Data Clustering: K means method

CPS 563 – Data Visualization

Dr. Tam Nguyen

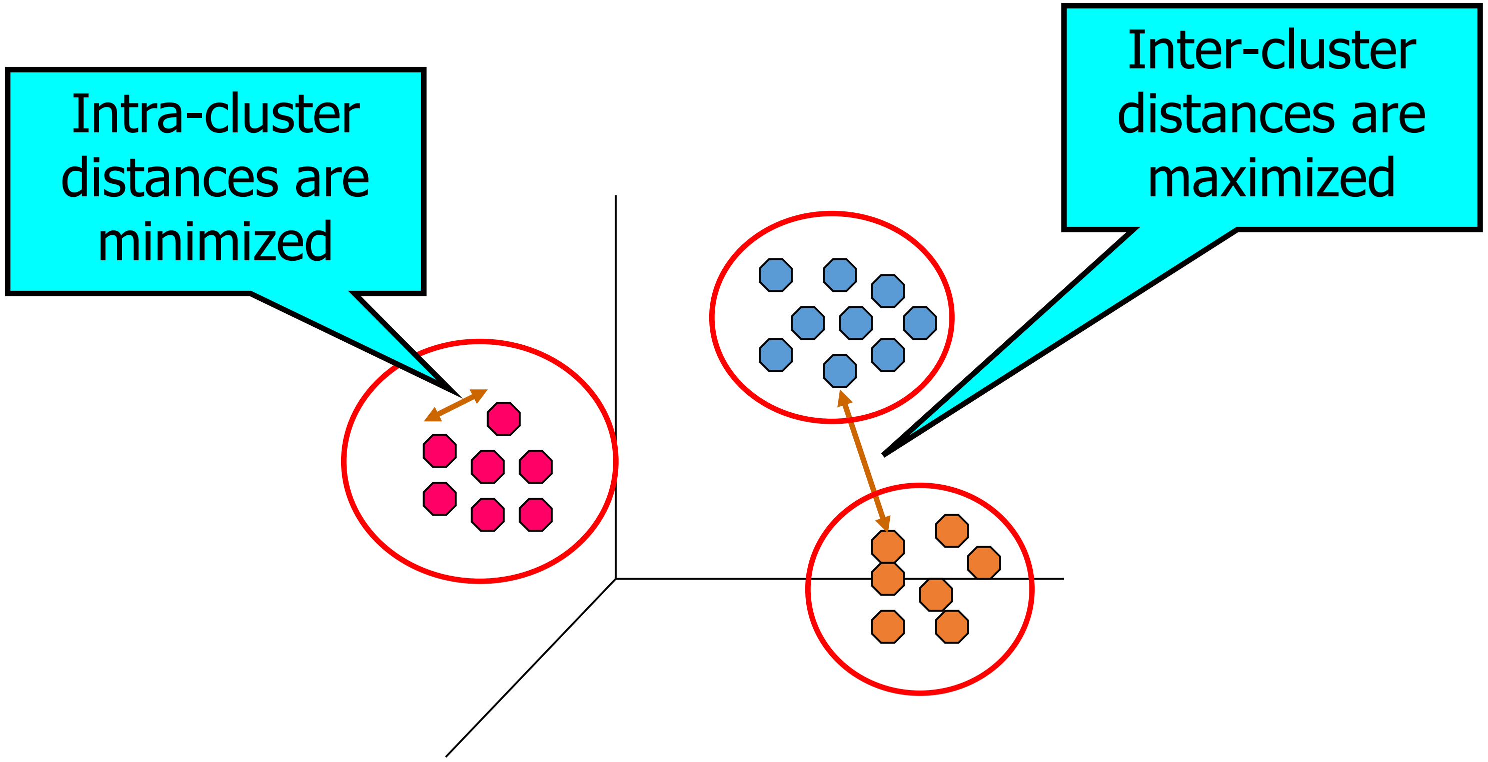
tamnguyen@udayton.edu

University of Dayton

What is Cluster Analysis?

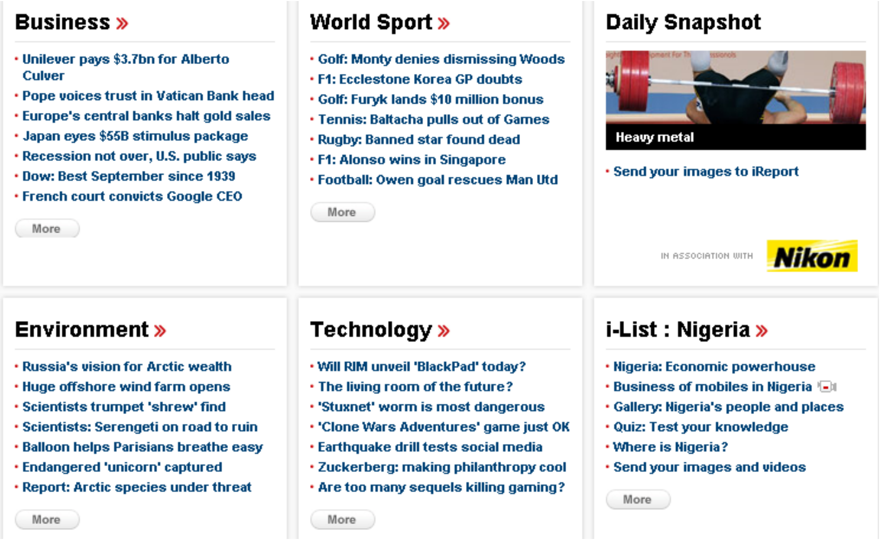
Implicit class label, not pre-defined!

- 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

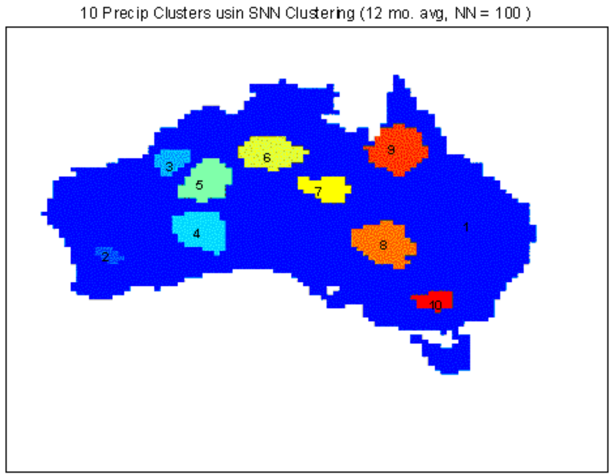


Applications of Cluster Analysis

- **Better understanding & search**
  - Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations



- **Visualization**
  - Reduce the size of large data sets



Clustering rain fall amount in Australia

Notion of a Cluster can be Ambiguous

How many clusters?

Six Clusters

Two Clusters

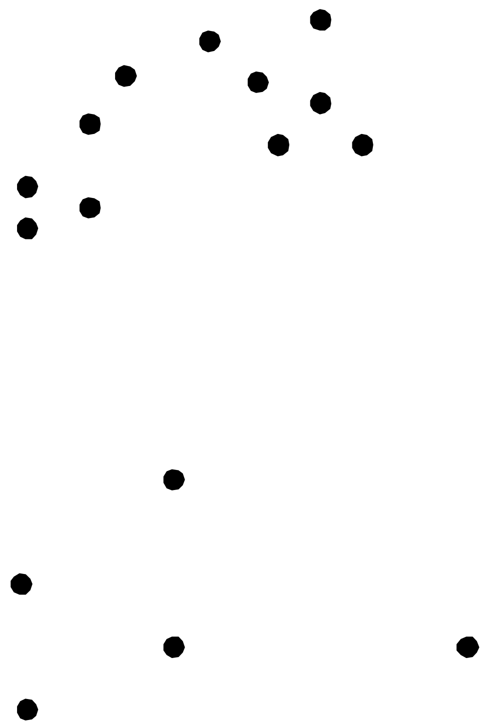
Four Clusters

Types of Clustering

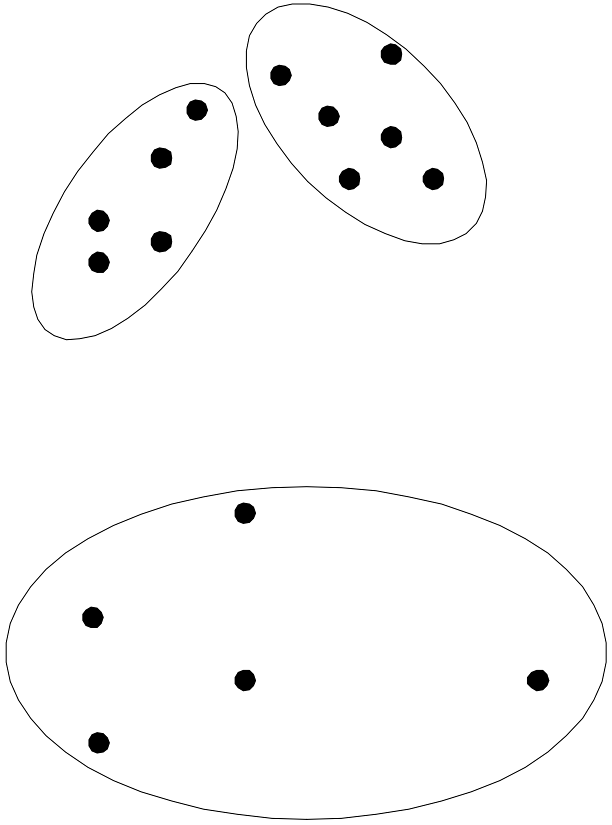
- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
- Partitional Clustering
  - A division of data points into non-overlapping subsets (clusters) such that each data point is in exactly one subset
- Hierarchical clustering
  - A set of nested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

