資料科學導論

K-means Data Clustering

Cheng-Te Li (李政德)

Dept. of CSIE & Statistics

National Cheng Kung University

chengte@ncku.edu.tw

What is a Cluster (群)?

- "A group of the same or similar elements gathered or occurring closely together"

Galaxy clusters

Birdhouse clusters

Cluster munition

Cluster computing

Cluster lights

Hongkeng Tulou cluster

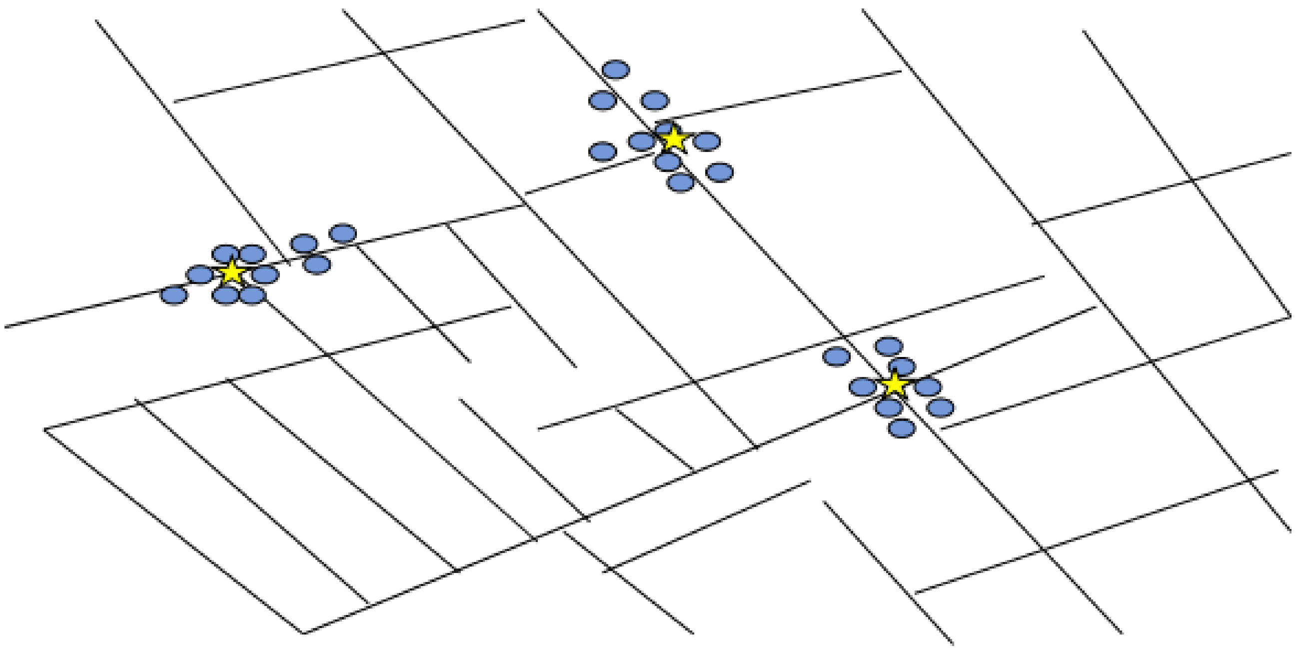
Clustering

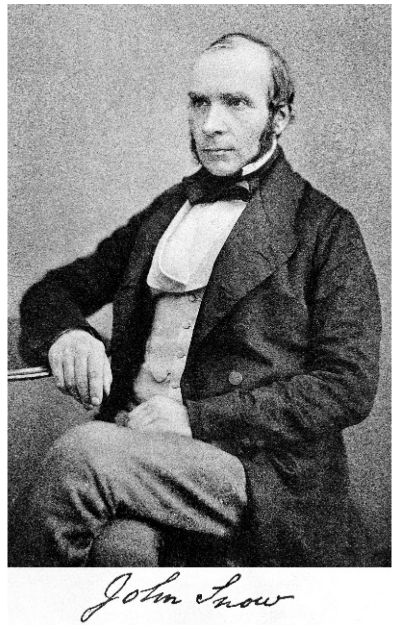
Unsupervised Learning

Given a collection of (unlabeled) objects, find meaningful groups

First Application of Clustering

- John Snow, a London physician plotted the location of cholera (霍亂) deaths on map during an outbreak in 1850s
- The locations indicated that cases were clustered around certain intersections where there were polluted wells -- thus exposing both the problem and the solution





Clustering for Image Compression

701,554 bytes

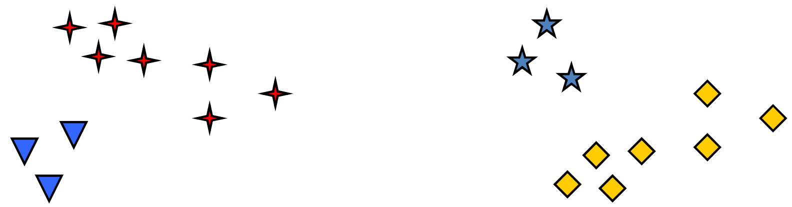
127,292 bytes

Clusters in 2D

How many clusters?

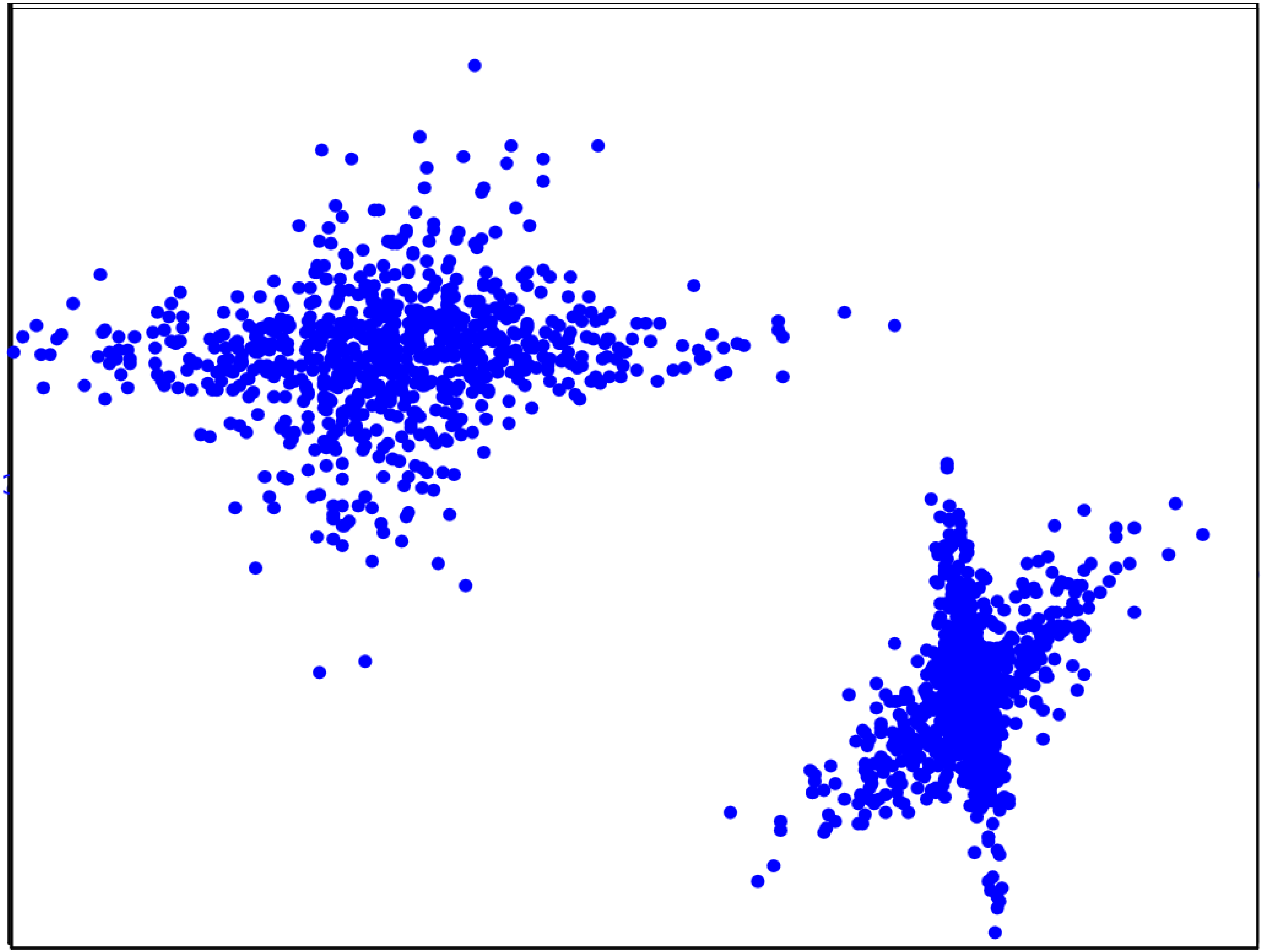
Six Clusters

Two Clusters



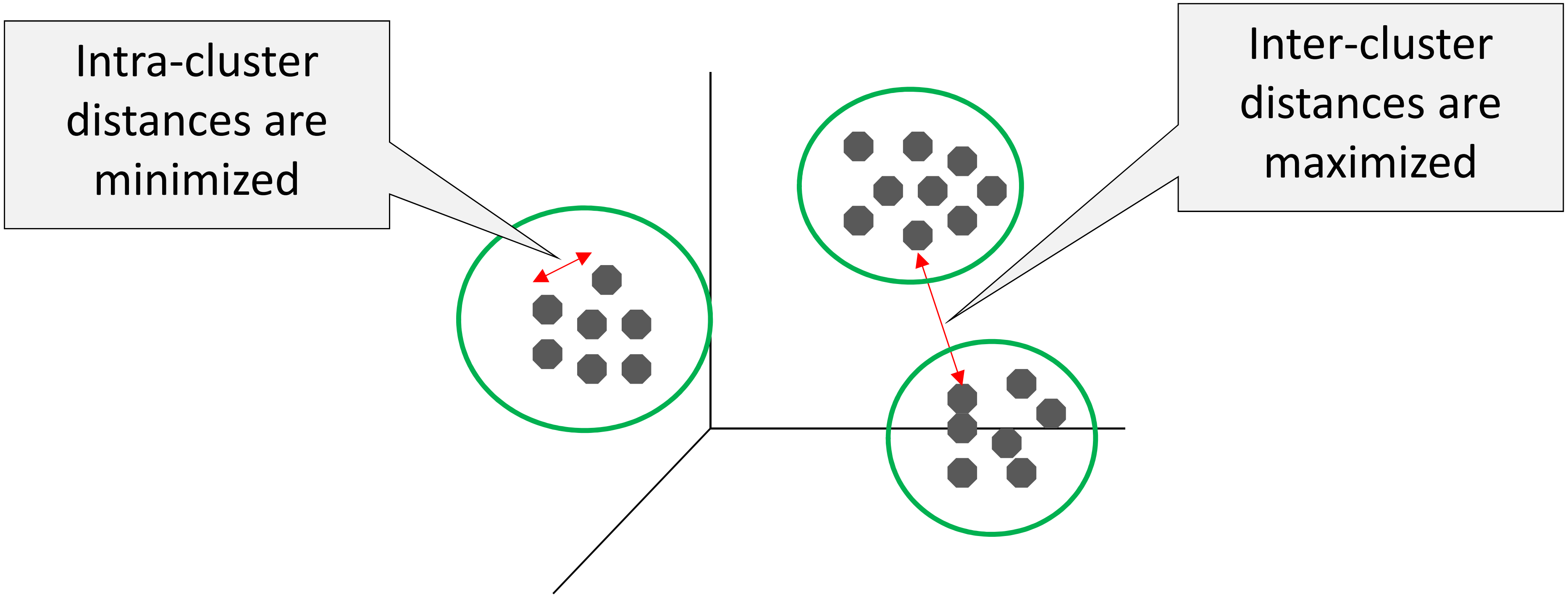
Four Clusters

Clusters in 2D



What is **Clustering?**

- A grouping of objects such that the objects in a group (**cluster**) are
  - **Similar** (or related) to one another in a group and
  - **Different** from (or unrelated to) the objects in other groups



