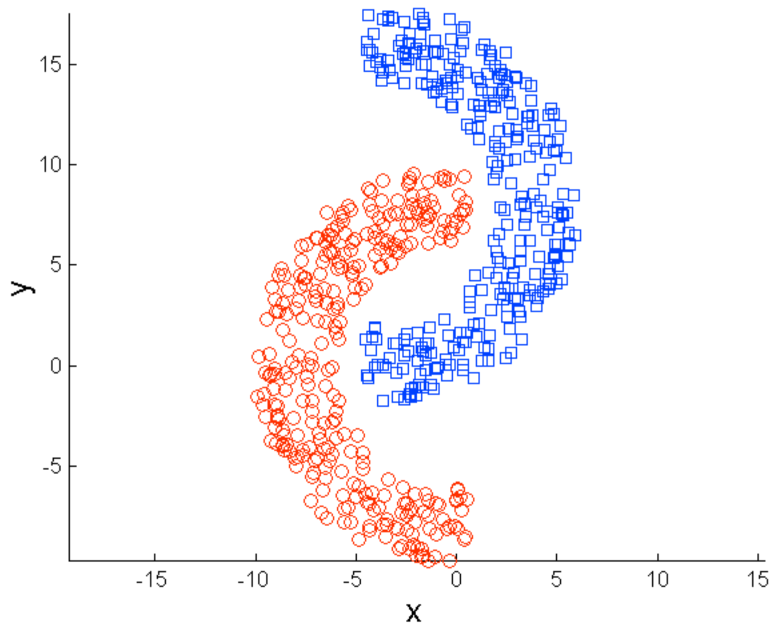


Original Points

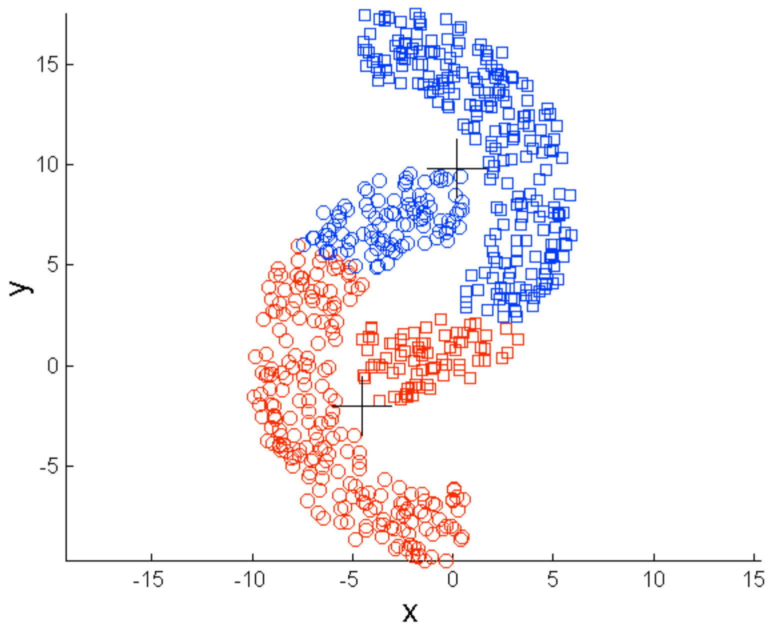


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes



Original Points

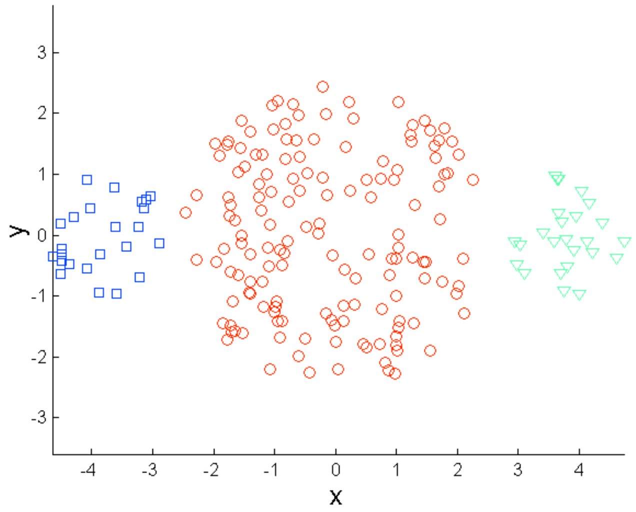


K-means (2 Clusters)