Cluster number is determined as you like

Partitional Clustering

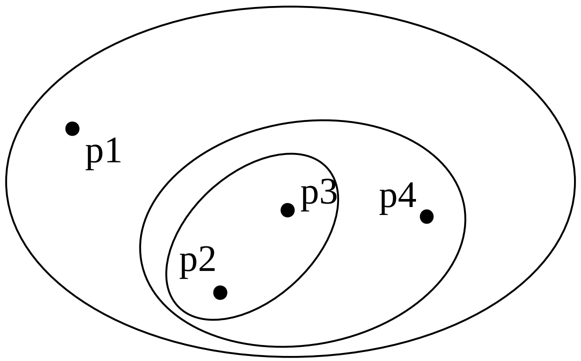


Original Points

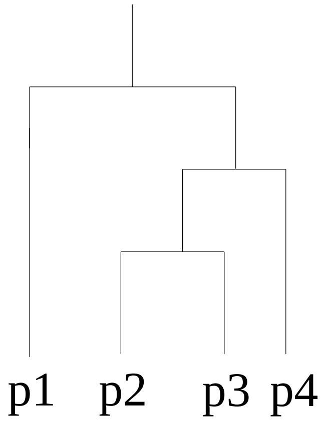


A Partitional Clustering

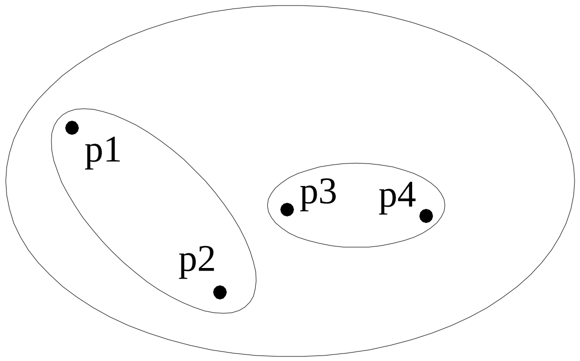
Hierarchical Clustering



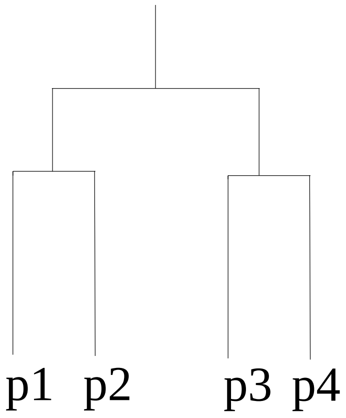
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

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

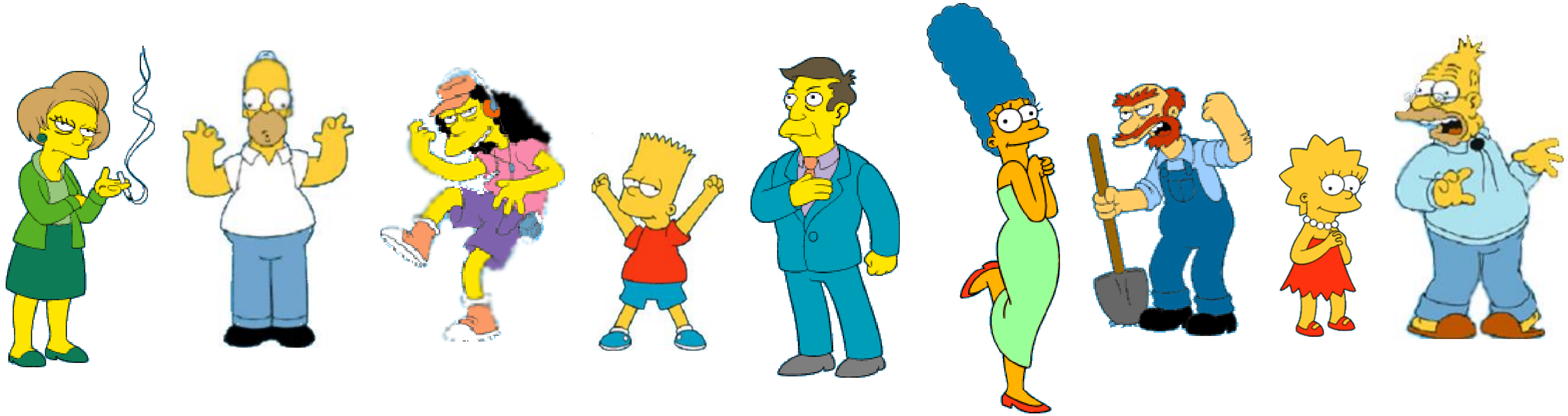
"How" is the key!
Discuss!

```
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
```

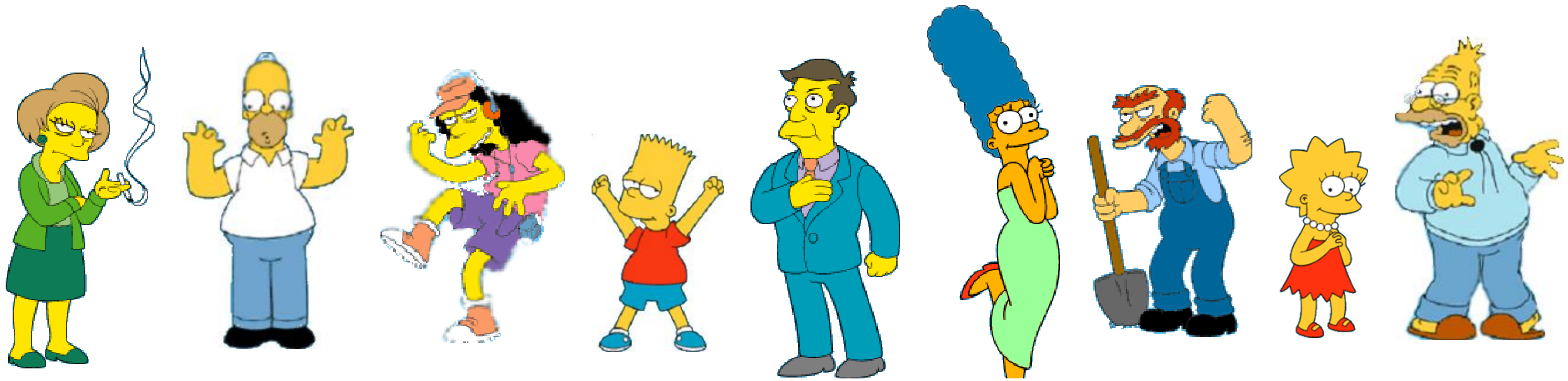
K-means Clustering – Details

- 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’
- Complexity is $O(n * K * I * d)$
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of features

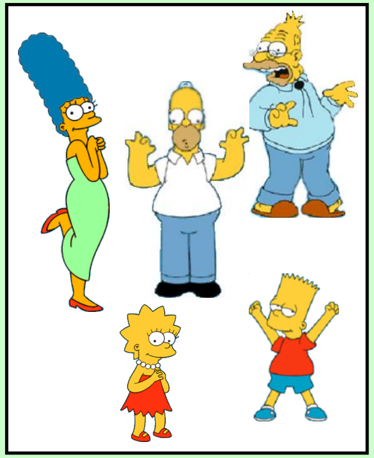
Why is choosing distance metric important?



What is a natural grouping among these objects?

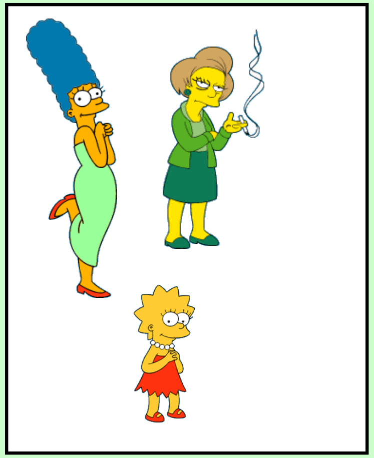


Clustering depends on the distance metric



Simpson's Family

School Employees



Females

Males

Another example

Clustering depends on the distance metric

Marvel DC Fly Run Billionaire Will be a billionaire

Distance Metrics

- The Minkowski metric is a generalization of a Euclidean distance:

$$L_p(\mathbf{a},\mathbf{b}) = \left(\sum_{k=1}^{d} \left| a_k - b_k \right|^p \right)^{1/p}$$

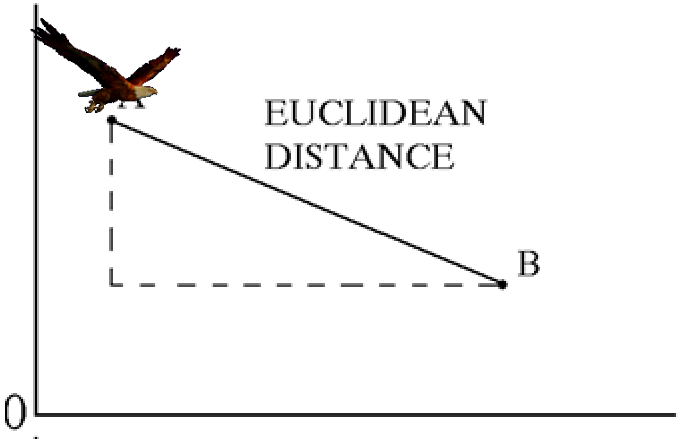
, where d is the number of feature dimensions, and is often referred to as the L_p norm.