Unsupervised Learning: Clustering

Raw data

Features/Variables

$f_1, f_2, f_3, \ldots, f_n$

Extract Features

$f_1, f_2, f_3, \ldots, f_n$

$f_1, f_2, f_3, \ldots, f_n$

Group into classes/clusters

Clusters

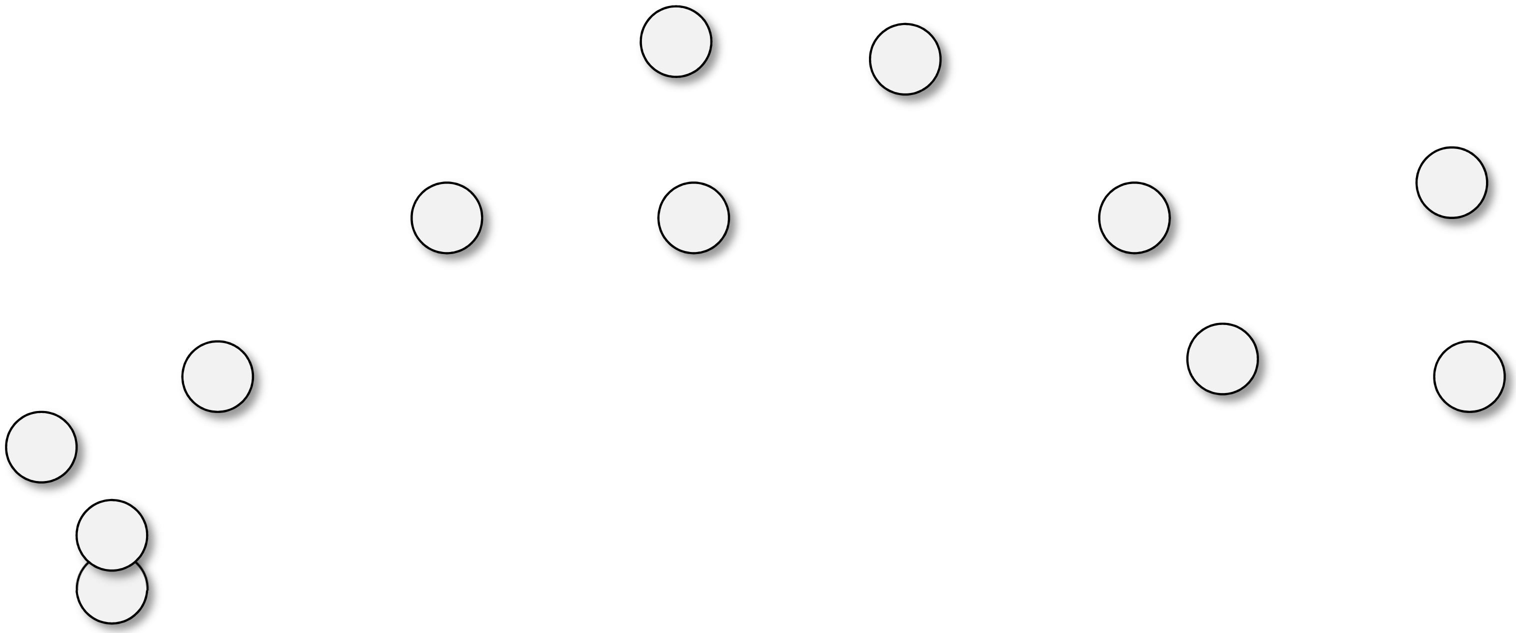
$f_1, f_2, f_3, \ldots, f_n$

$f_1, f_2, f_3, \ldots, f_n$

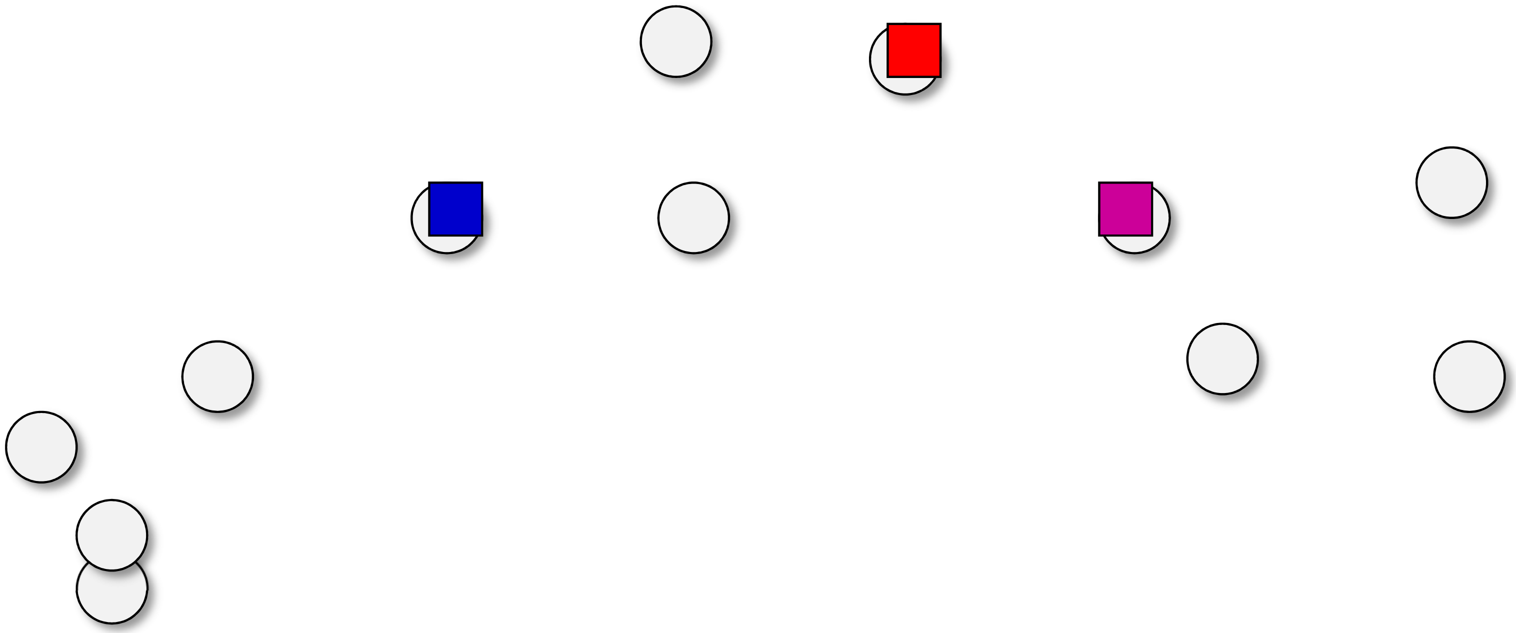
No "supervision",

we're only given data and want to find natural groupings

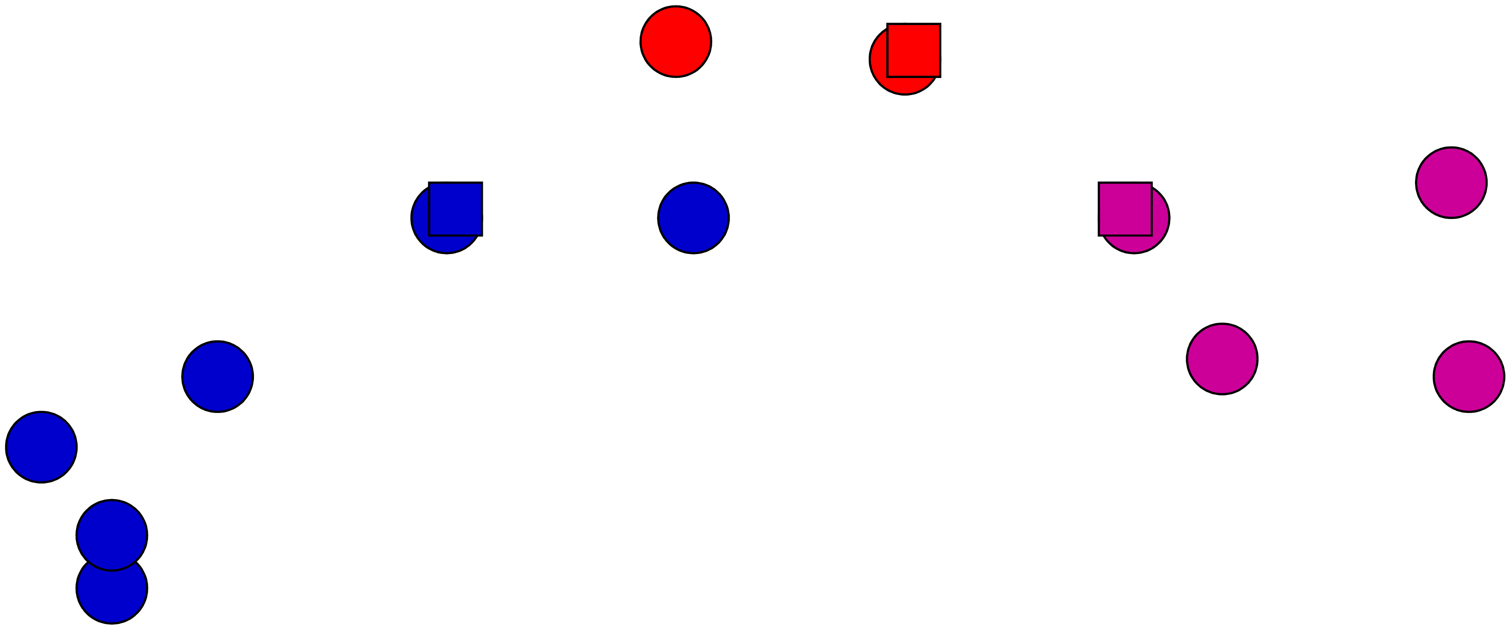
K-means: An Example



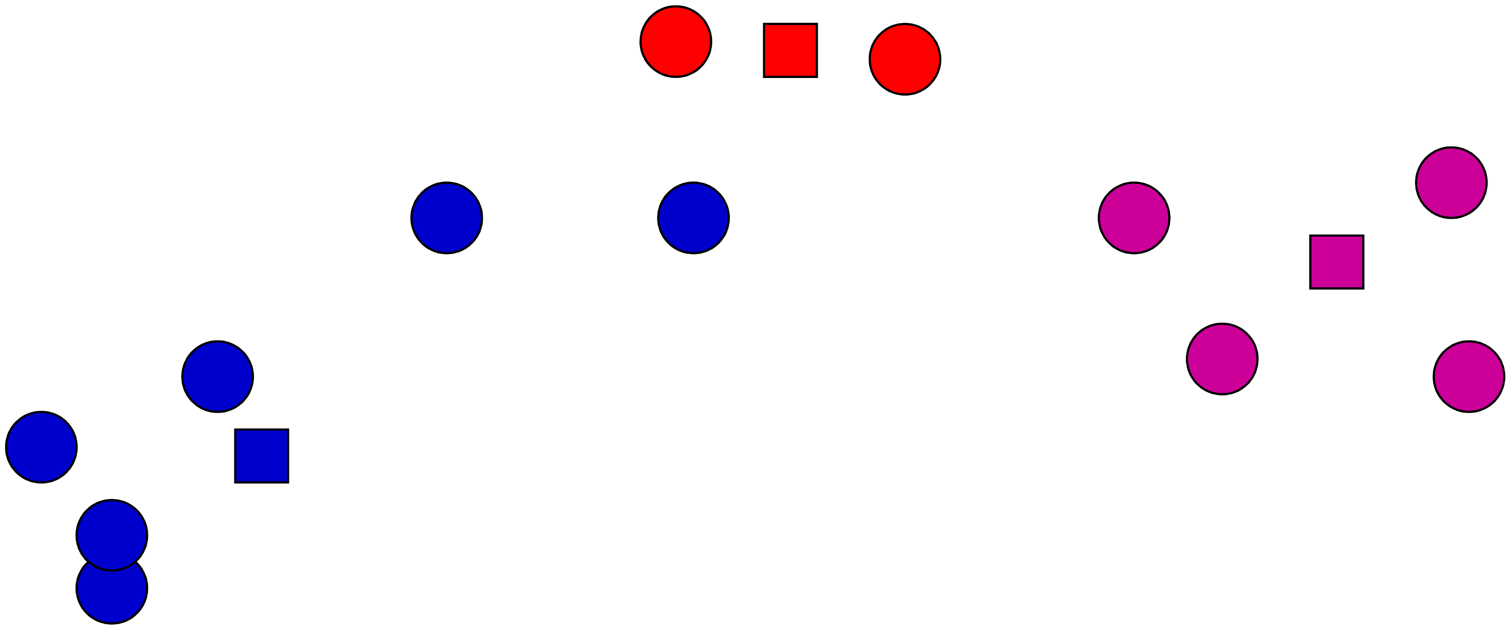
K-means: Initialize Centers **Randomly**



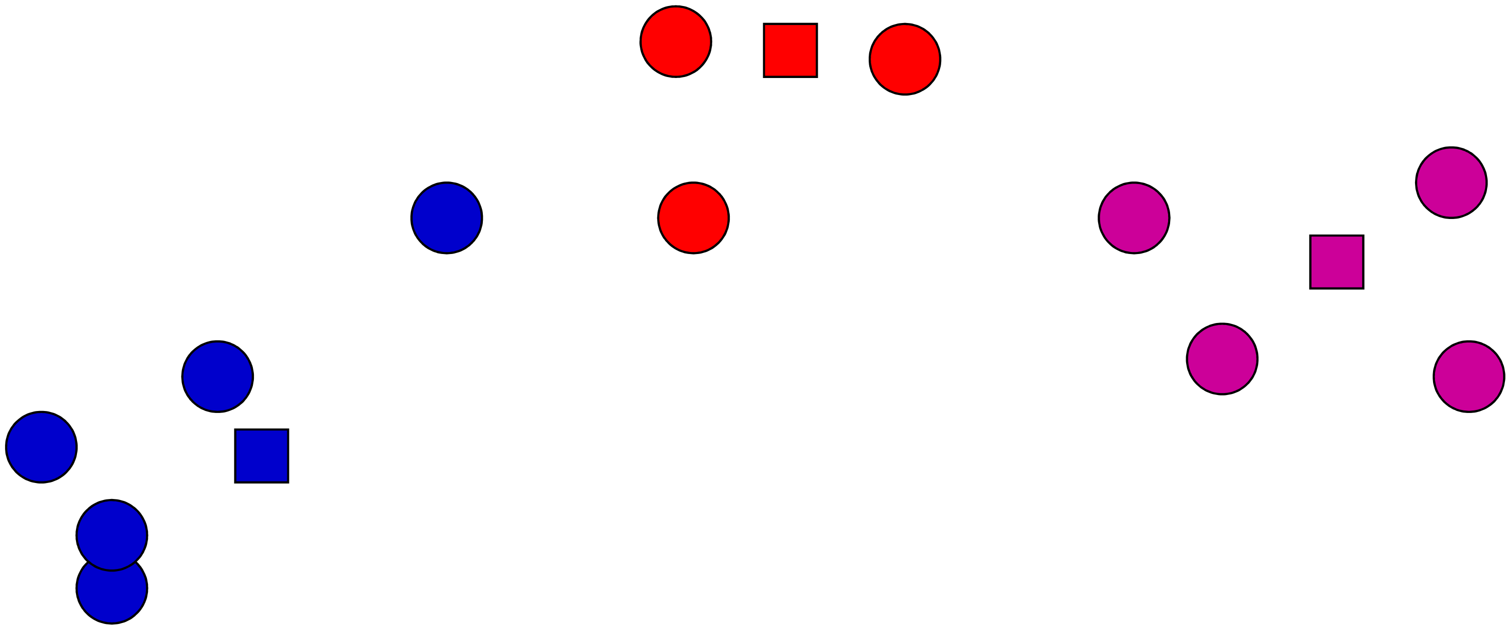
K-means: Assign Points to the Nearest Center



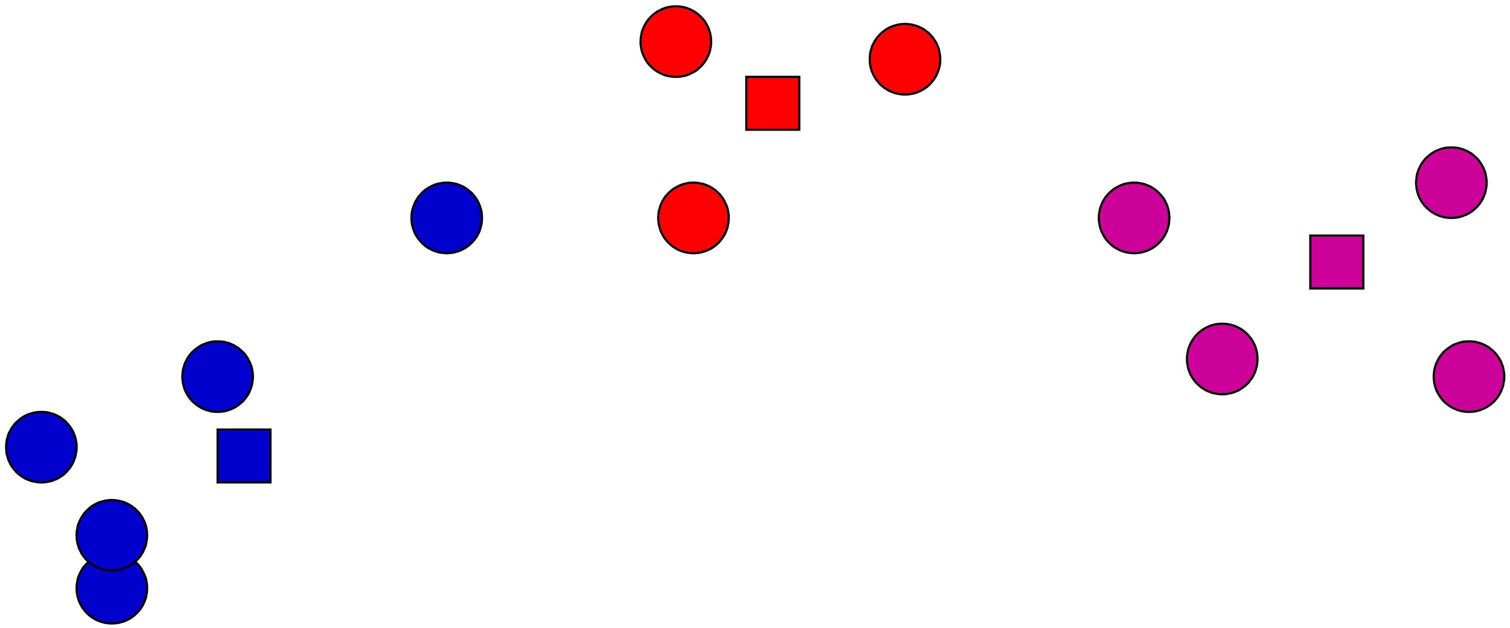
K-means: Readjust Centers



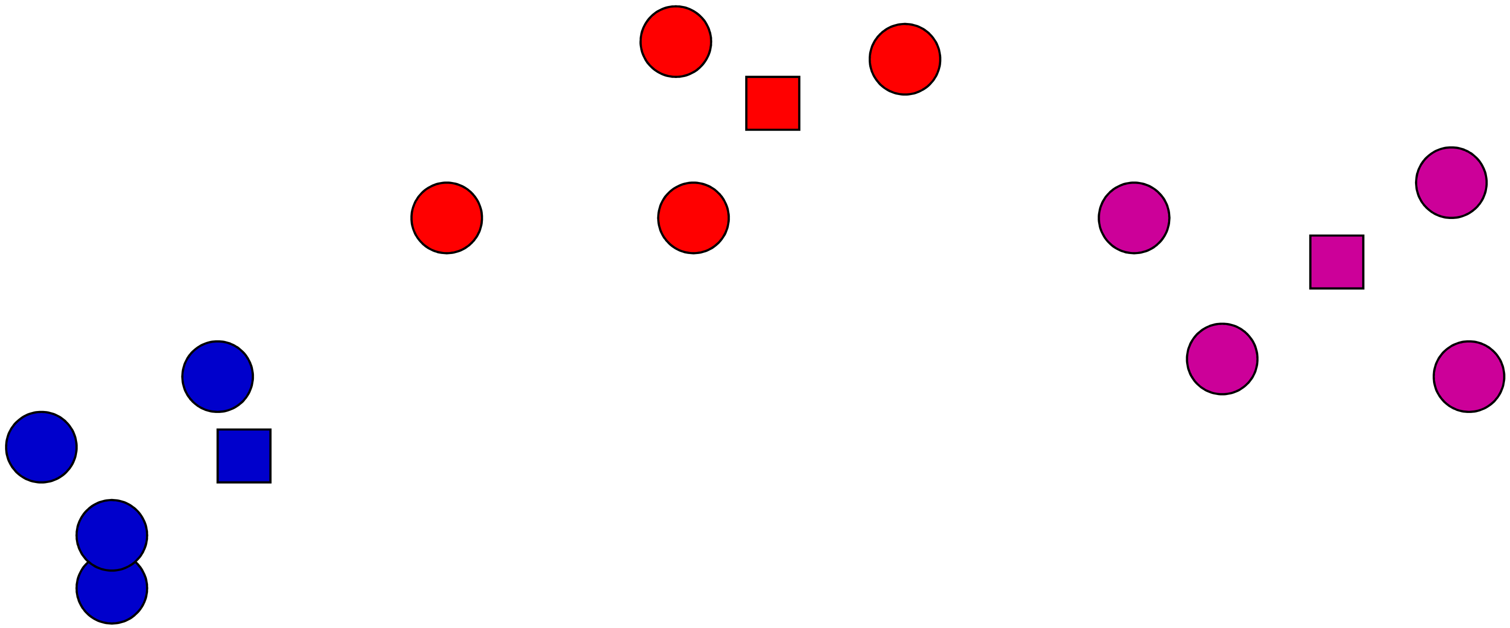
K-means: Assign Points to the Nearest Center