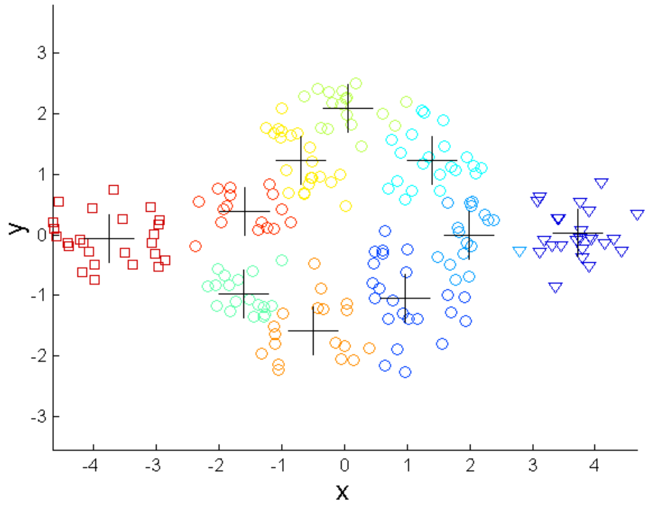
Overcoming K-means Limitations

One solution is to use many clusters

Find parts of clusters, but need to put together

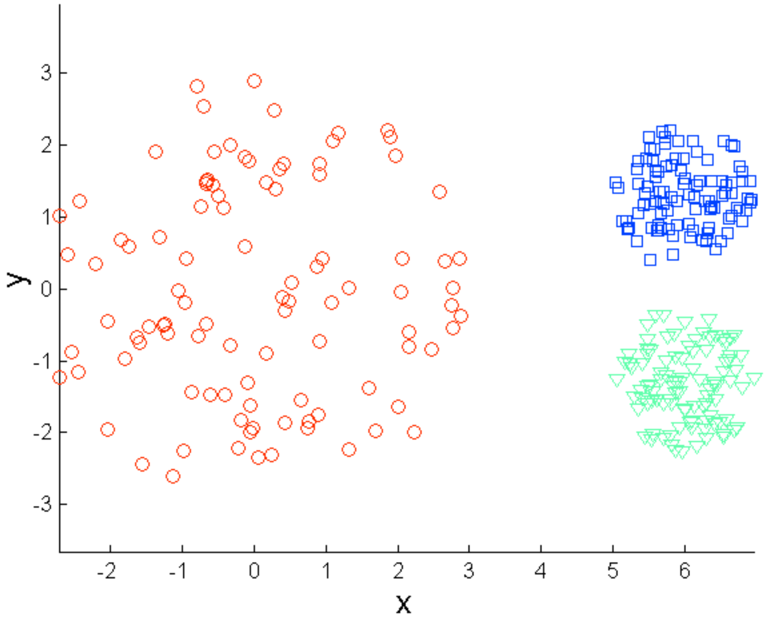


Original Points

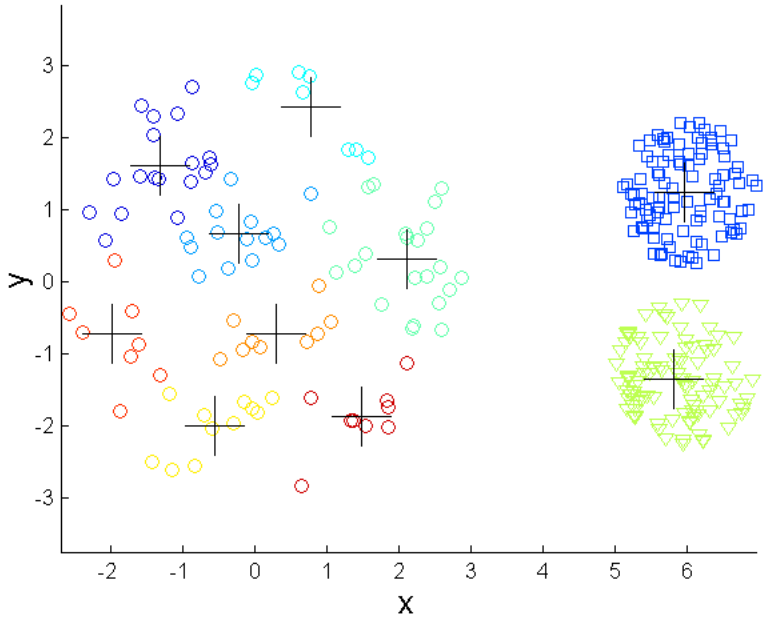