- Special cases:
  - L_1: absolute, cityblock, or Manhattan distance
  - L_2: Euclidian distance

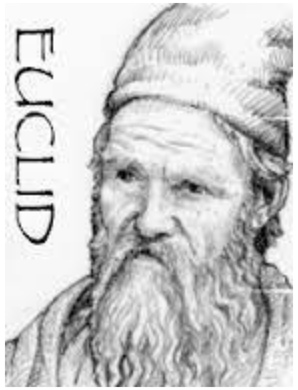
Distance Metrics

- Euclidean Distance: $$dist(\mathbf{a},\mathbf{b}) = \left(\sum_{k=1}^{d} \left(a_k - b_k \right)^2 \right)^{1/2}$$

x_2

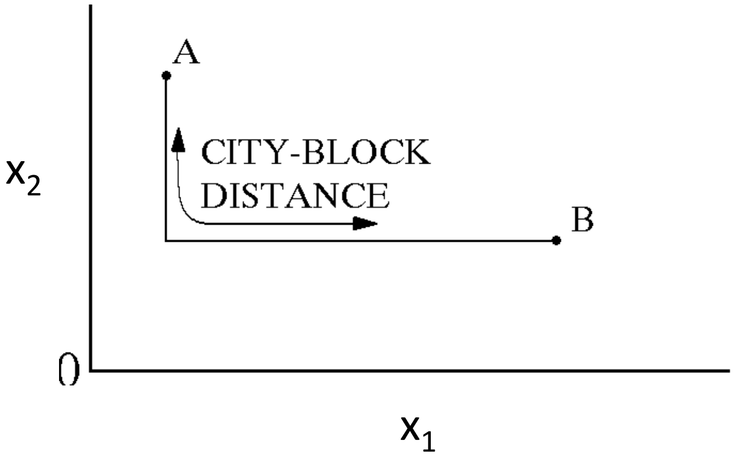


x_1



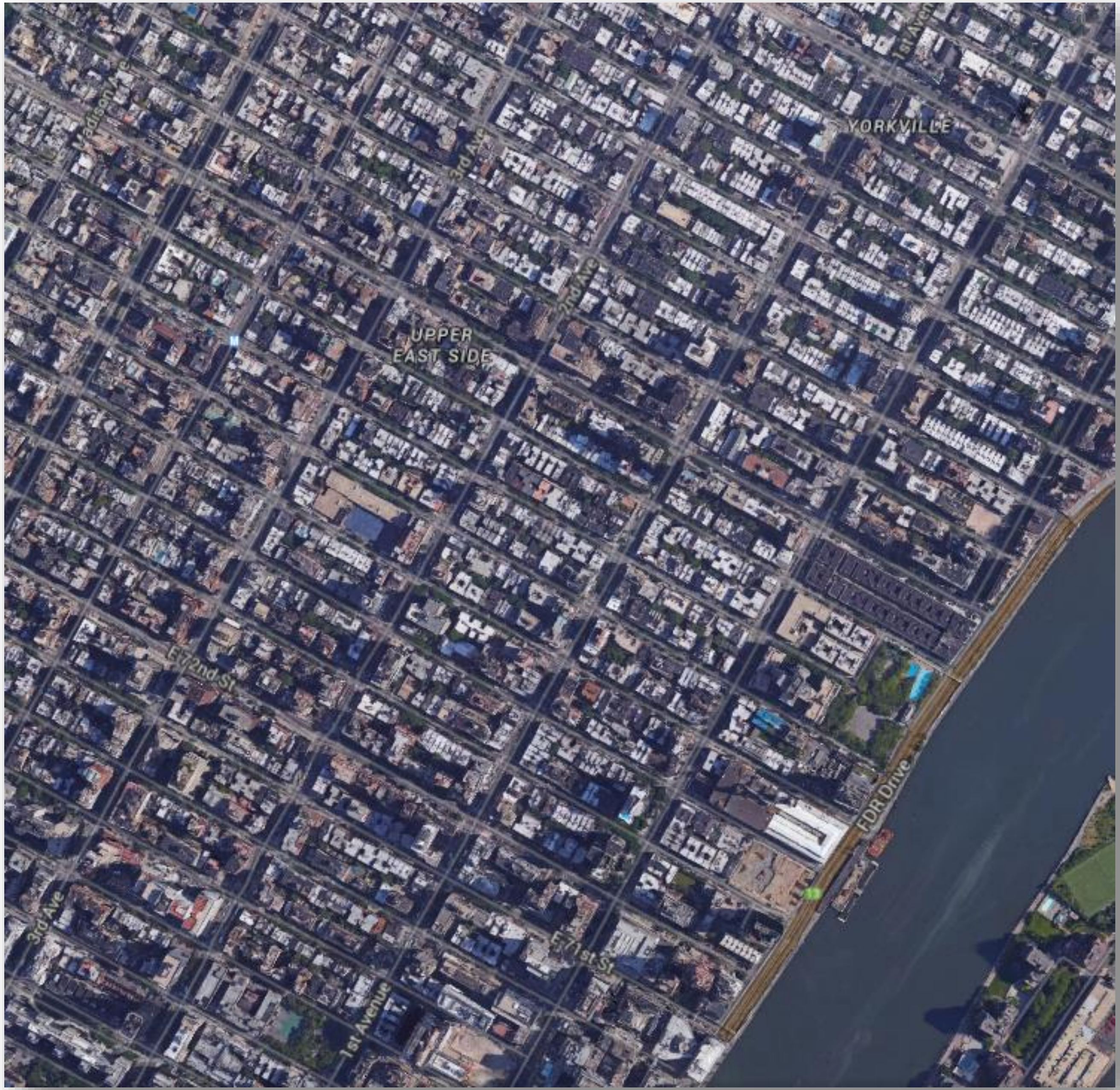
Distance Metrics

- Manhattan distance: $dist(\mathbf{a},\mathbf{b}) = \sum_{k=1}^{d} \left| a_k - b_k \right|$

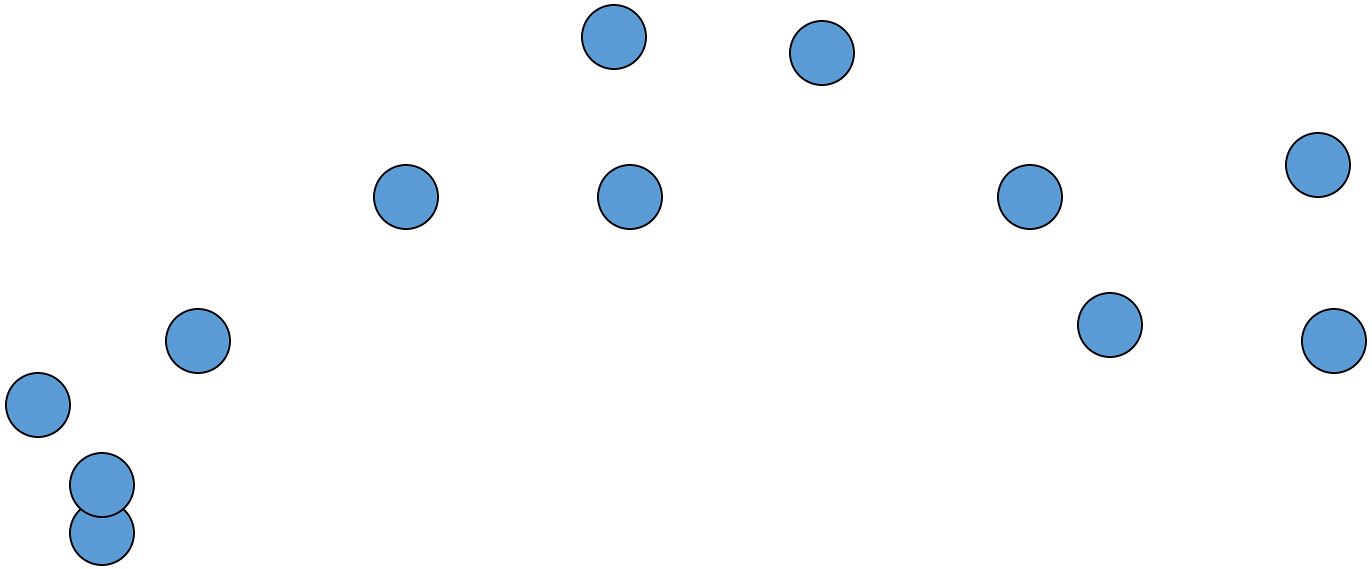


Distance Metrics

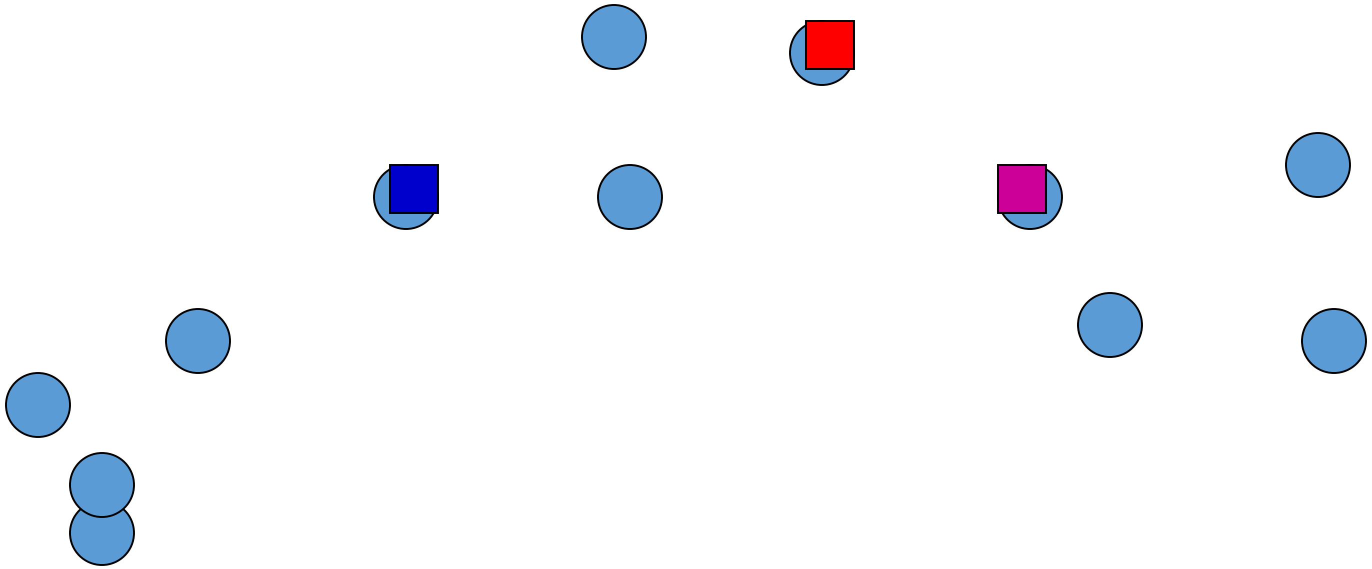
- It is named Manhattan distance because it is the shortest distance a car would drive in a city laid out in square blocks, like Manhattan.