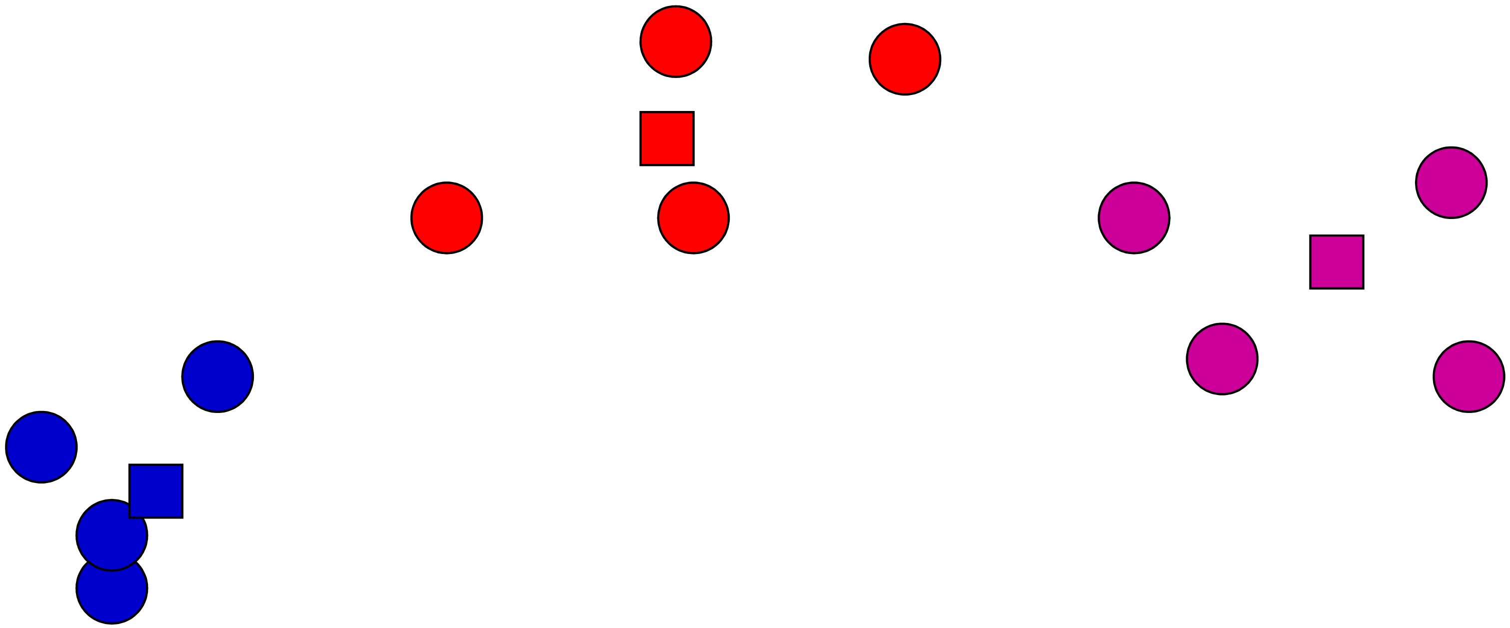
K-means: Readjust Centers



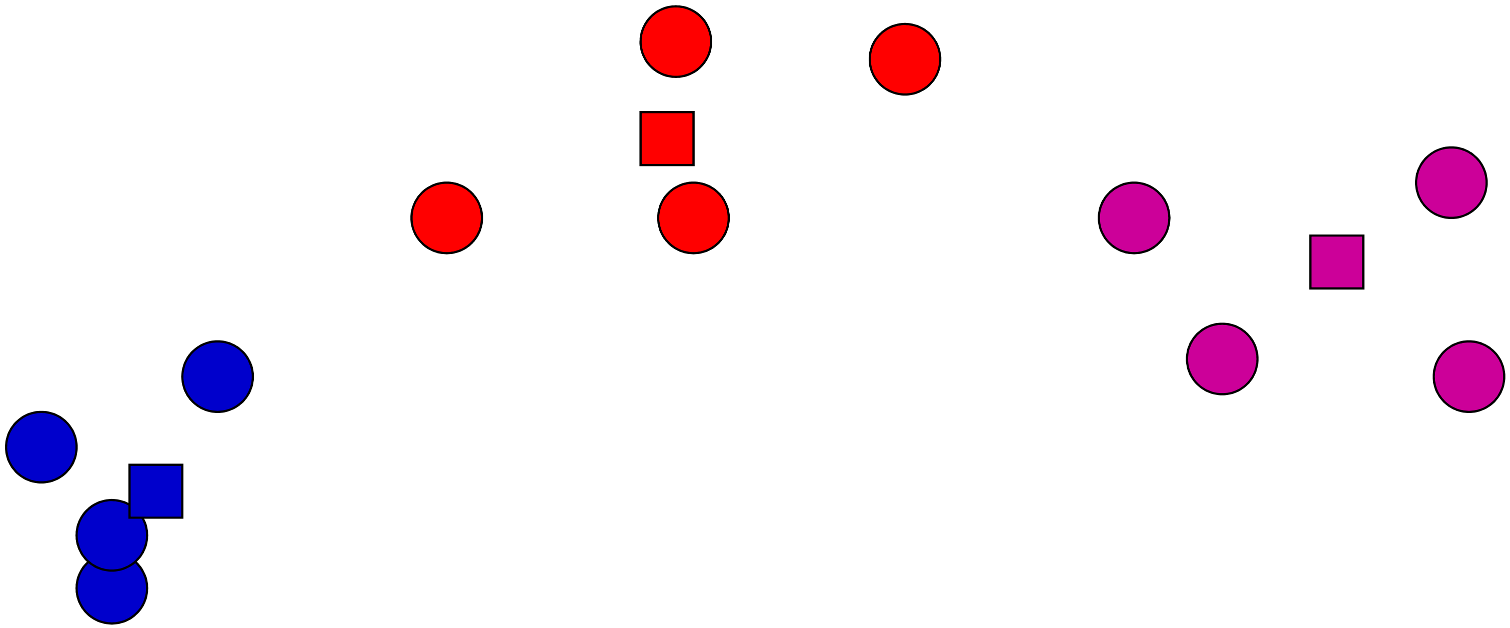
K-means: Assign Points to the Nearest Center



K-means: Readjust Centers



K-means: Assign Points to the Nearest Center



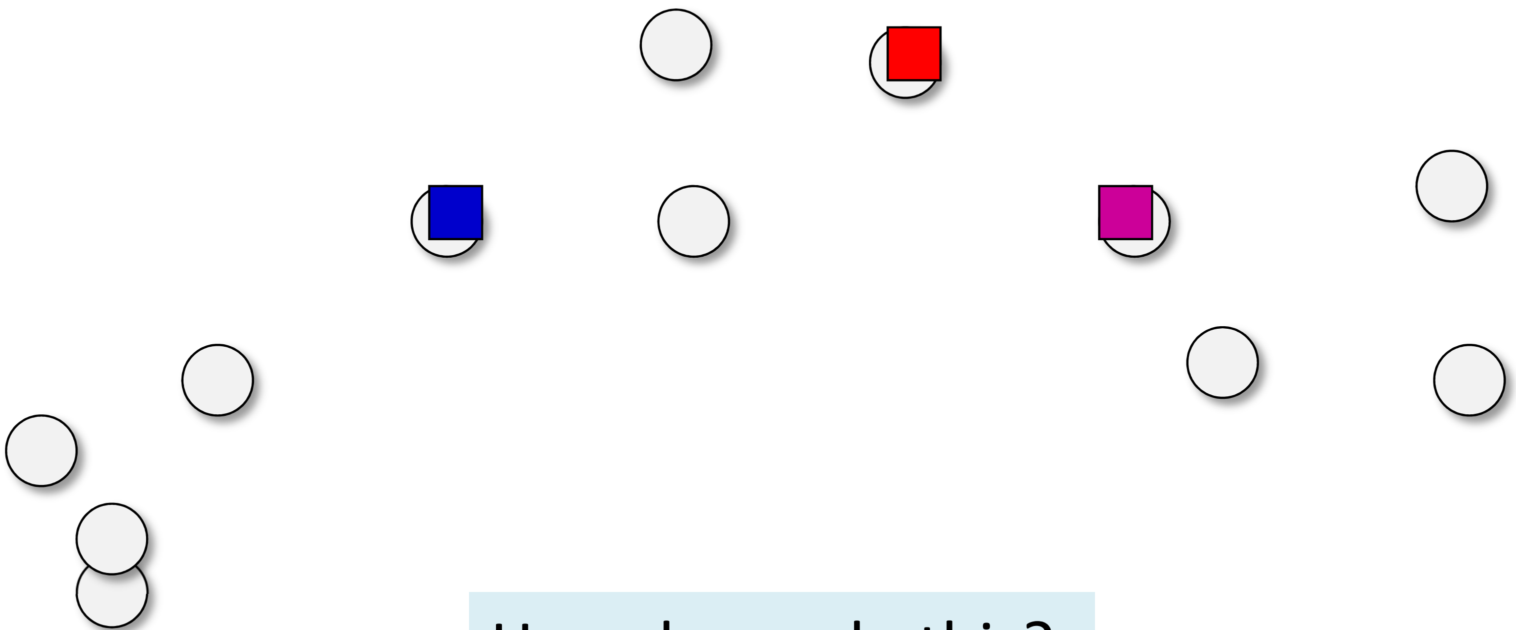
No changes: Done

K-means

Iterate:

Centroid

- Assign/cluster each point to the $\boxed{\textbf{closest}\ \text{center}}$
- Recalculate centers as the mean of the points in a cluster



How do we do this?

K-means

Iterate:

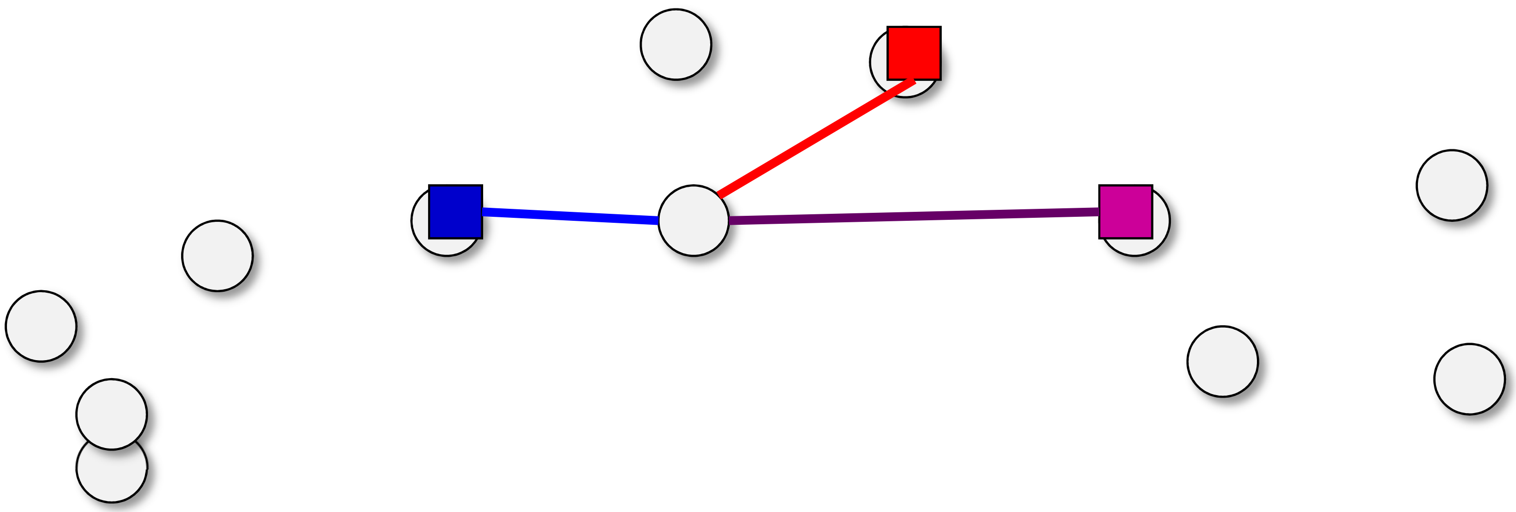
- **Assign/cluster each point to the closest center** (Centroid)

  Iterate over each point:

  (1) Get **distance** to each cluster center

  (2) Assign to the closest center

- Recalculate centers as the mean of the points in a cluster



What distance measure should we use?

Distance Measures

- Euclidean

  Good for spatial data

$$d(x,y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$$

- Cosine Distance

  Correlated with the angle between two vectors

  between 0 and 1

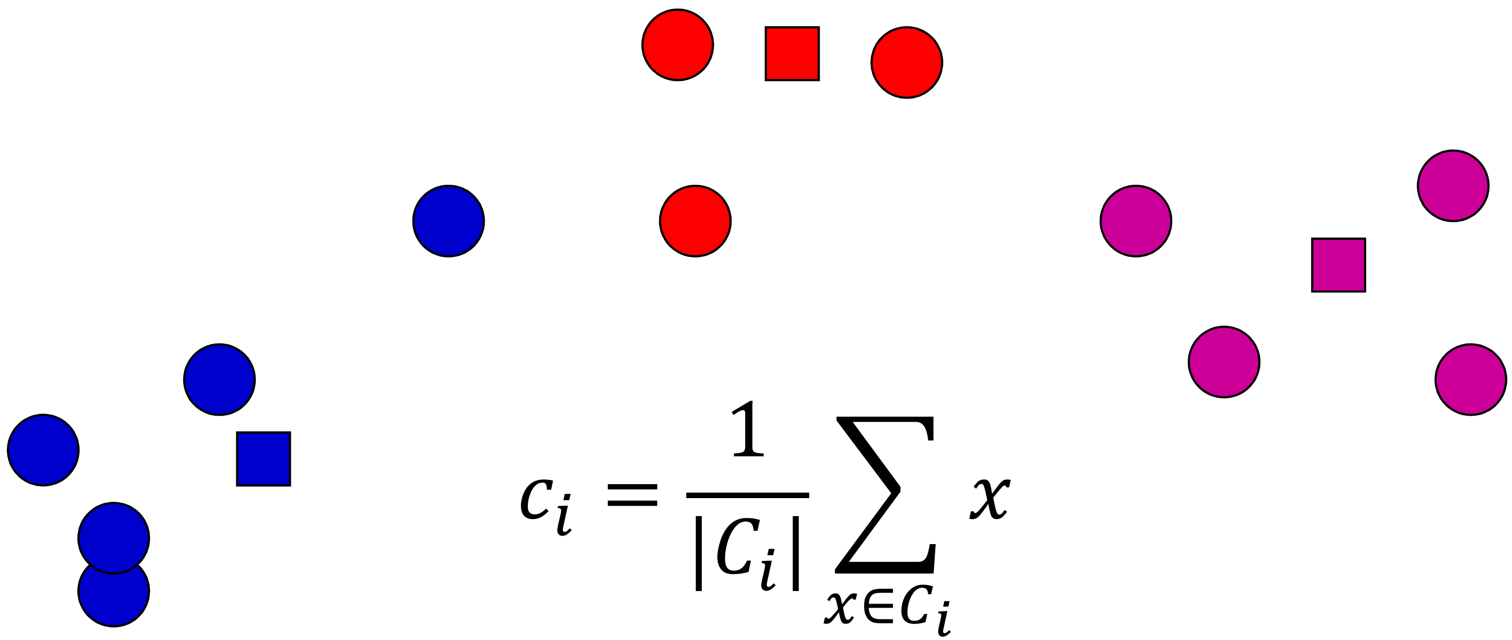
$$d(x,y) = 1 - sim(\boldsymbol{x},\boldsymbol{y})$$

$$sim(x,y) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} = \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} {x_i}^2}\sqrt{\sum_{i=1}^{d} {y_i}^2}}$$

K-means

Iterate:

- Assign/cluster each example to $\boxed{\text{the } \textbf{closest} \text{ center}}$ **Centroid**
- Recalculate centers as the mean of the points in a cluster



$$c_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

K-means Clustering

- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, $\boldsymbol{K}$, must be specified
- The objective is to find:
  - K centroids and
  - the assignment of points to clusters/centroids

  so as to **minimize the sum of distances of the points to their respective centroid**