K-means Clusters

Overcoming K-means Limitations

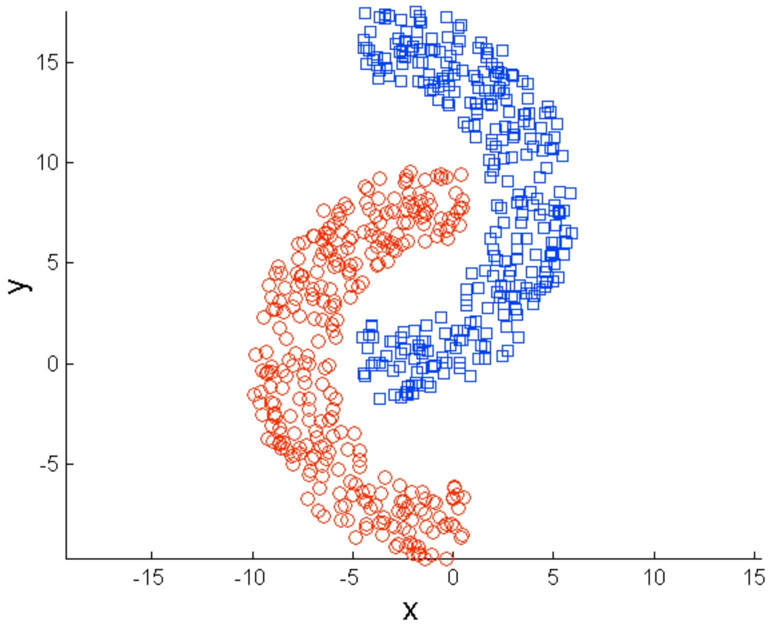


Original Points

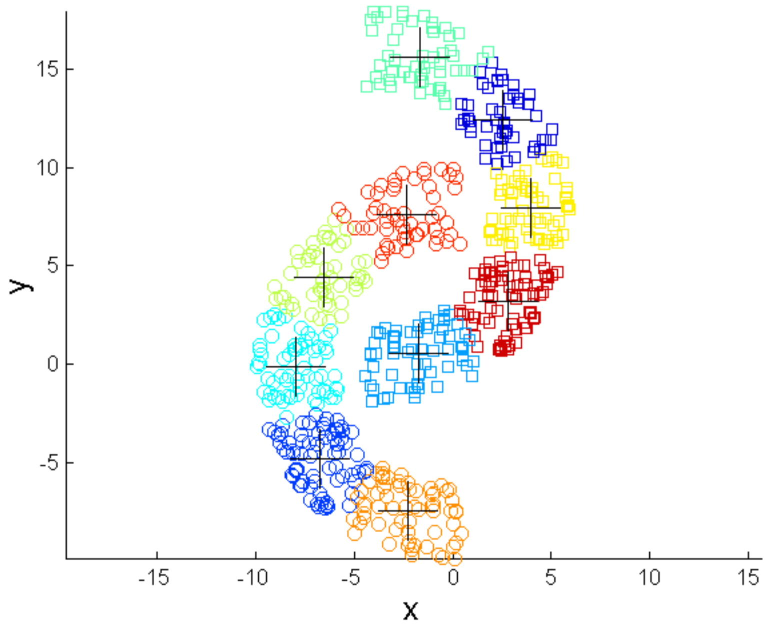


K-means Clusters

Overcoming K-means Limitations



Original Points



K-means Clusters