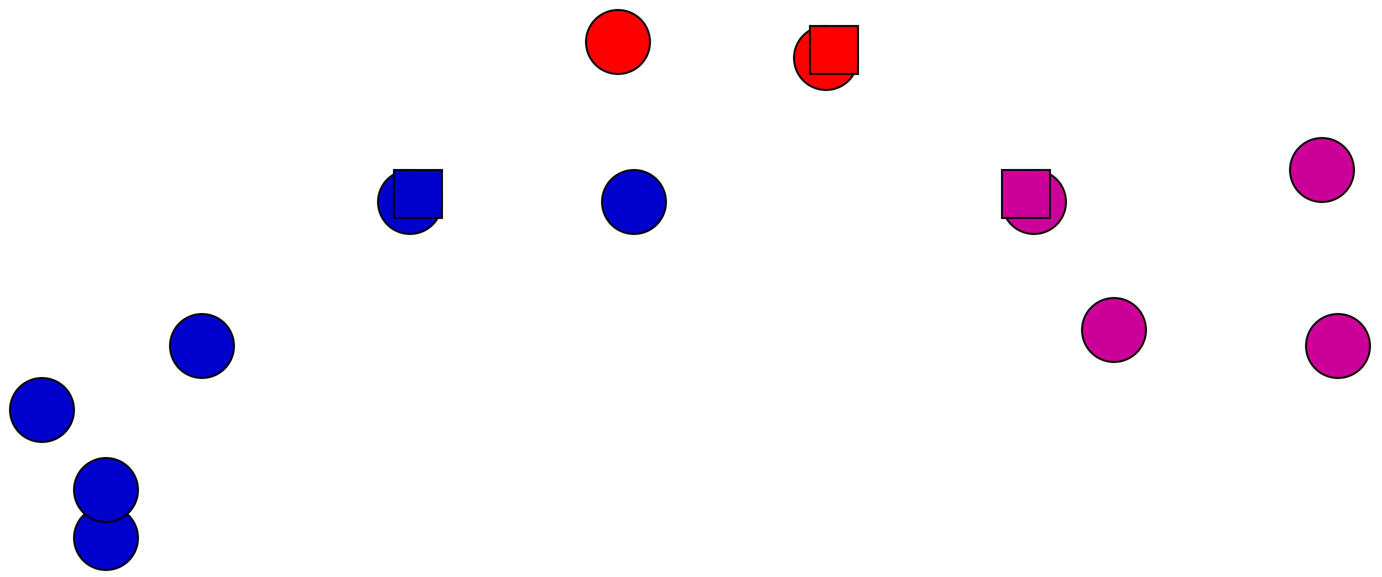
K-means: an example



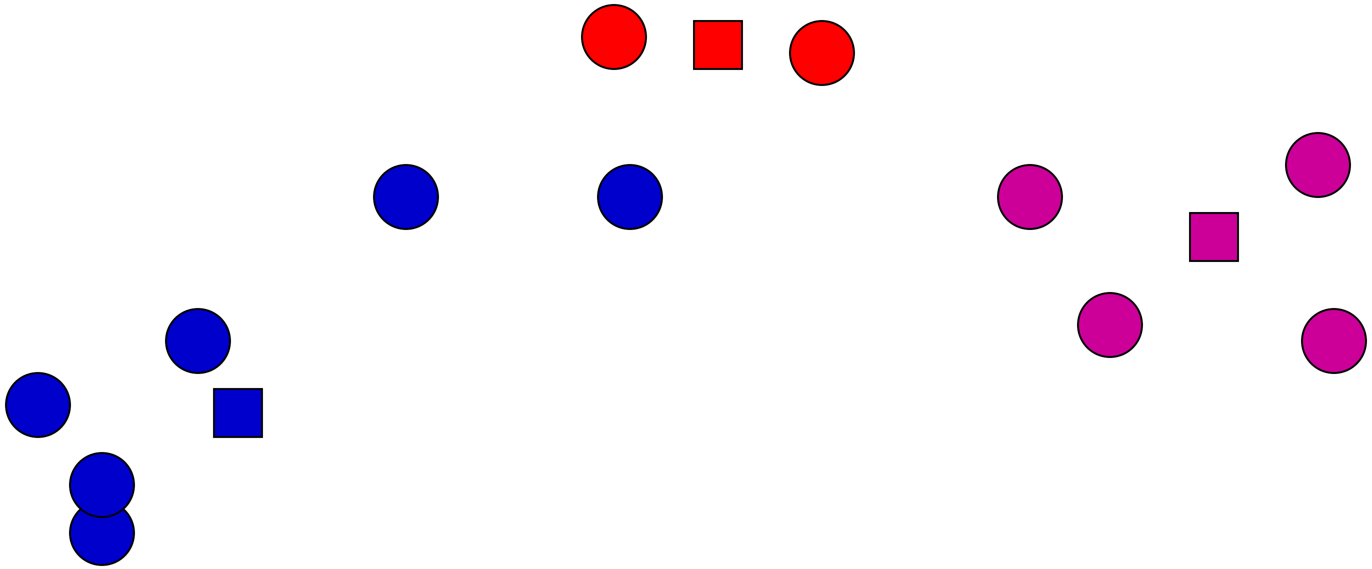
K-means: Initialize centers randomly



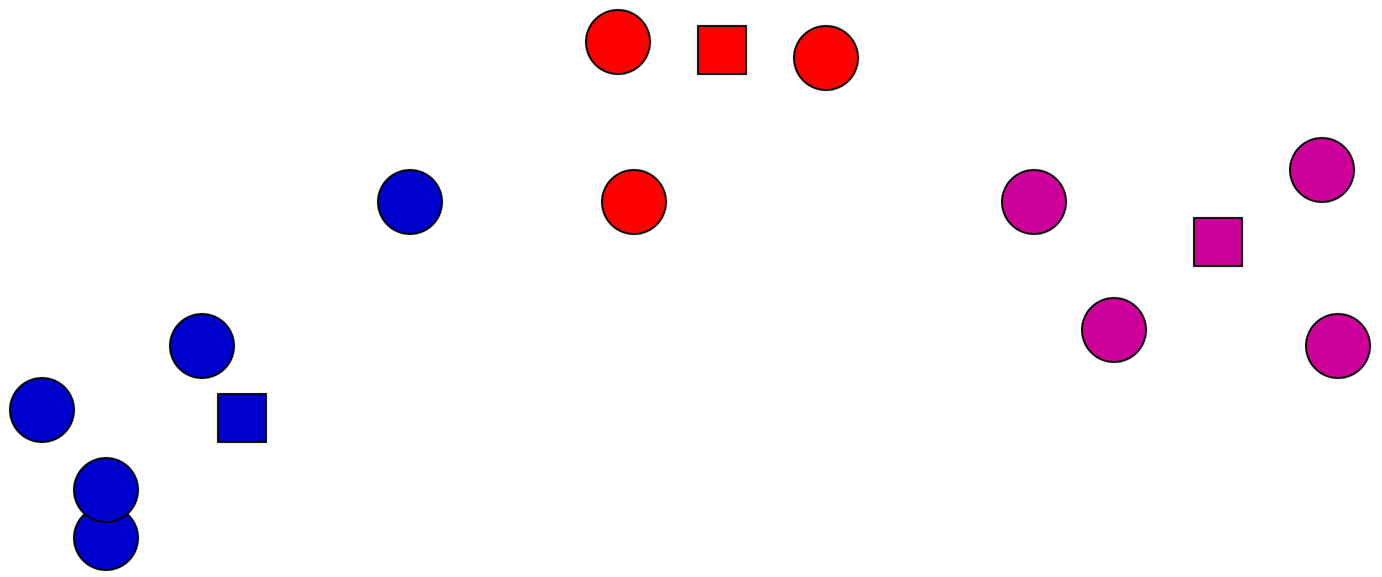
K-means: assign points to nearest center