K-means Clustering

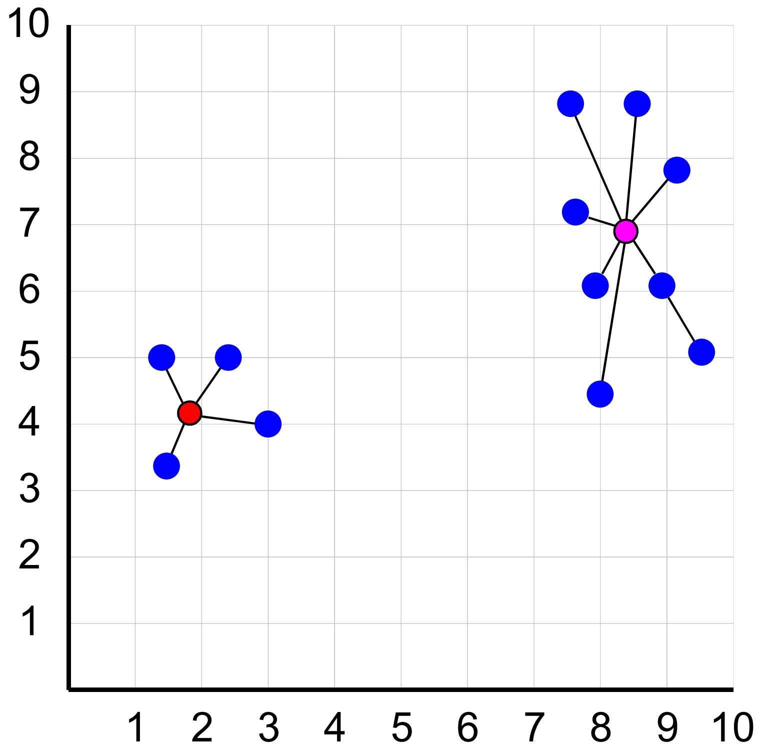
- **Definition:** Given a set X of n objects and an integer K, group the points into K clusters $C = \{C_1, C_2, \dots, C_K\}$ such that **Sum of Squared Error (SSE)**

$$Cost(C) = Error(C) = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, c_i)^2$$
$$= Loss(C)$$

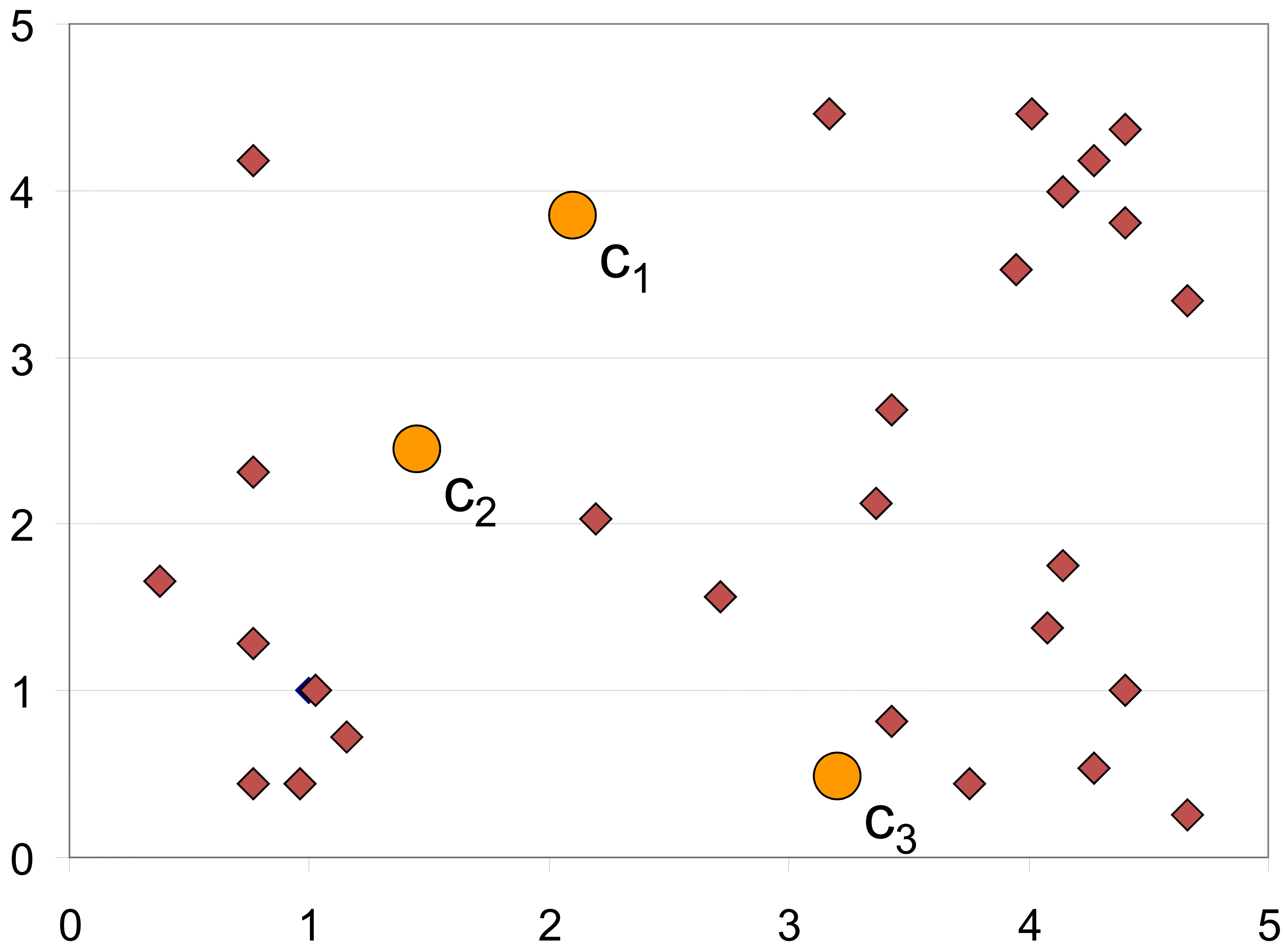
is minimized,
where c_i is the centroid of the points in cluster C_i
  - Given two clusterings,
    we will choose the one with the smallest error
- Note: we need to find both the grouping into clusters and the centroids per cluster

K-means Algorithm

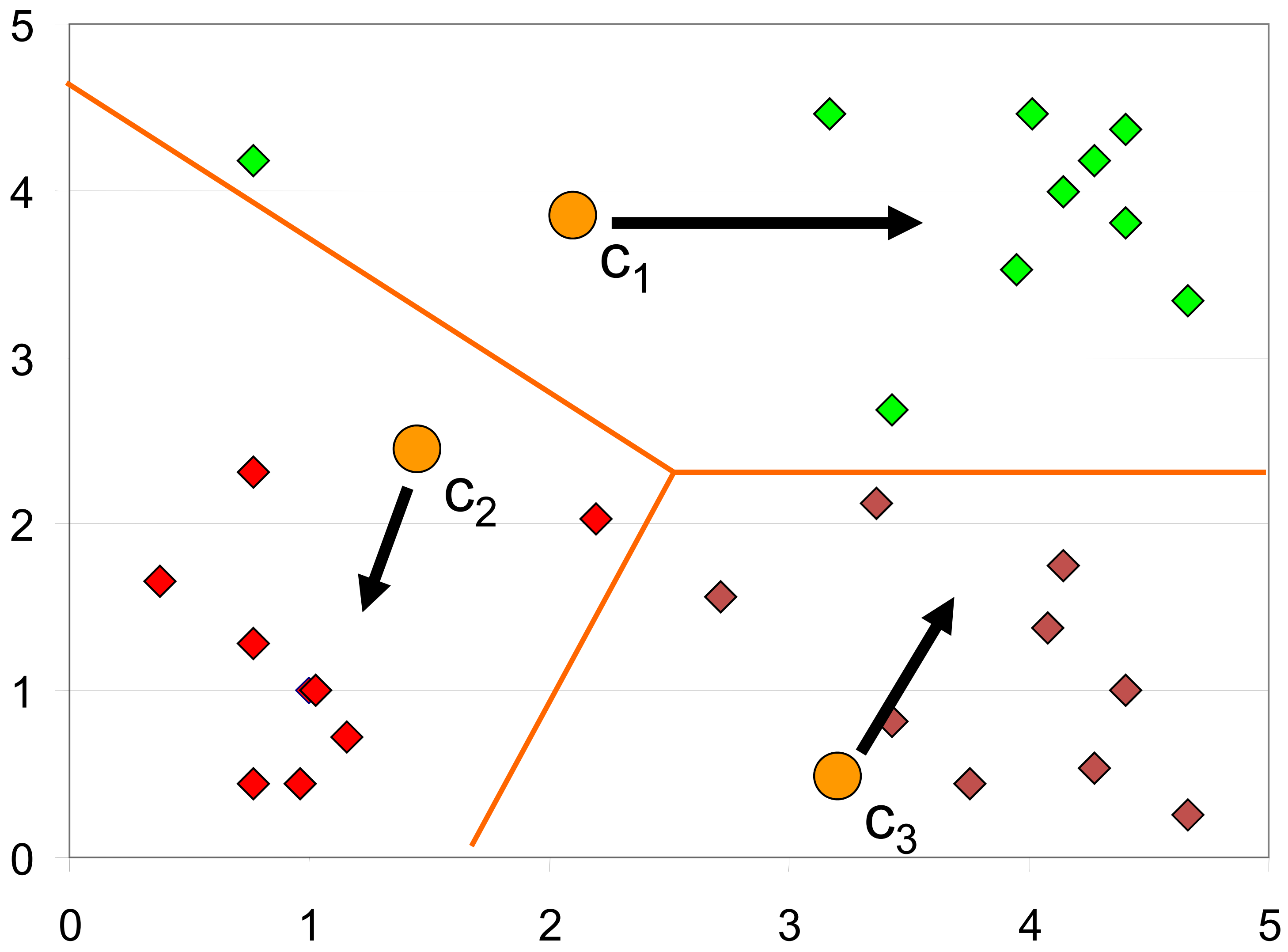
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change