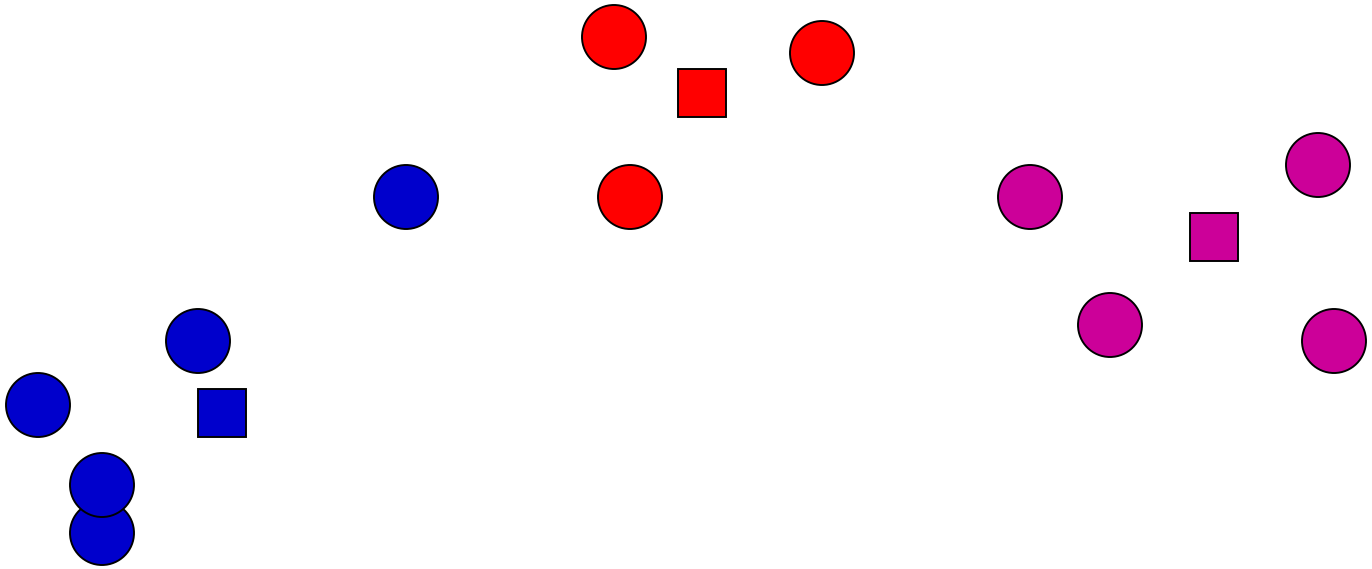
K-means: readjust centers



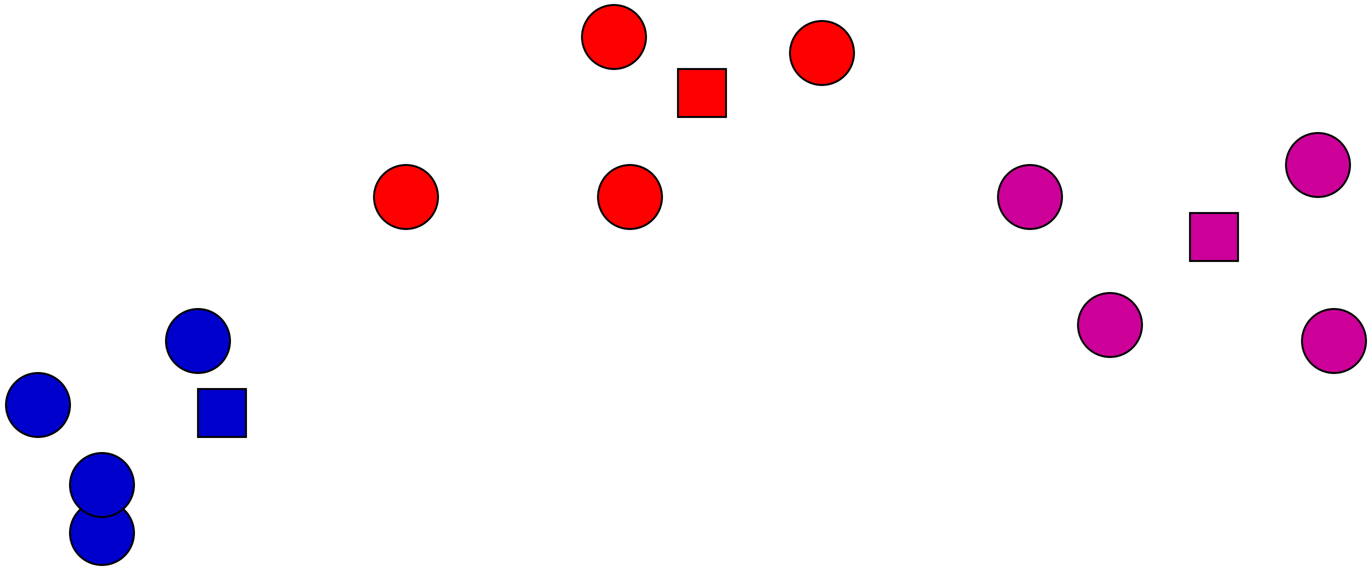
K-means: assign points to nearest center



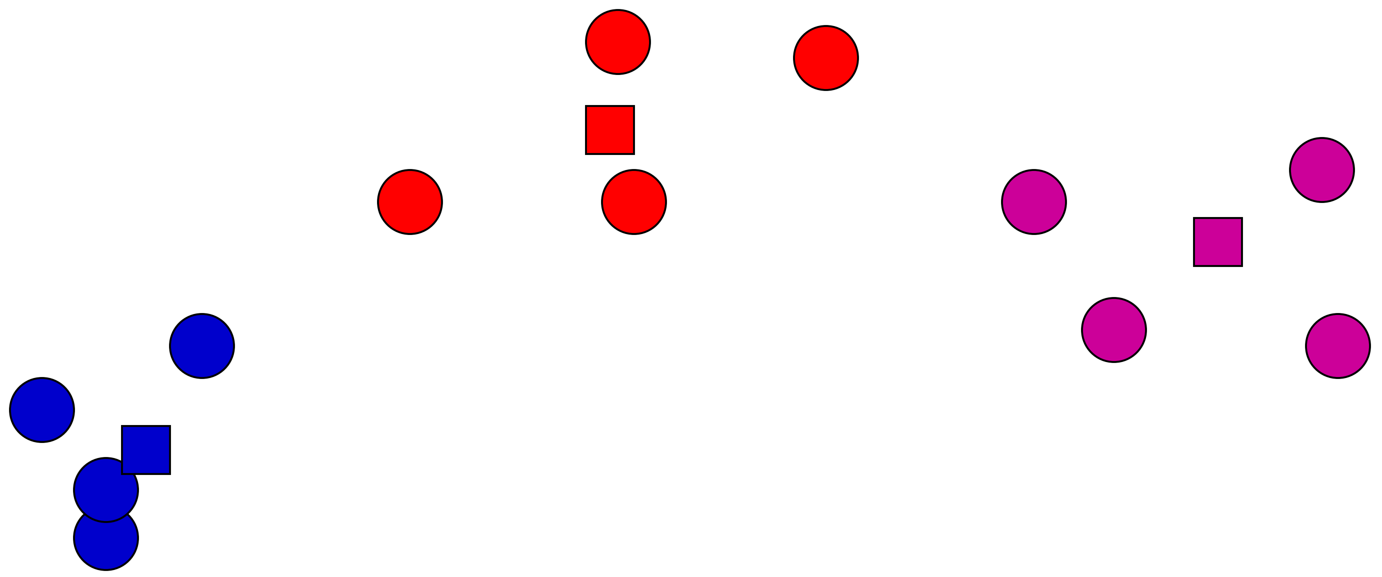
K-means: readjust centers



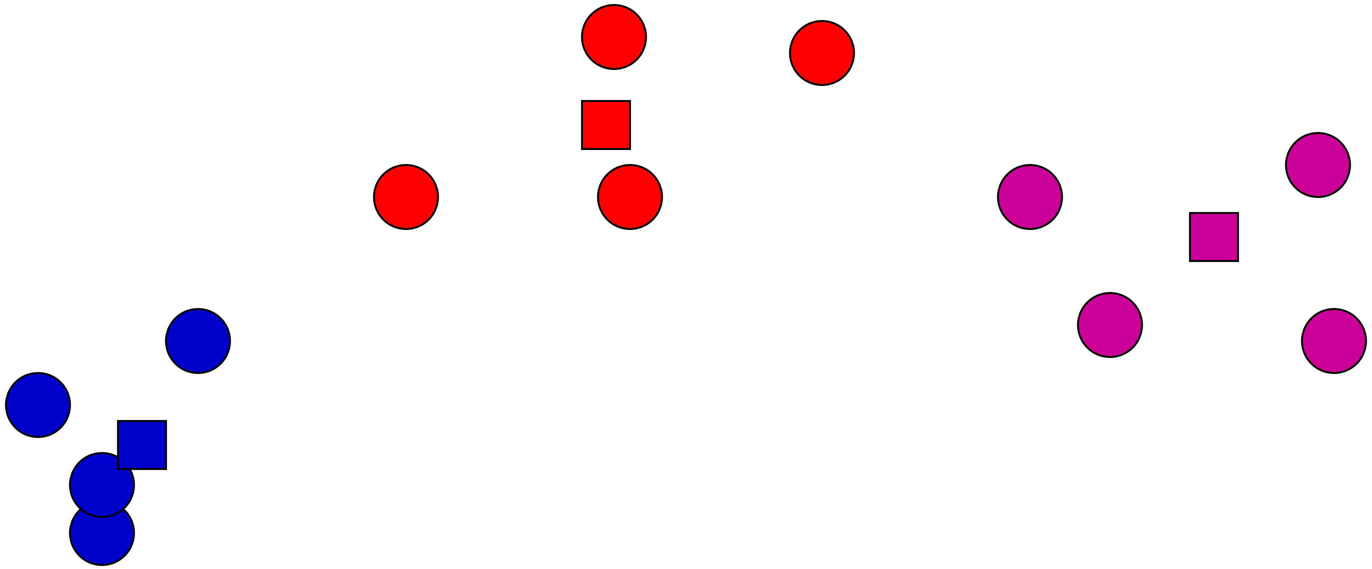
K-means: assign points to nearest center



K-means: readjust centers

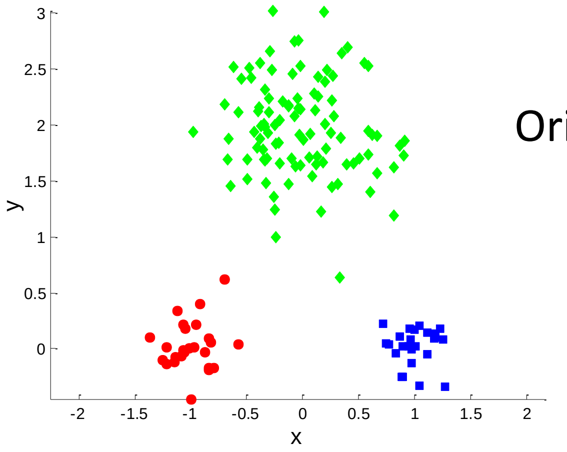


K-means: assign points to nearest center

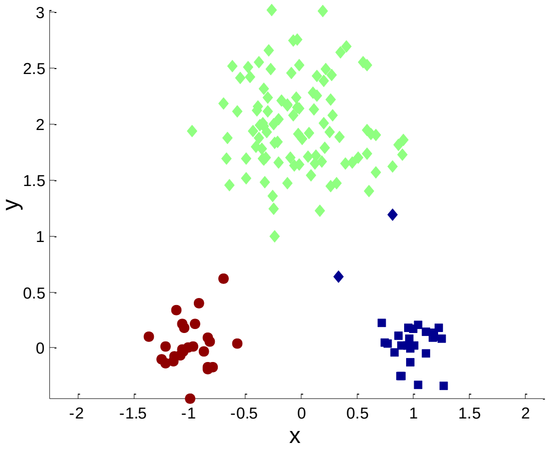


No changes: Done

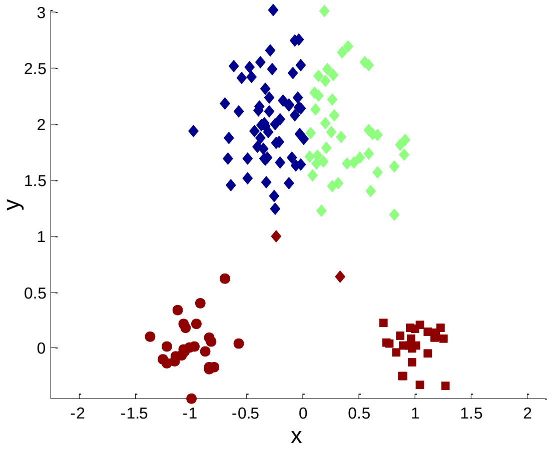
Two different K-means Clusterings



Original Points



Optimal Clustering



Sub-optimal Clustering

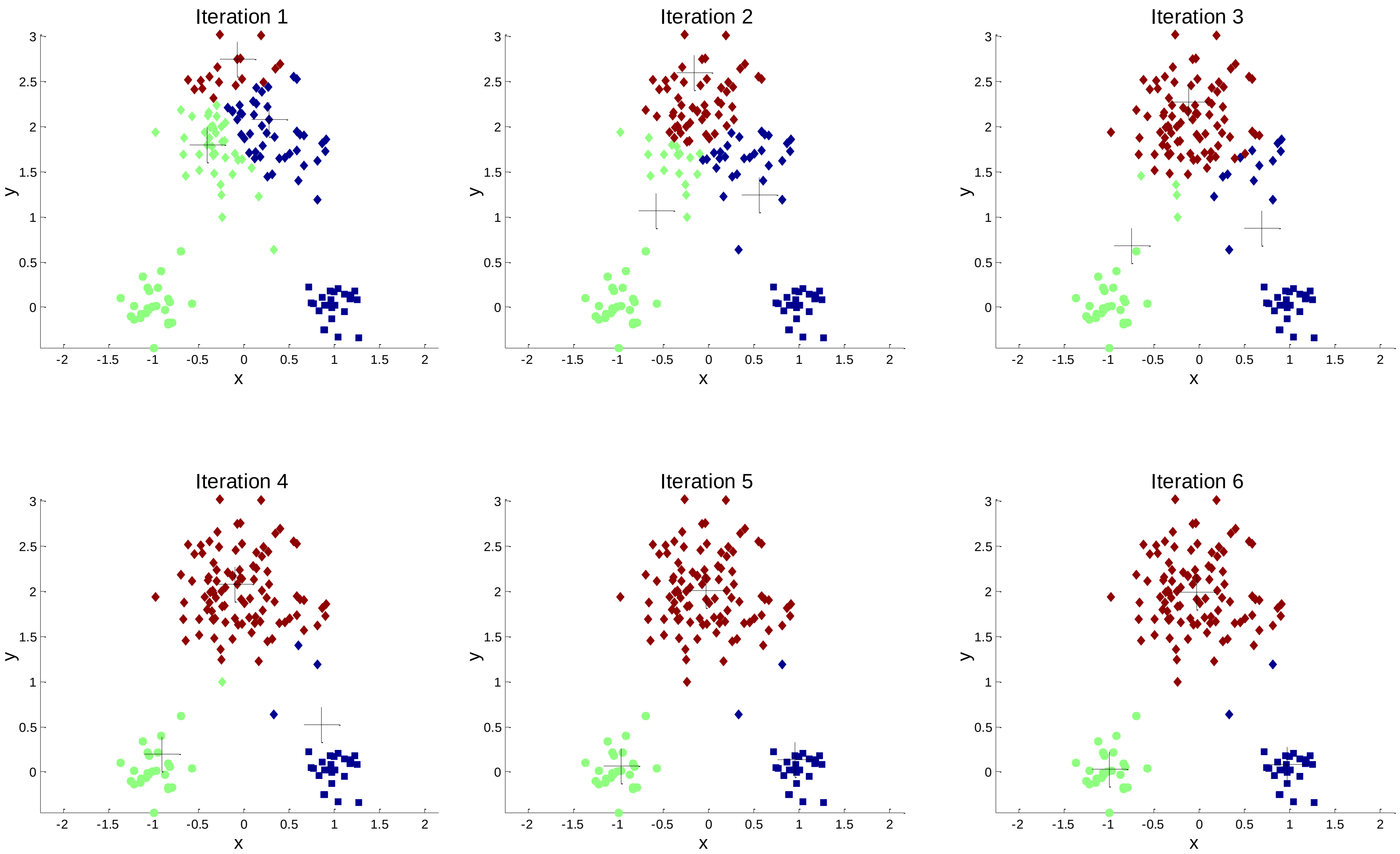
Evaluating K-means Clusters

- 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:

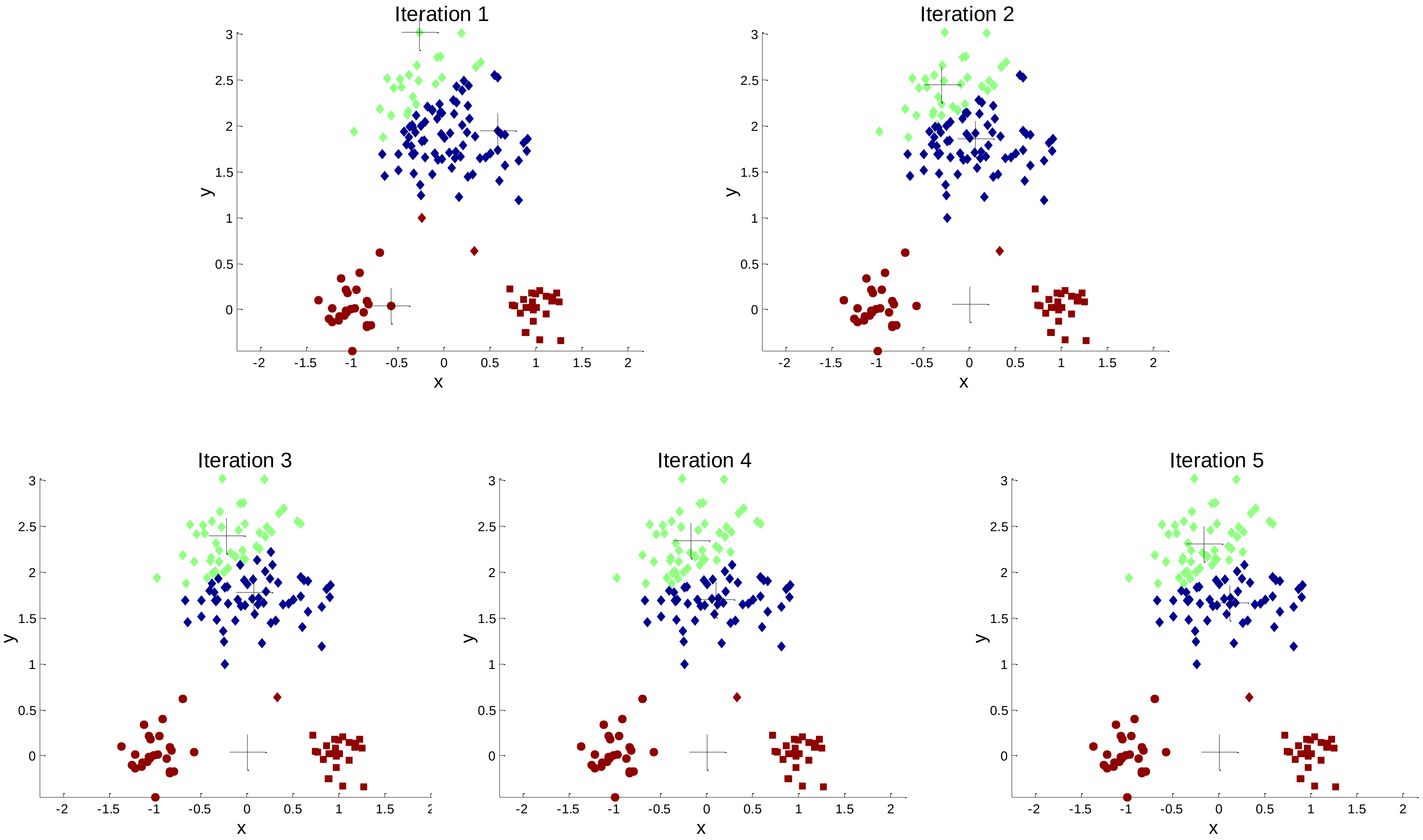
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x is a data point in cluster C_i and m_i is the representative point for cluster C_i
    - m_i corresponds to the center (mean) of the cluster mostly
  - Given many clusterings, we can choose the one with the smallest error