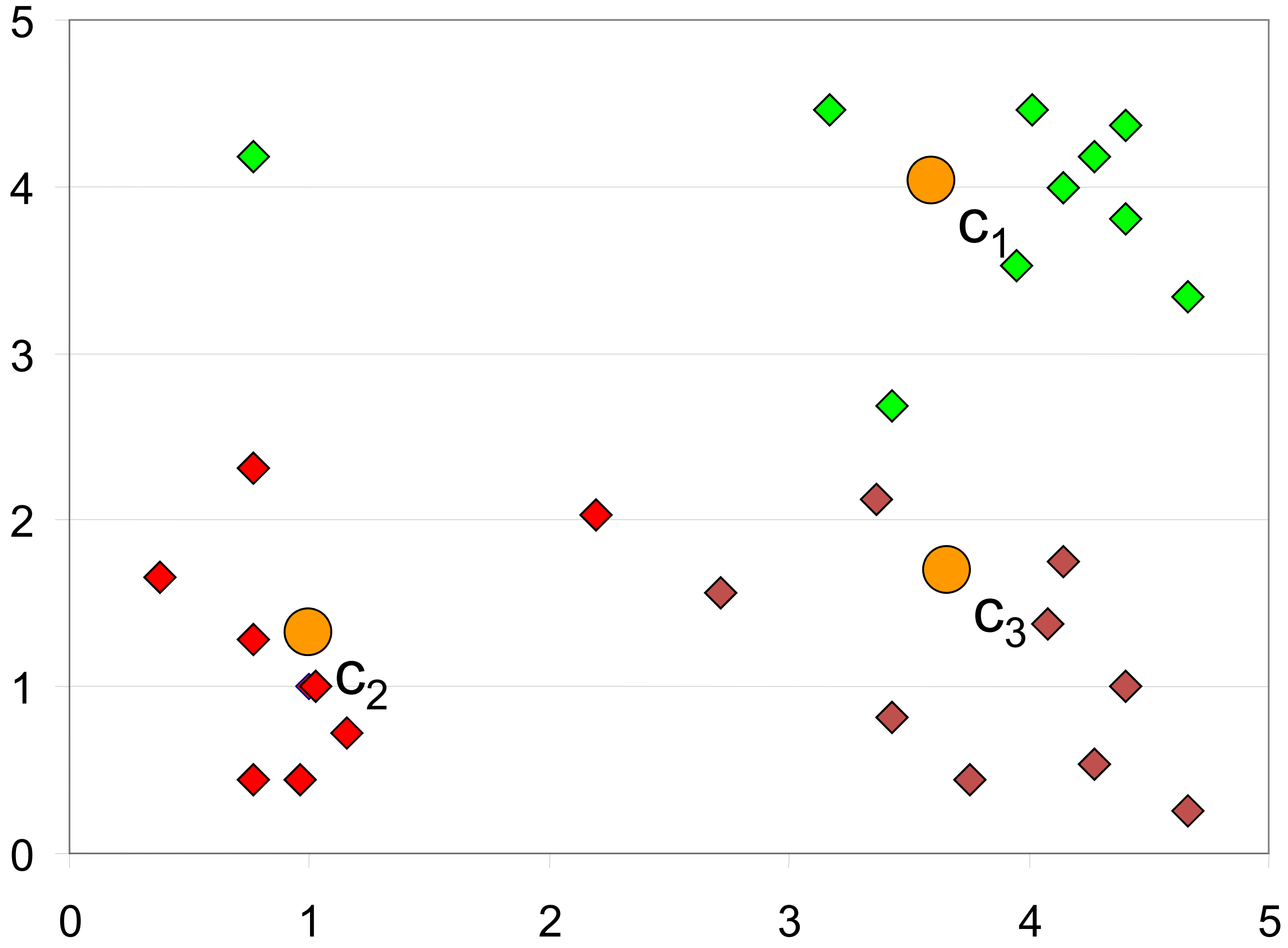
K-means Clustering: Step 1



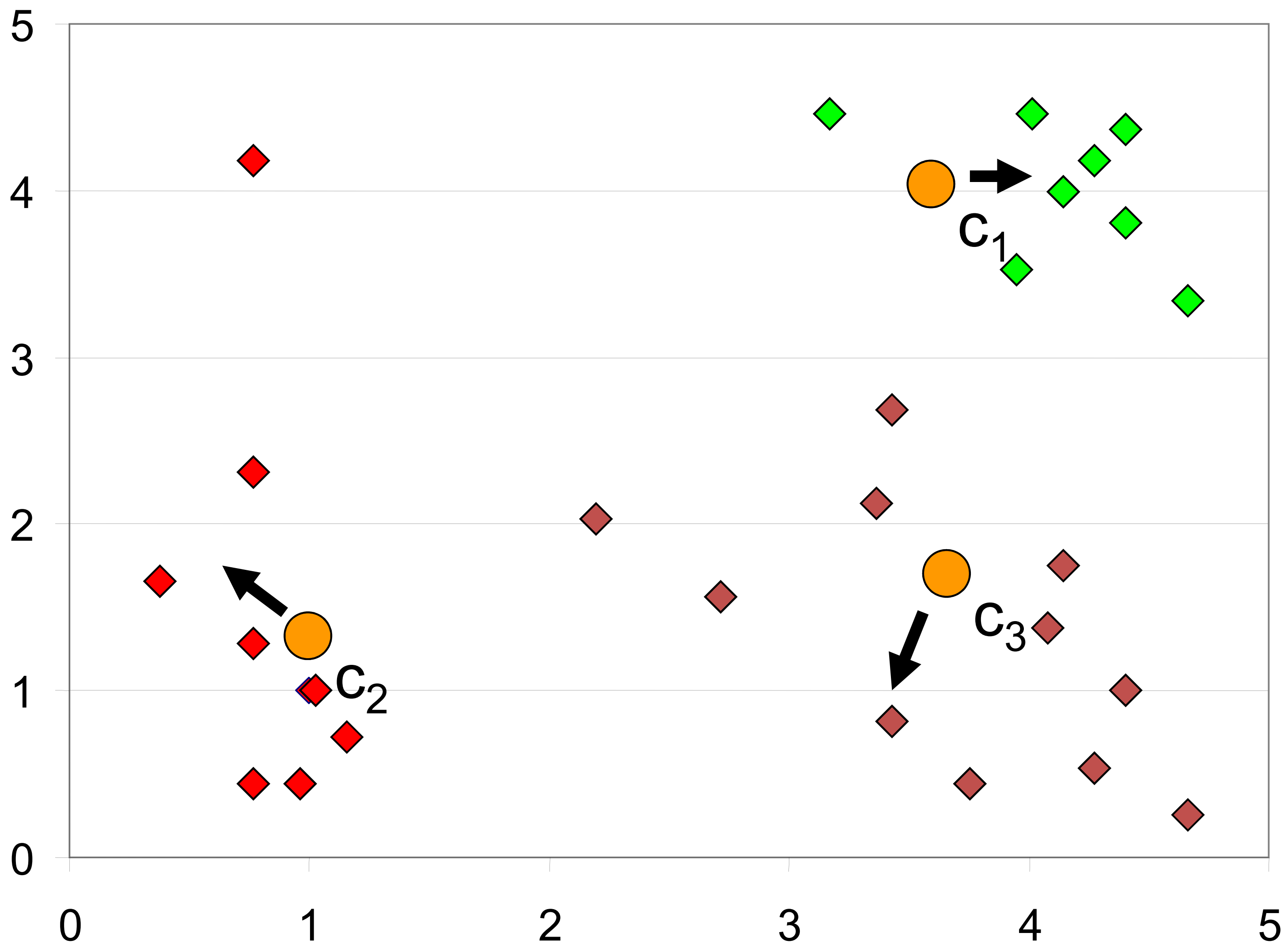
K-means Clustering: Step 2



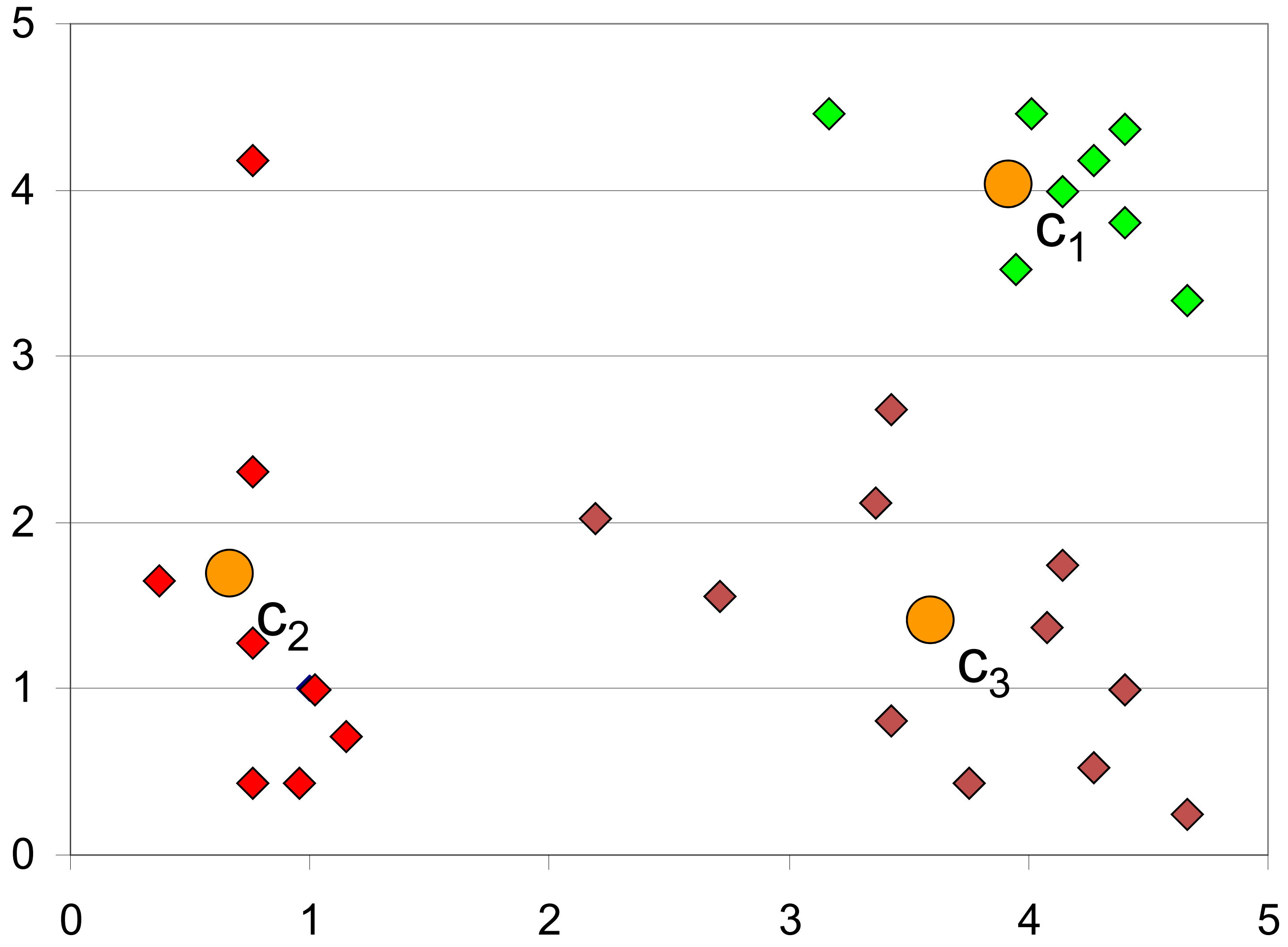
K-means Clustering: Step 3



K-means Clustering: Step 4

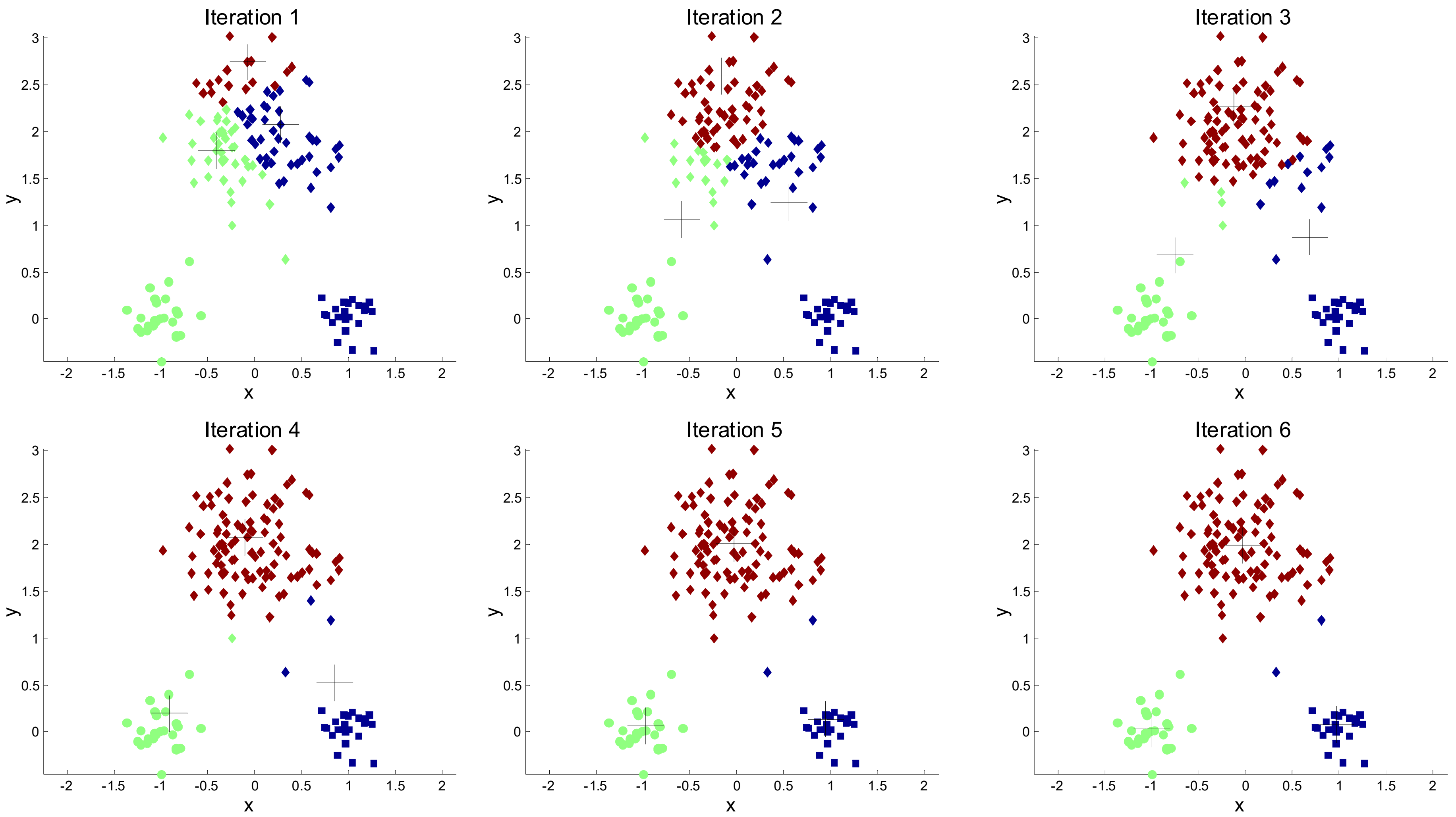


K-means Clustering: Step 5



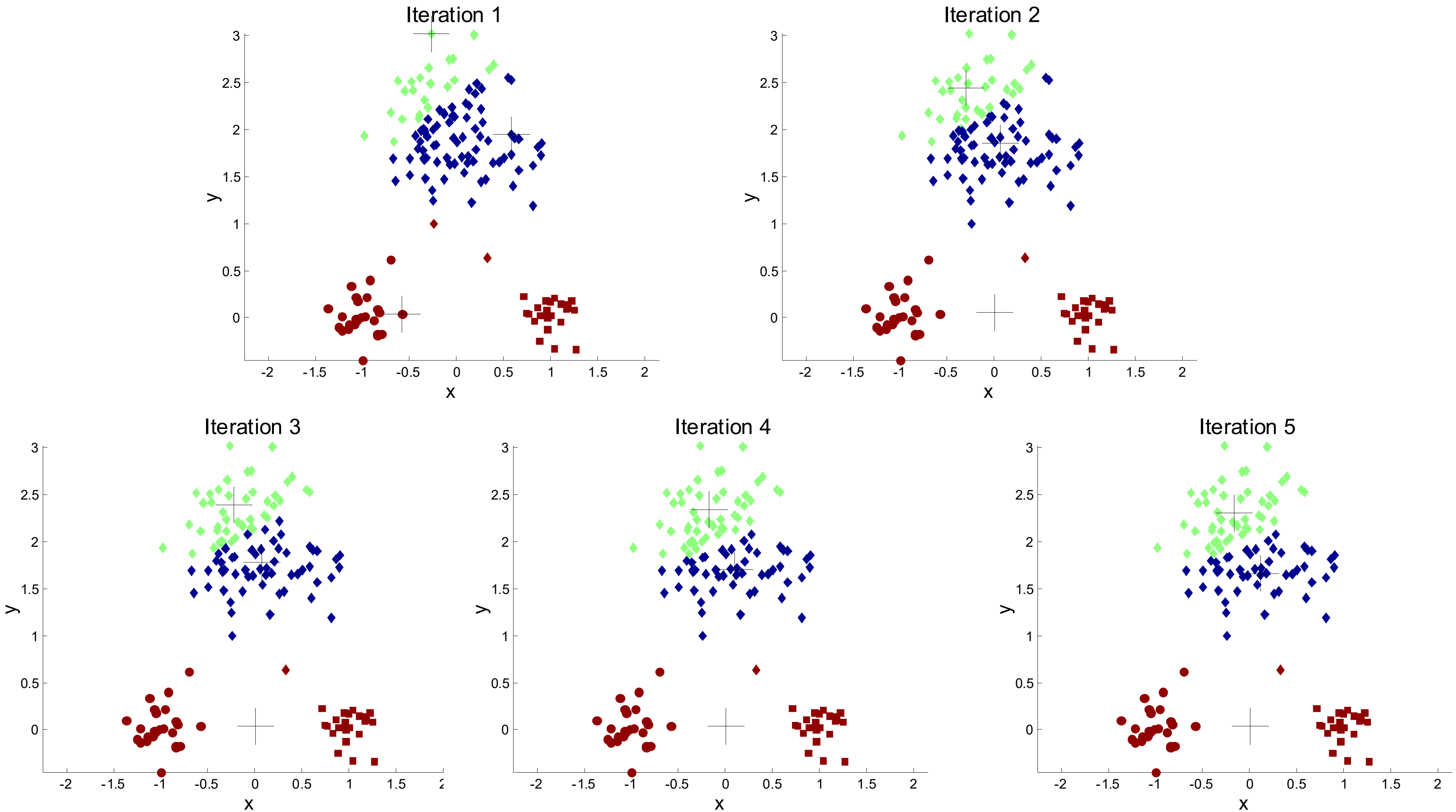
Issue 1: Initialization of Centroids

- Initial centroids are often chosen **randomly**
- Clusters produced vary from one run to another



Issue 1: Initialization of Centroids

- Initial centroids are often chosen **randomly**
- Clusters produced vary from one run to another



Issue 1: Initialization of Centroids

- Do multiple runs and
  select the clustering with the smallest SSE error
- Select original set of points by methods
  other than random
  - E.g., pick the most distant (from each other) points
    as cluster centers
    (**Furthest Centers Heuristic** algorithm)

Issue 1: Initialization of Centroids

$c_1 = pick\ a\ random\ point$

$for\ i = 2\ to\ K:$

$c_i = point\ that\ is\ furthest\ from\ \boldsymbol{any}\ previous\ centers$

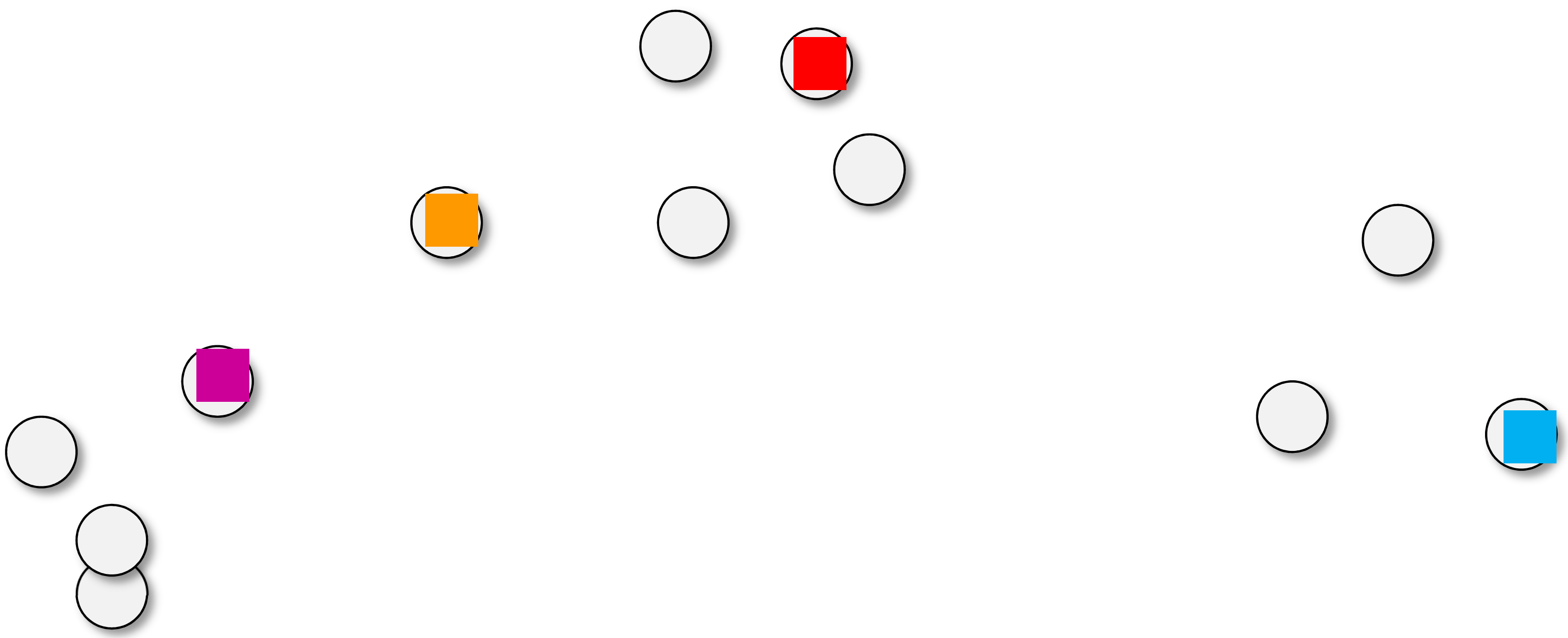
$$c_i = \underset{x}{\operatorname{argmax}} \min_{c_j: 1 < j < i} d(x, c_j)$$