Importance of Choosing Initial Centroids



Importance of Choosing Initial Centroids …



Problems with Selecting Initial Points

- If there are K '**real**' clusters then the chance of selecting one centroid from each cluster is small.

  - Chance is relatively small when K is large
  - Sometimes the initial centroids will readjust themselves in 'right' way, and sometimes they don't

Solutions to Initial Centroids Problem

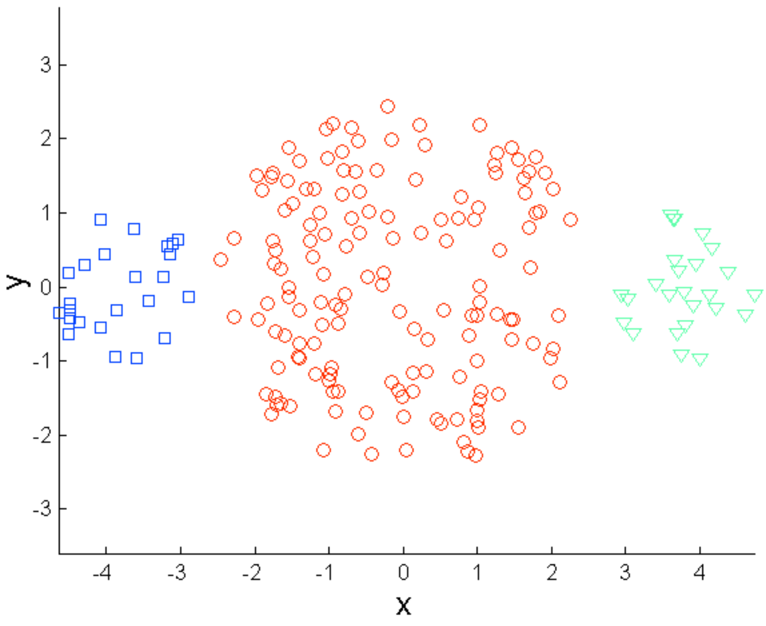
- Multiple runs
  - Helps, but probability is not on your side
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Postprocessing
  - Eliminate small clusters that may represent outliers
  - Split ‘loose’ clusters, i.e., clusters with relatively high SSE
  - Merge clusters that are ‘close’ and that have relatively low SSE

Limitations of K-means

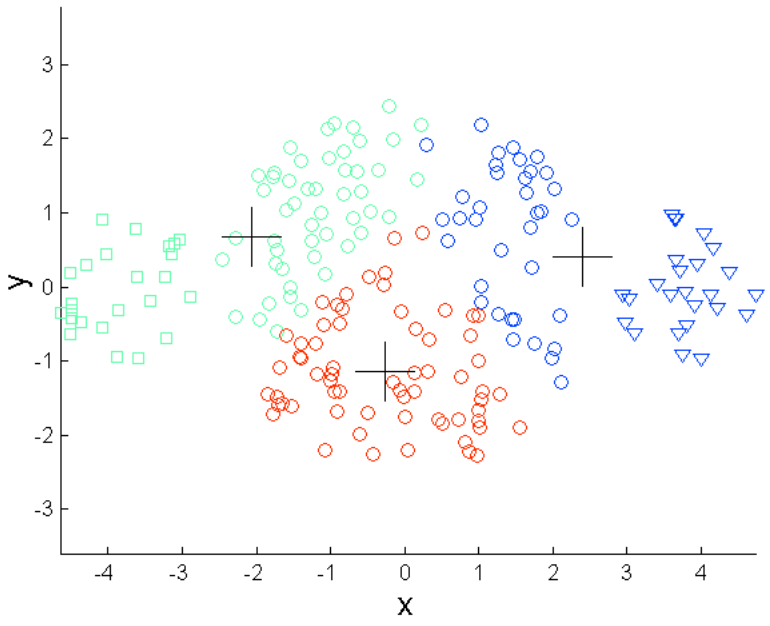
- K-means has problems when clusters are of differing
  - Sizes
  - Densities

- K-means has problems when the data contains outliers (not belonging to any cluster).

Limitations of K-means: Differing Sizes

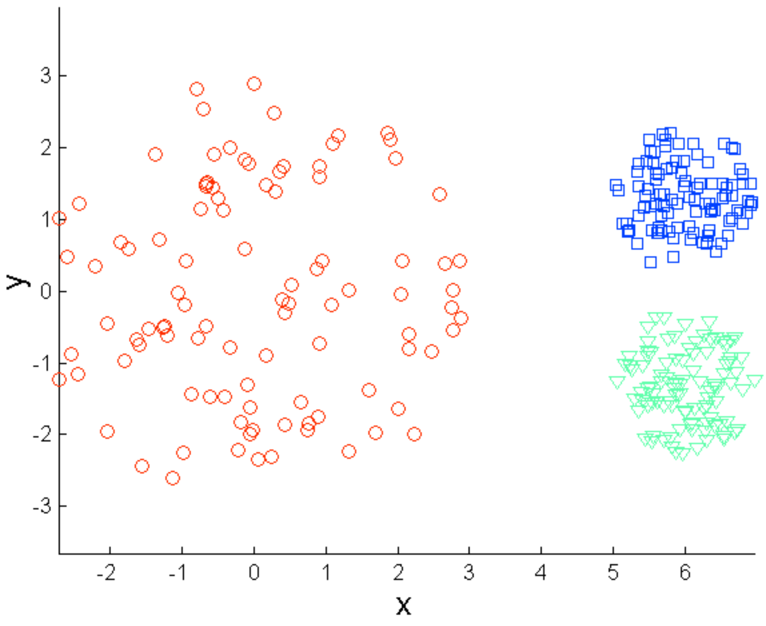


Original Points

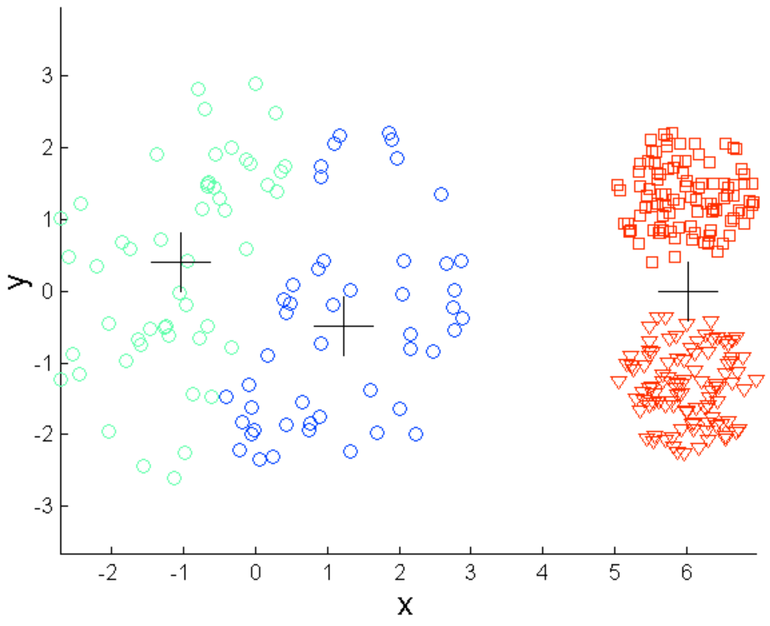


K-means (3 Clusters)

Limitations of K-means: Differing Density

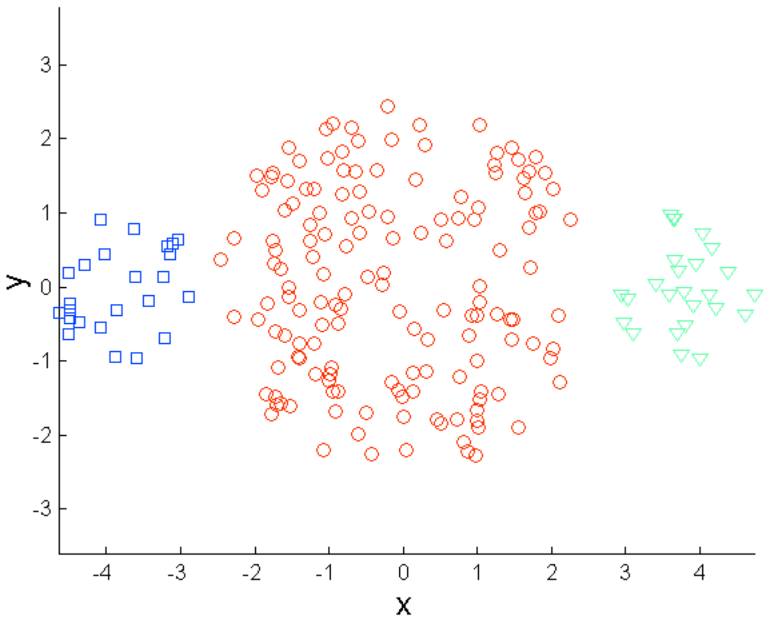


Original Points

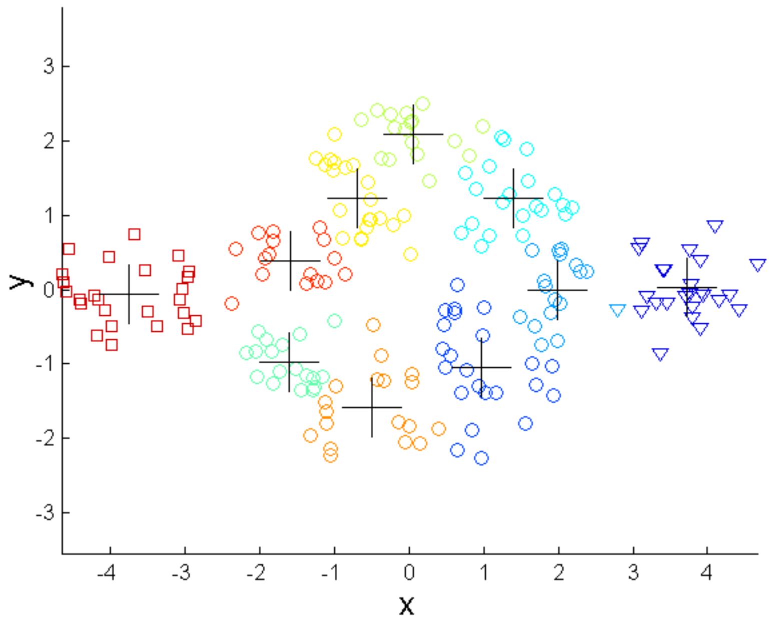


K-means (3 Clusters)

Overcoming K-means Limitations



Original Points

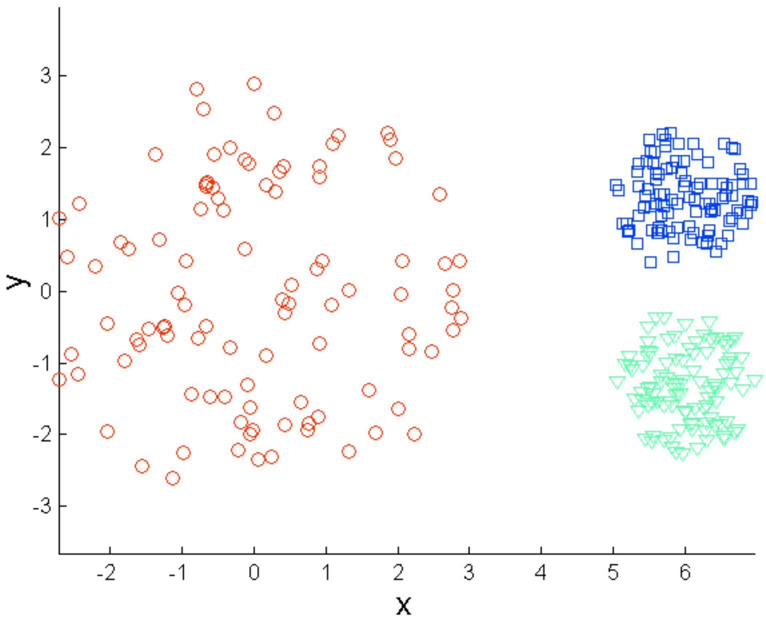


K-means Clusters

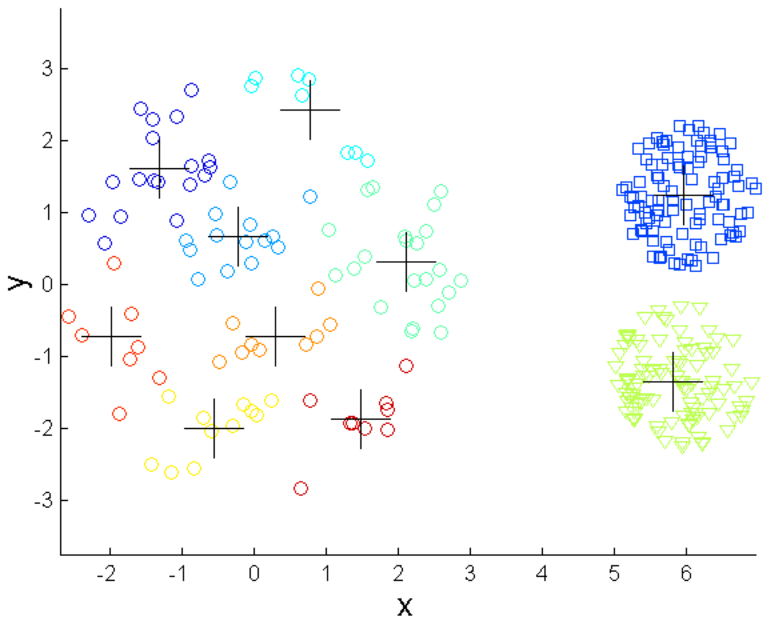
One solution is to use many clusters.

Find parts of clusters, but need to put together.

Overcoming K-means Limitations



Original Points



K-means Clusters

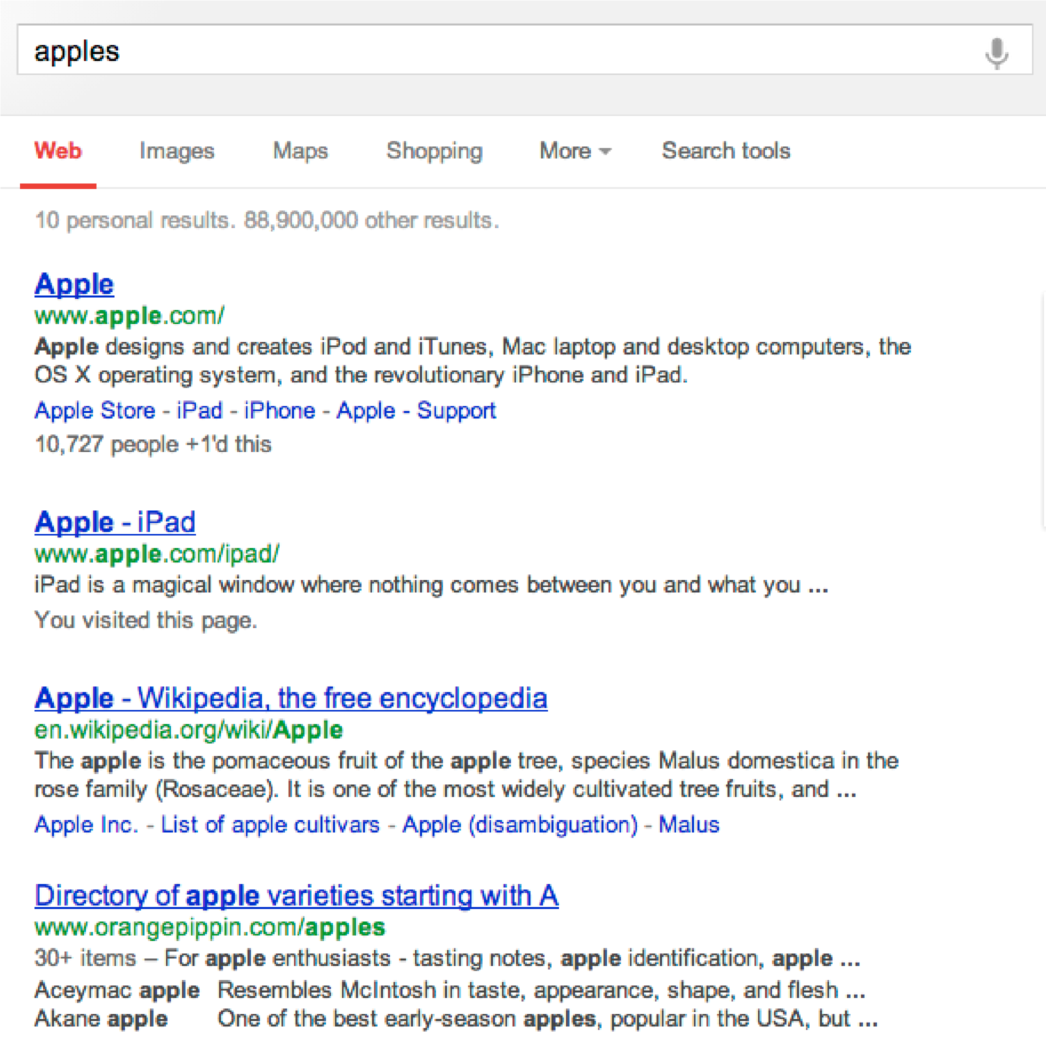
More Clustering Applications: Face Clustering

…

Face photos

Clustering

Search result clustering



Pixel Clustering

Image pixels are represented by 3D vectors of R,G,B values. The vectors are grouped to K = 10, 3, 2 clusters, and represented by the mean values of the respective clusters.

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad \mathbf{x} \in \mathrm{R}^3$$

Original image

$K = 10$

$K = 3$

$K = 2$

Q&A