point with the largest distance to any previous center

smallest distance from x to any previous center

Issue 1: Initialization of Centroids

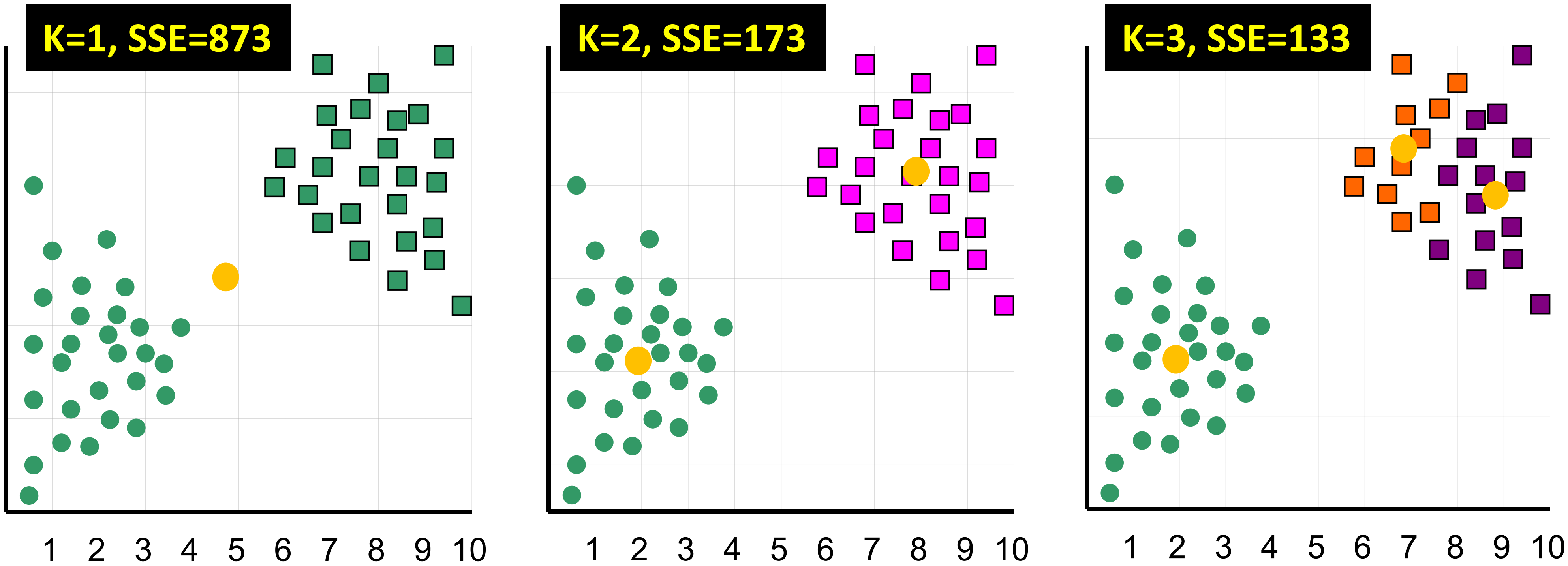
- Initialize furthest from centers



Furthest point from center

Issue 2: What about K?

- Idea 1: try different K and do validation
  - What should we optimize? SSE
- Idea 2: let the domain expert look at the clustering and decide if they like it
  - How should we show this to them?



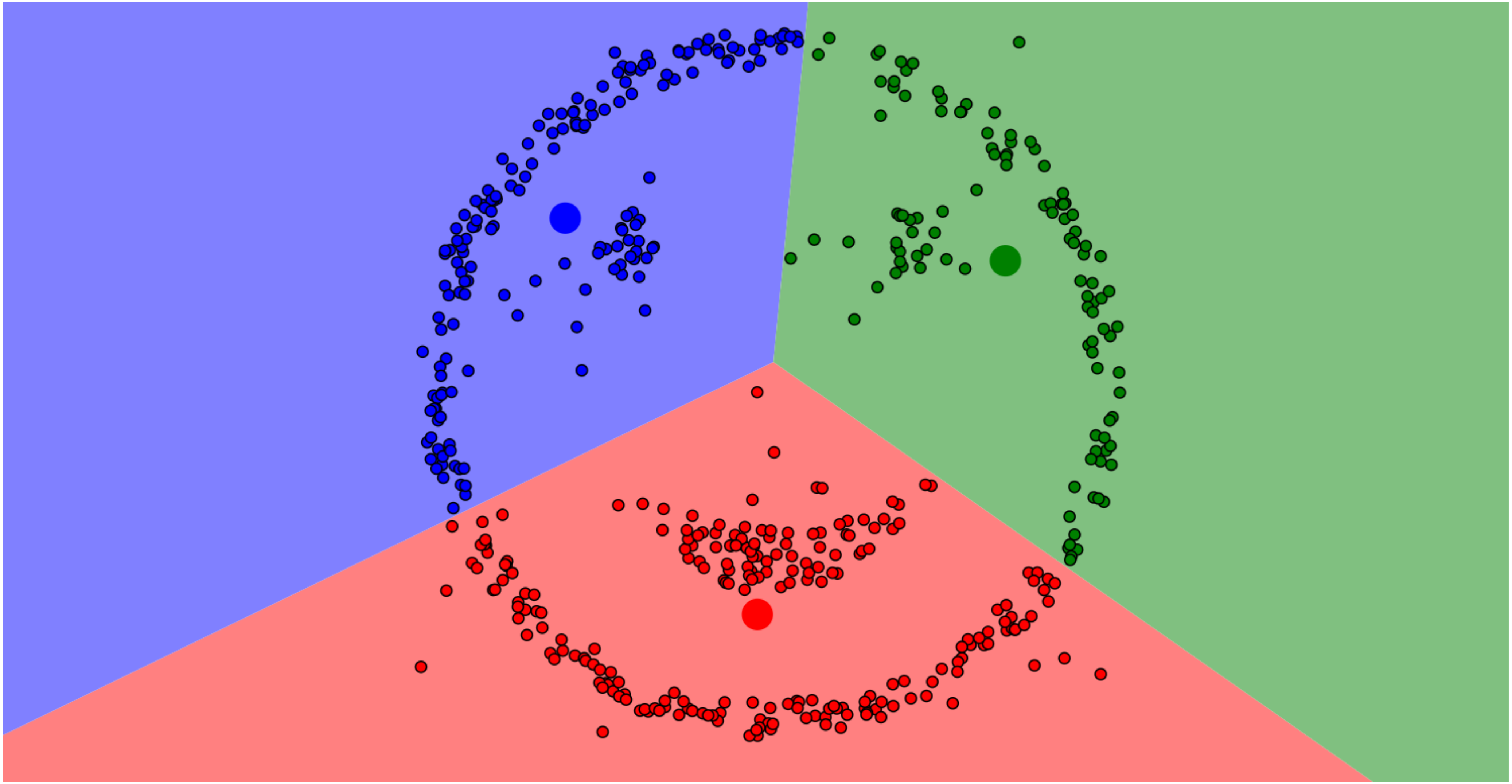
Issue 2: What about K?

- We can plot the objective function values for $K = 1 \sim 6$
  - The abrupt change at $k = 2$, is highly suggestive of two clusters
  - Known as "**knee finding**" or "elbow finding"

Issue 3: Convergence

- K-means converges for common similarity measures
  - Most of the convergence happens in few iterations
  - Often the stopping condition is changed to "until relatively few points change clusters"
- In general a fast and efficient algorithm

K-means DEMO

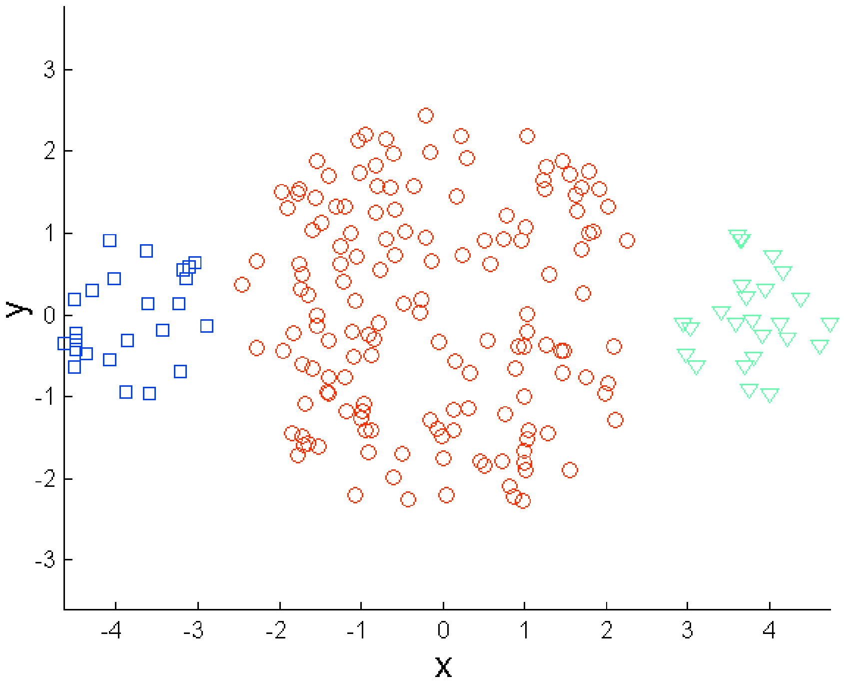


https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

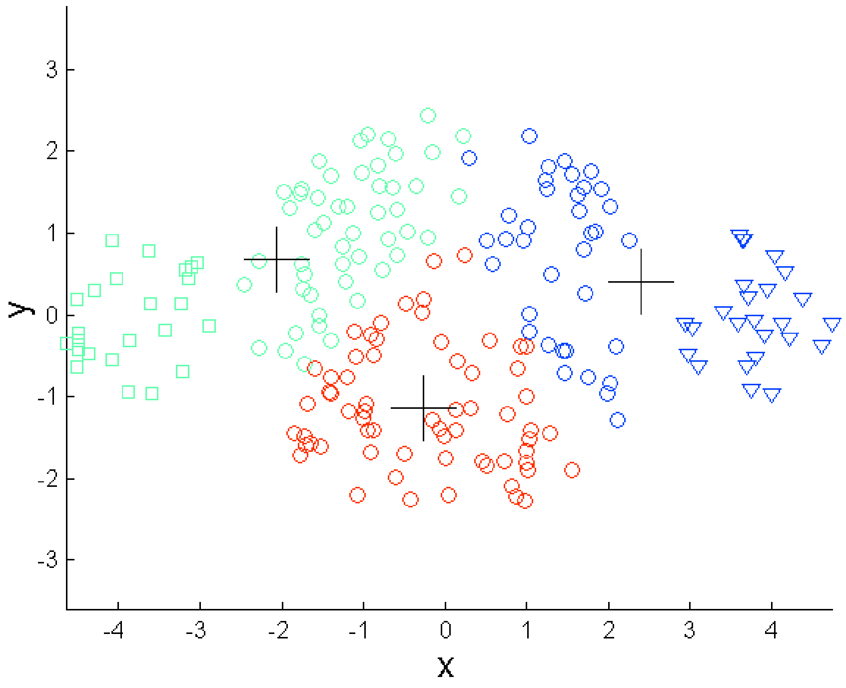
Limitation of K-means

- K-means has problems
  when clusters are of different:
  - Sizes
  - Densities
  - Non-globular shapes
- K-means also has problems
  when the data contains outliers

Limitations of K-means: Differing Sizes

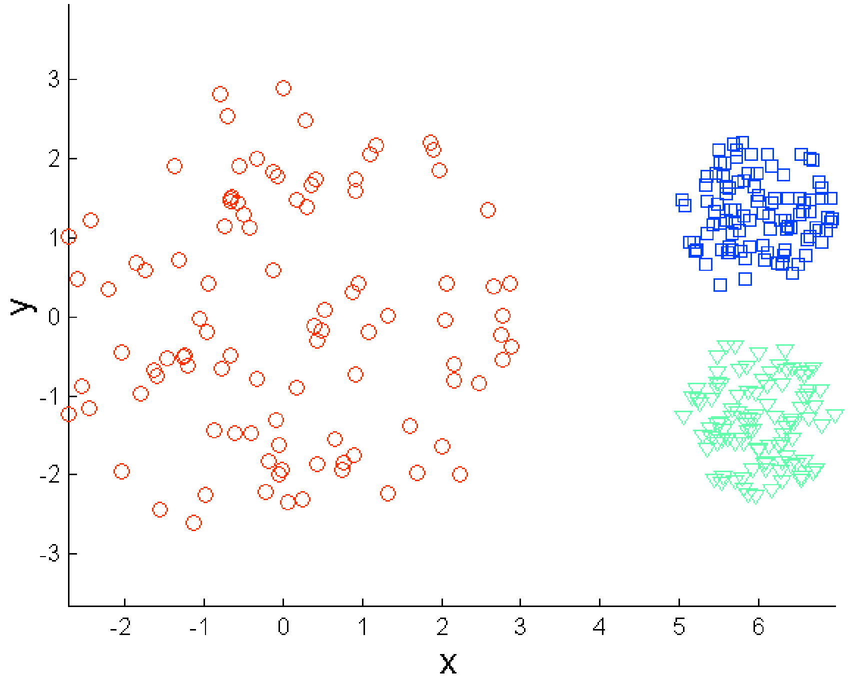


Original Points

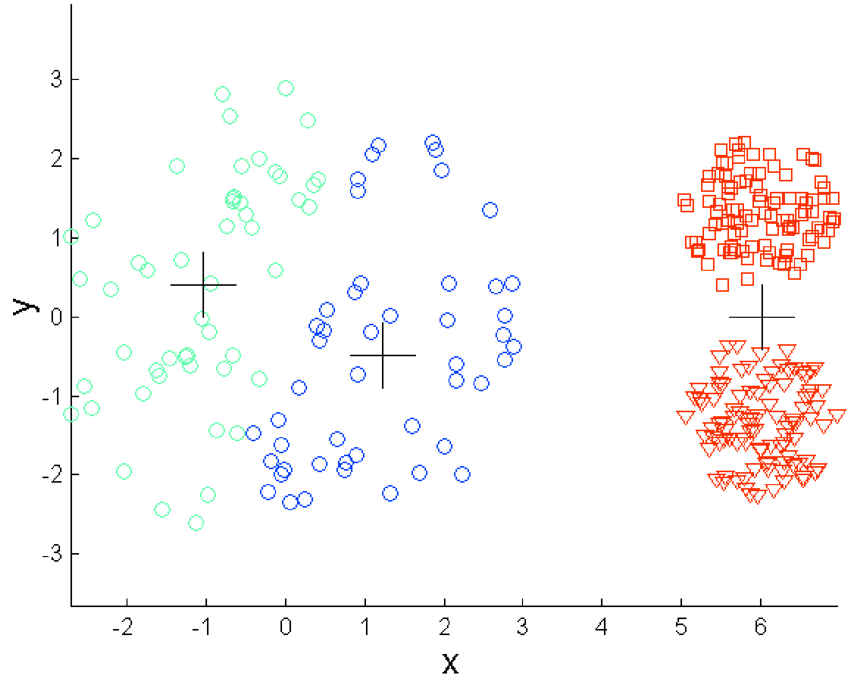


K-means (3 Clusters)

Limitations of K-means: Differing Density

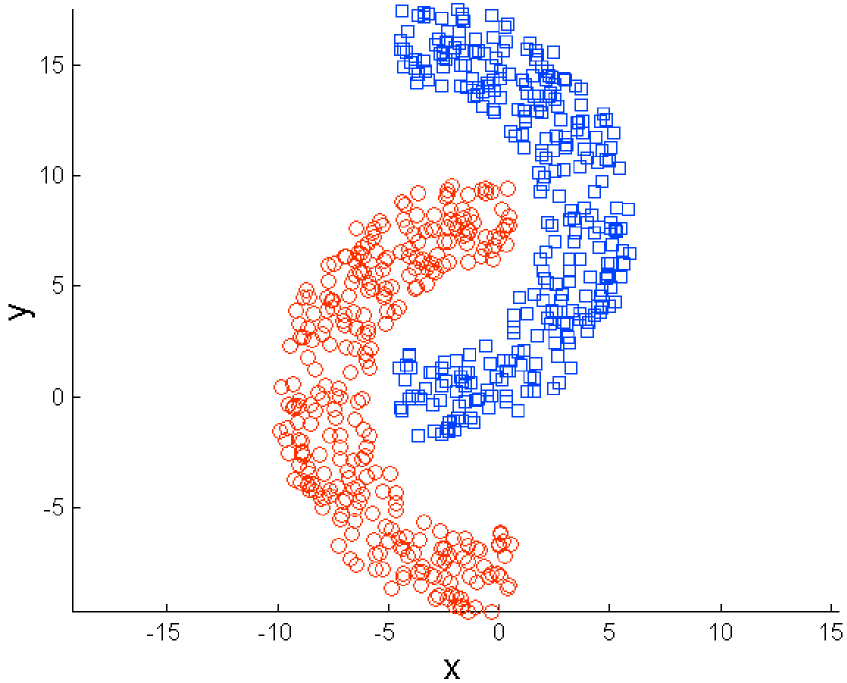


Original Points

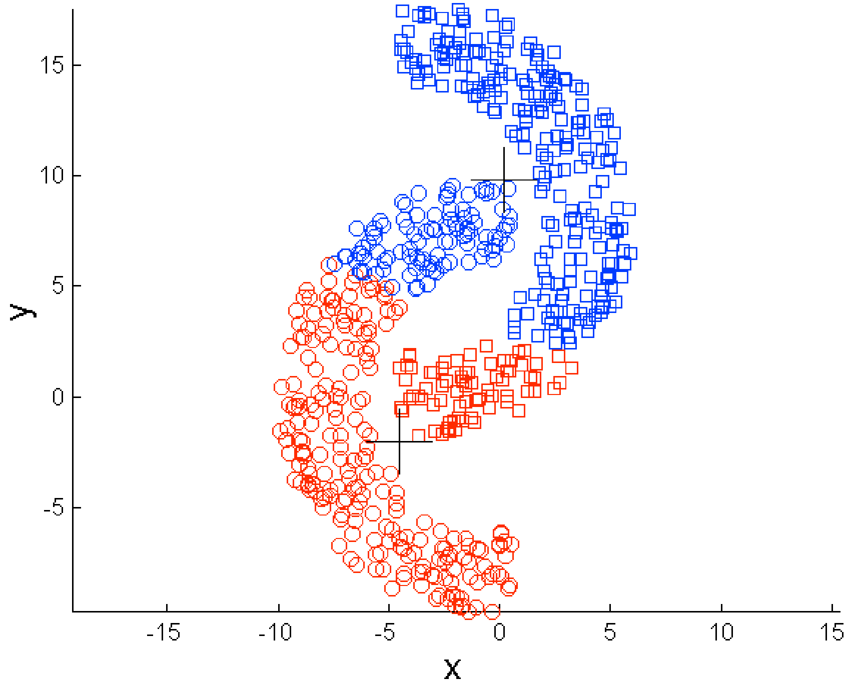


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes



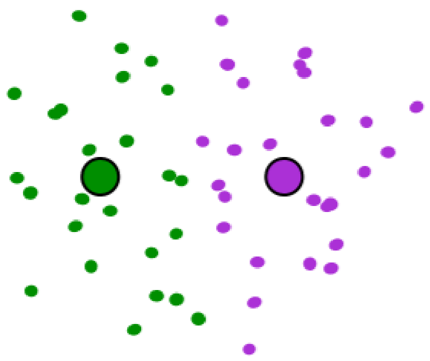
Original Points



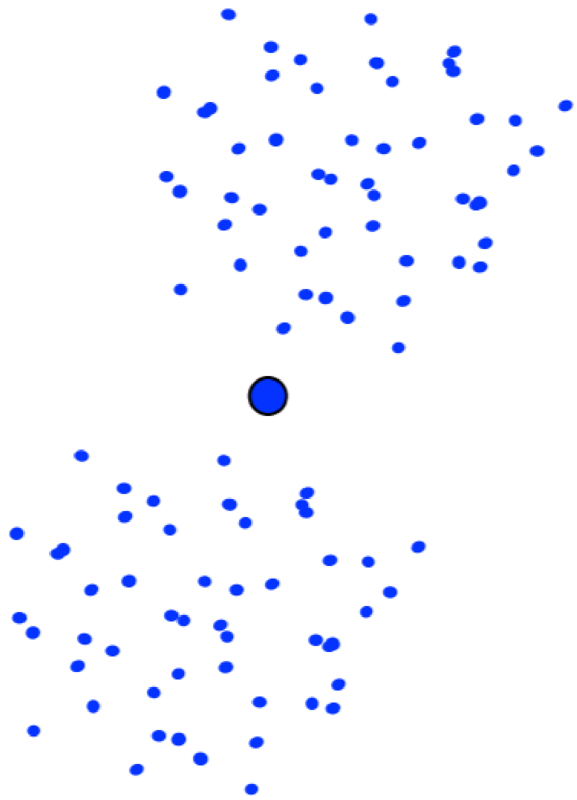
K-means (2 Clusters)

Limitations of K-means: Getting Stuck

A Local Optimum:

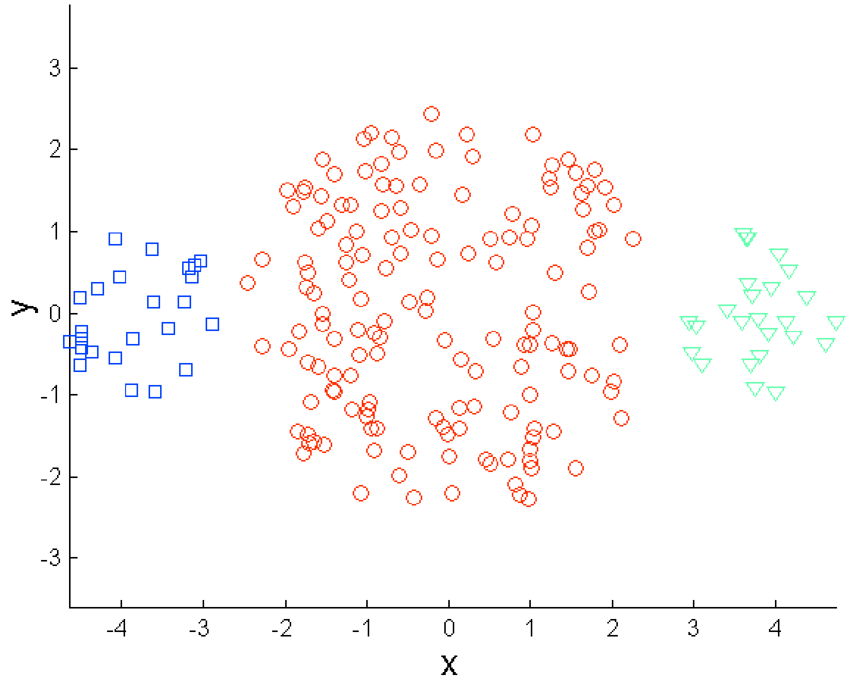


Would be better to
have one cluster here

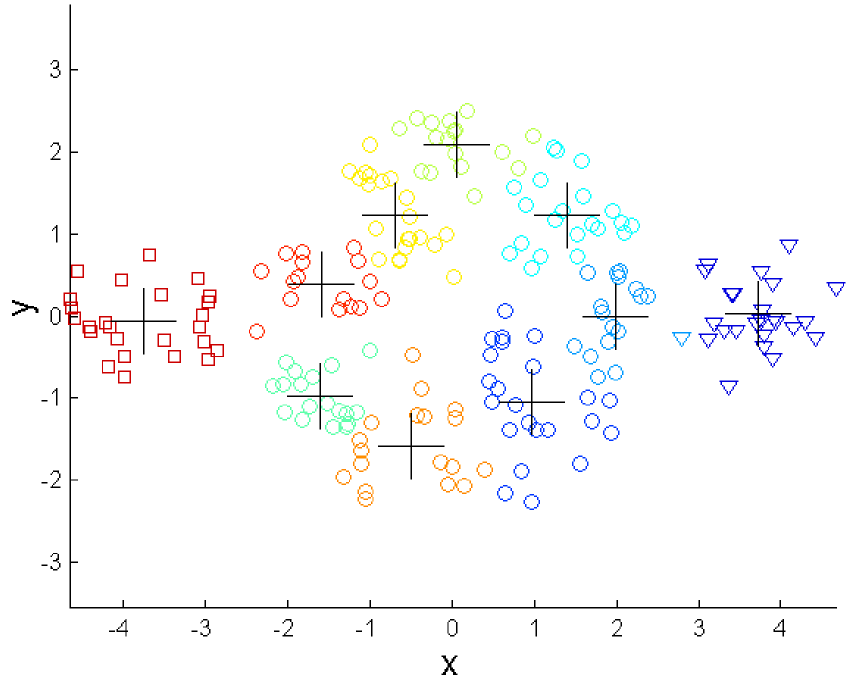


and two clusters here...

Overcoming K-means Limitations



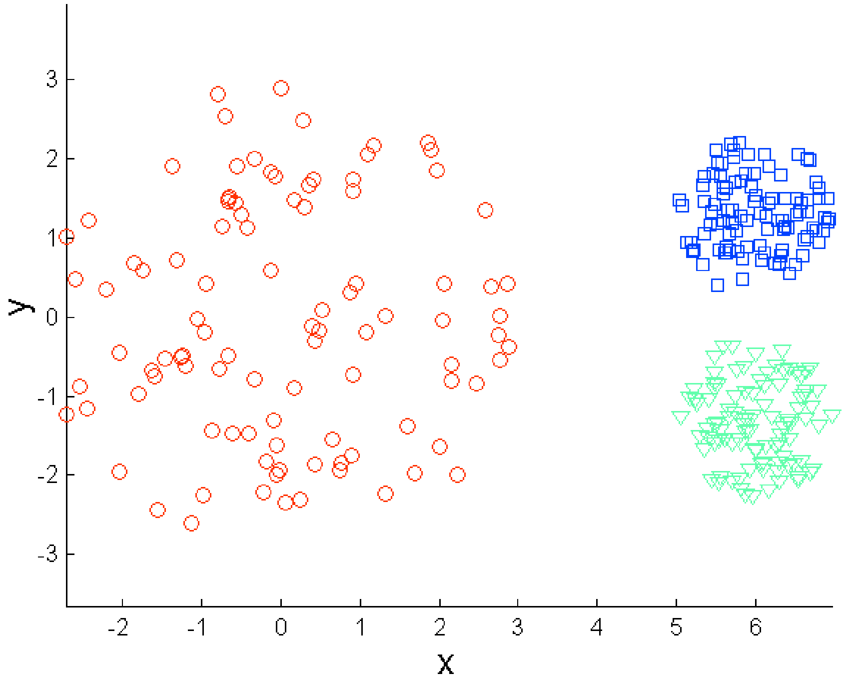
Original Points



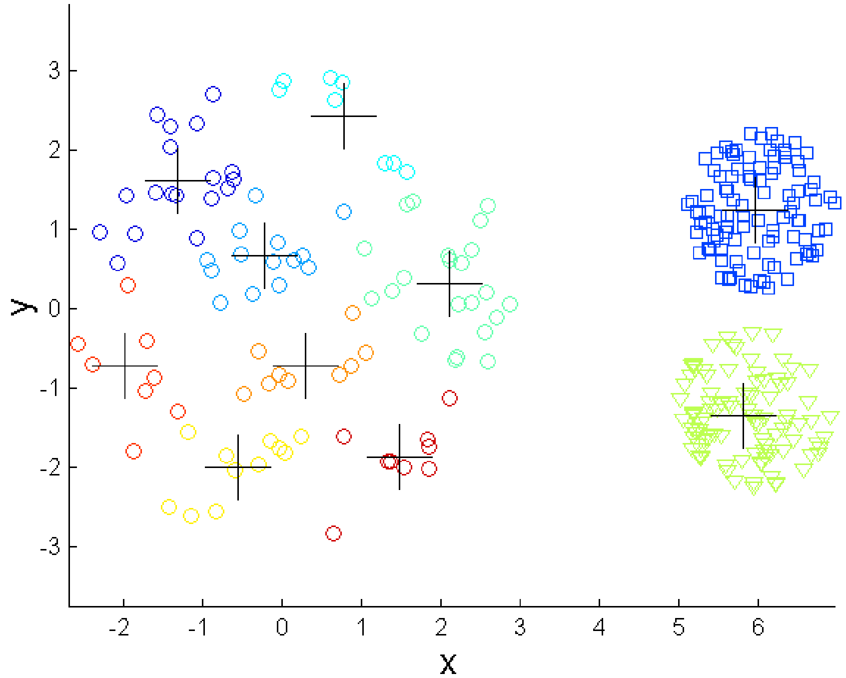
K-means (3 Clusters)

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

Overcoming K-means Limitations

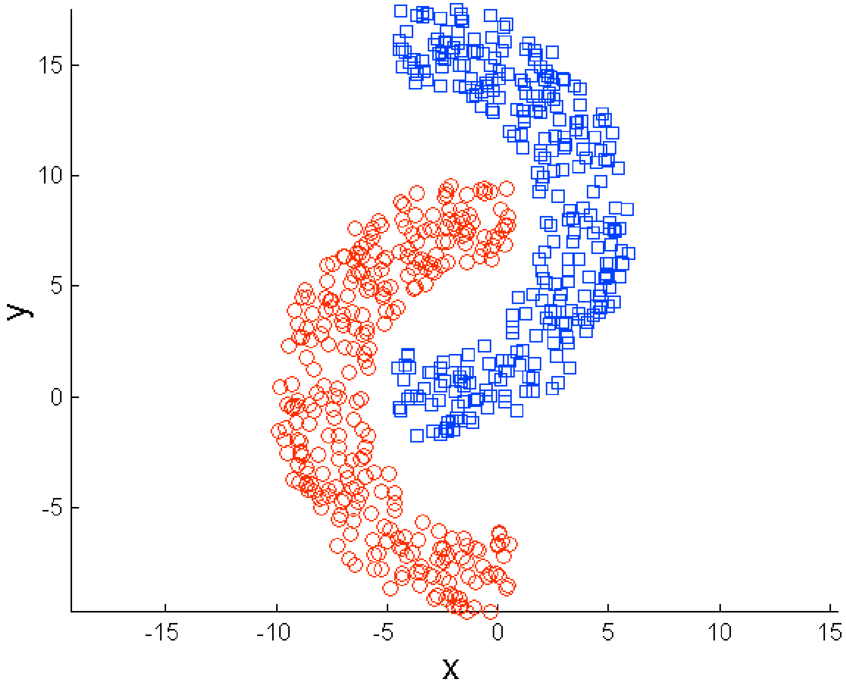


Original Points

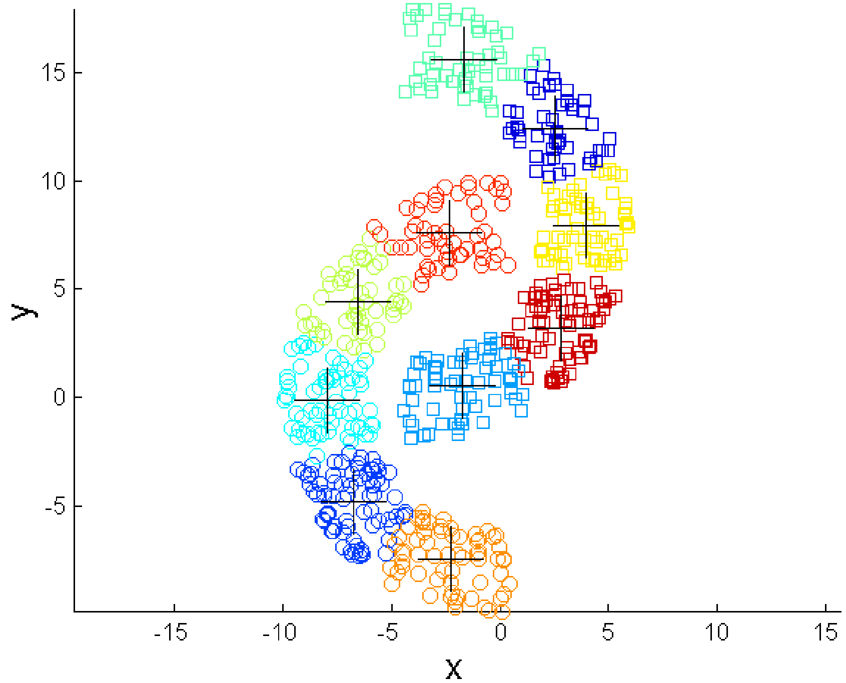


K-means (3 Clusters)

Overcoming K-means Limitations



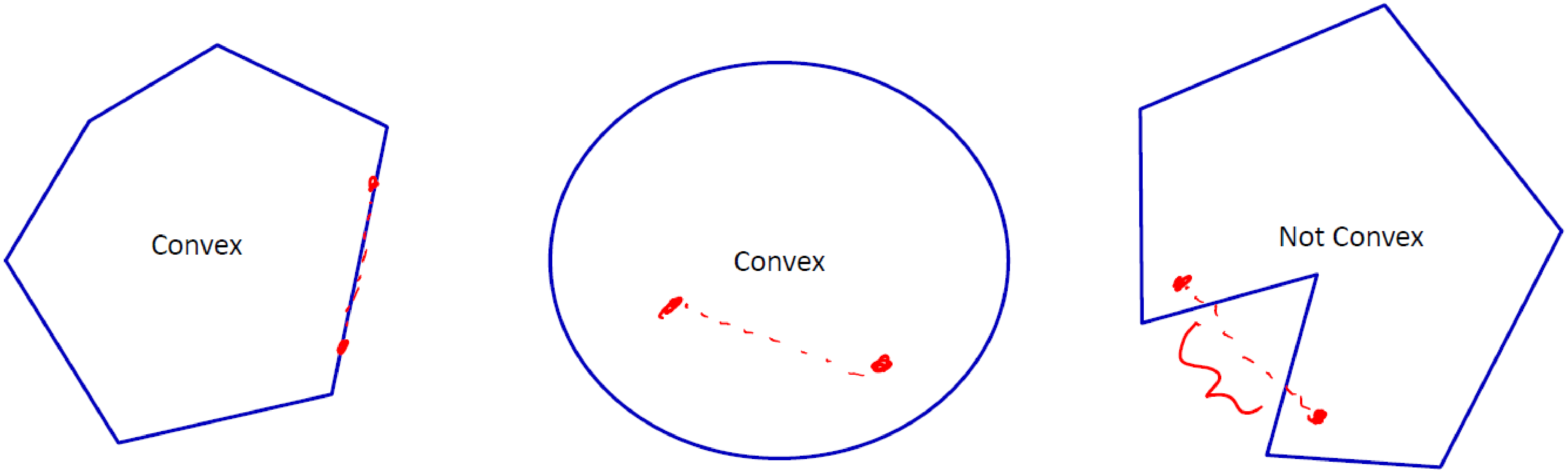
Original Points



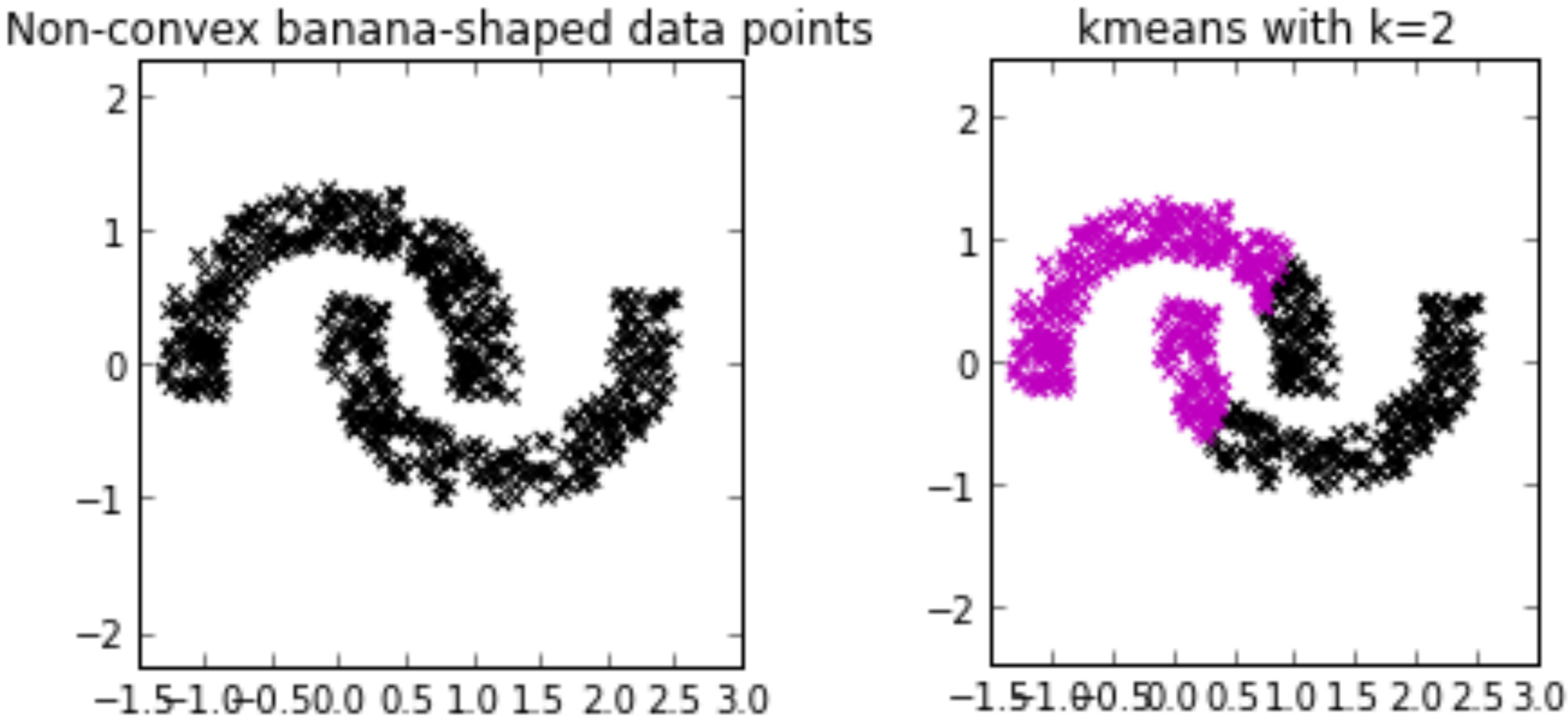
K-means (2 Clusters)

Convex Sets

- A set is convex if line between two points in the set stays in the set



K-means with Non-Convex Clusters

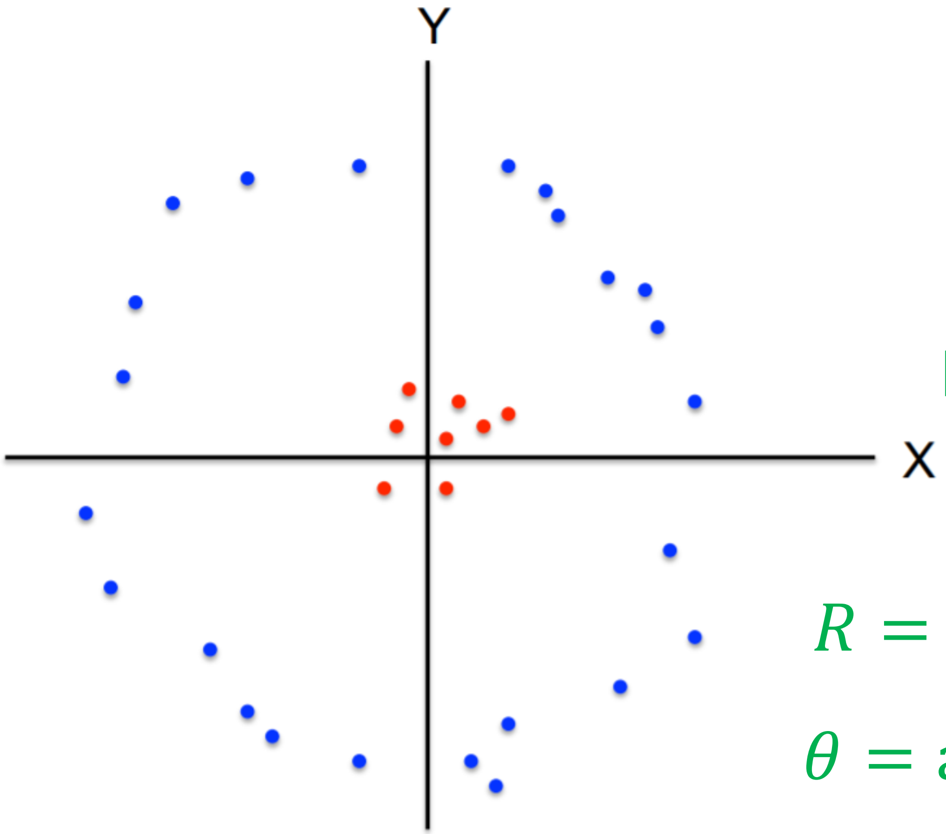


K-means cannot separate non-convex clusters

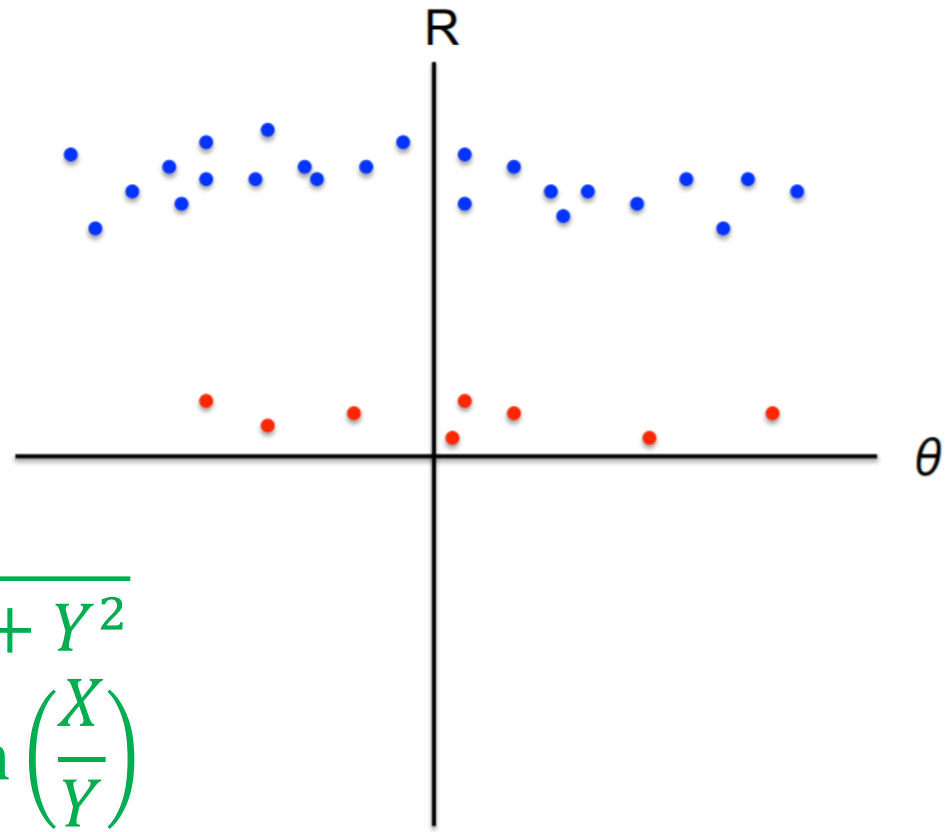
Overcoming K-means Limitations

Feature Transformation

K-means not able to properly cluster



Changing the features (distance function) can help



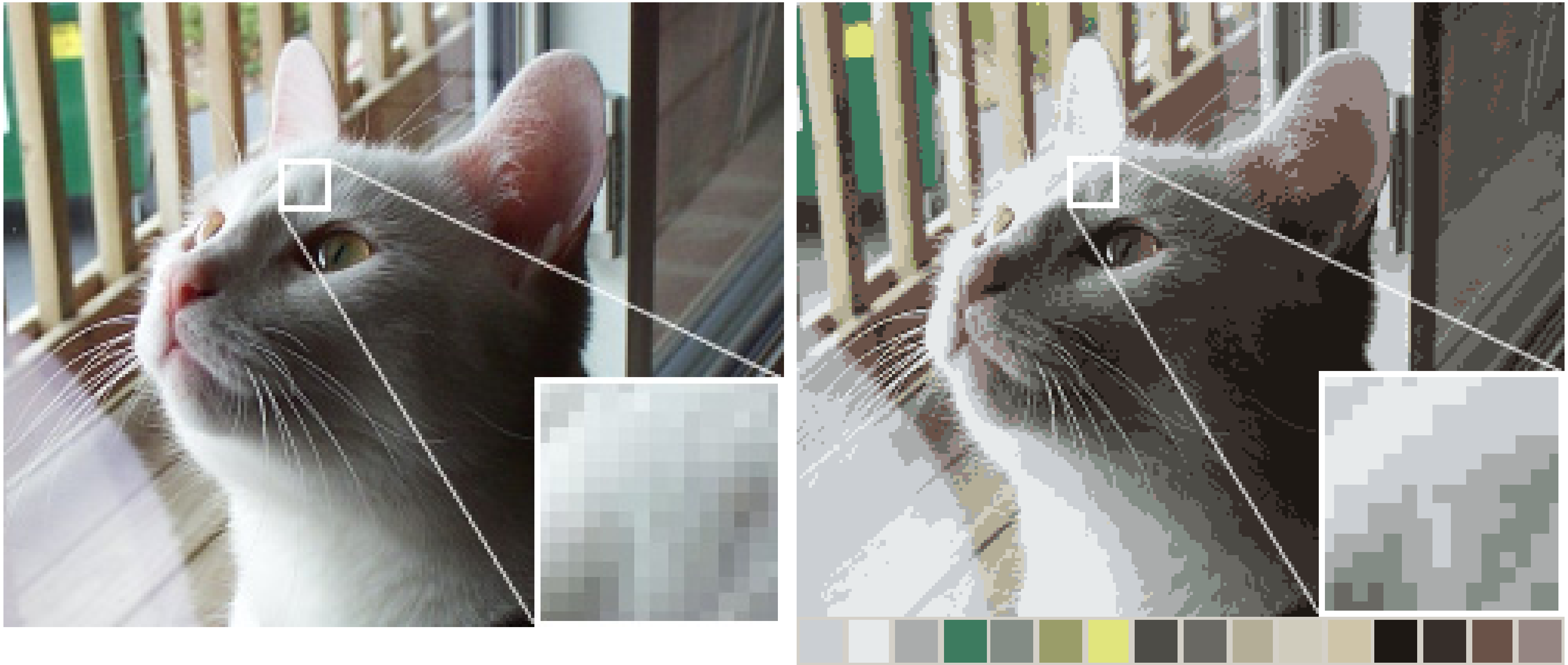
$$R = \sqrt{X^2 + Y^2}$$

$$\theta = \arctan\left(\frac{X}{Y}\right)$$

Applications: Color Quantization

Image Compression

gif, png

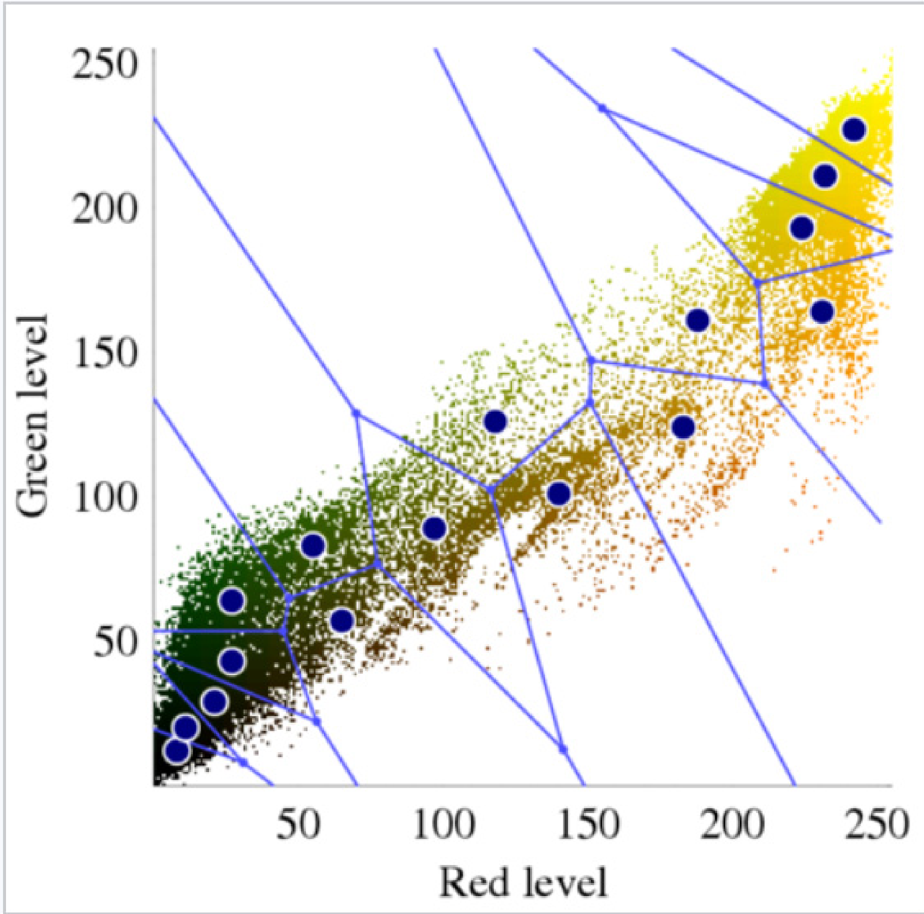


Another application: optimal palette generation

Applications: Color Quantization

https://en.wikipedia.org/wiki/Color_quantization

A small photograph that has had its blue channel removed. This means all of its pixel colors lie in a two-dimensional plane in the color cube.



The color space of the photograph to the left, along with a 16-color optimized palette produced by Photoshop. The Voronoi regions of each palette entry are shown.

Clustering is the Key to **Big Data** Problem

- Not feasible to **"label"** a large collection of objects
- No prior knowledge of the number and nature of groups (clusters) in the dataset
- Clusters may evolve over time
- Clustering provides efficient browsing, search, recommendation and organization of data