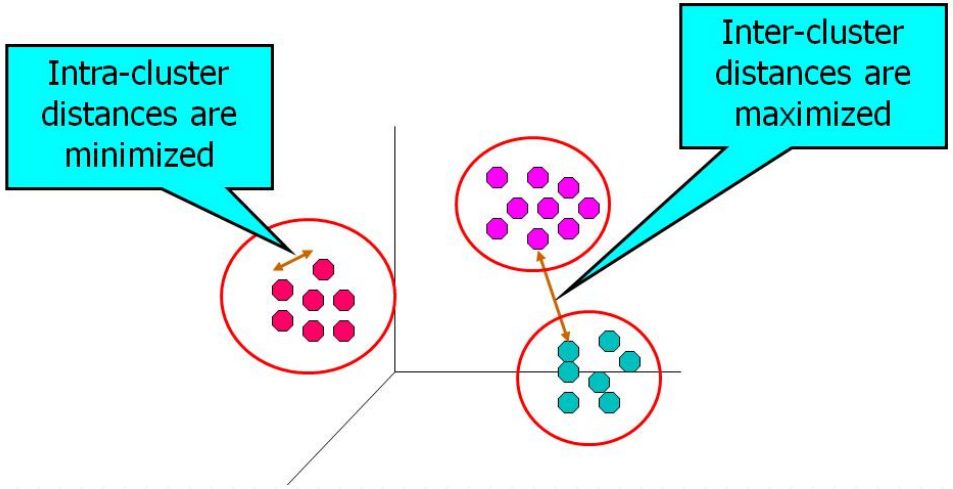
Clustering & Association

Clustering - Overview

- What is cluster analysis?
  - Grouping data objects based only on information found in the data describing these objects and their relationships
  - Maximize the similarity within objects in the same group
  - Maximize the difference between objects in different groups



Source: http://apandre.wordpress.com/visible-data/cluster-analysis/

Clustering - Overview

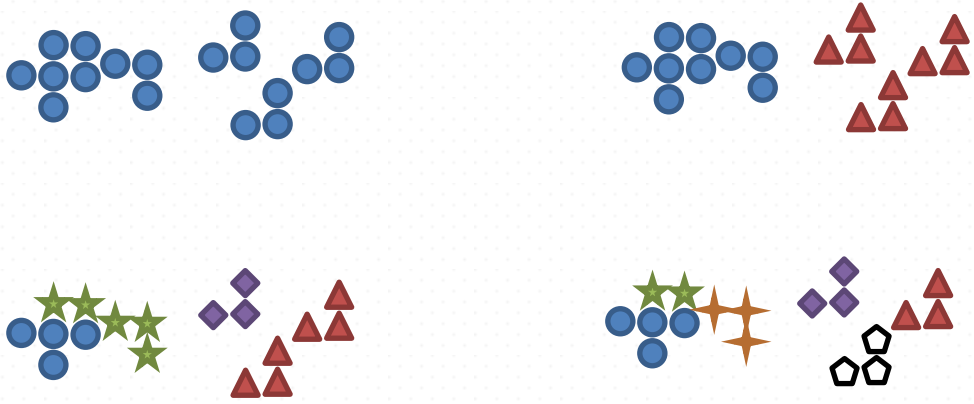
- What is cluster analysis?
  - **Similarity:**
    - Numerical measure of “alikeness” of two instances
    - Greater if more alike
    - Can be normalized to [0,1]
  - **Dissimilarity:**
    - How different two instances are
    - Lower for “alike” instances
  - These are usually expressed in terms of a distance function
  - Different weights can be assigned to different features
  - Hard to define what “similar enough” means. The answer is typically highly subjective.

Clustering - Overview

- What is cluster analysis?
  - Different from classification (supervised learning) in that the labels for each instance are derived only from the data
  - For that reason, cluster analysis is referred to as *unsupervised classification*

Clustering - Overview

- Not well defined at times:

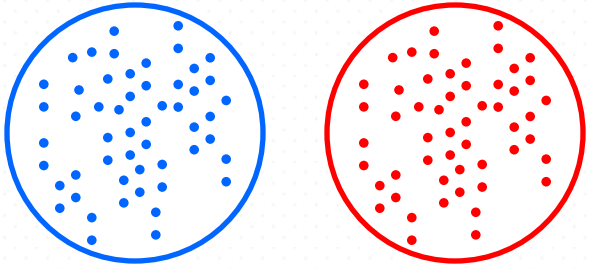


- How do we really know how many clusters should exist in the above example?

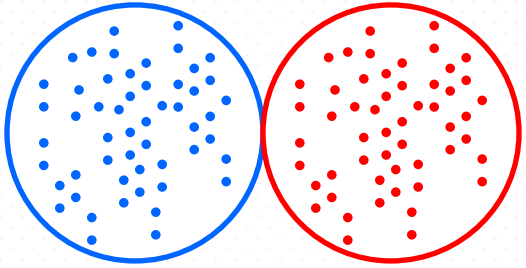
Clustering - Overview

- Different types of clusterings:
  - Hierarchical versus Partitional
  - Exclusive versus Overlapping versus Fuzzy
  - Complete versus Partial
- Different types of clusters:
  - Well separated
  - Prototype-Based
  - Density-Based
  - Shared-Property

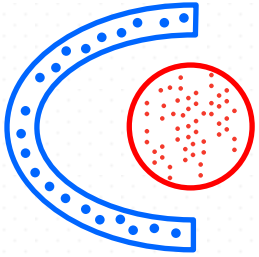
Clustering - Overview



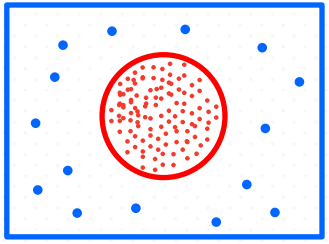
(a) Well separated clusters



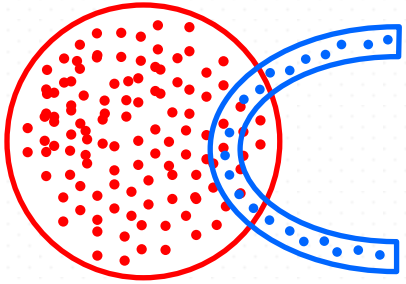
(b) Center-based clusters



(c) Contiguity-based clusters



(d) Density-based clusters



(e) Conceptual clusters

Clustering

1. K-means clustering
2. Hierarchical clustering
3. Evaluating clusters

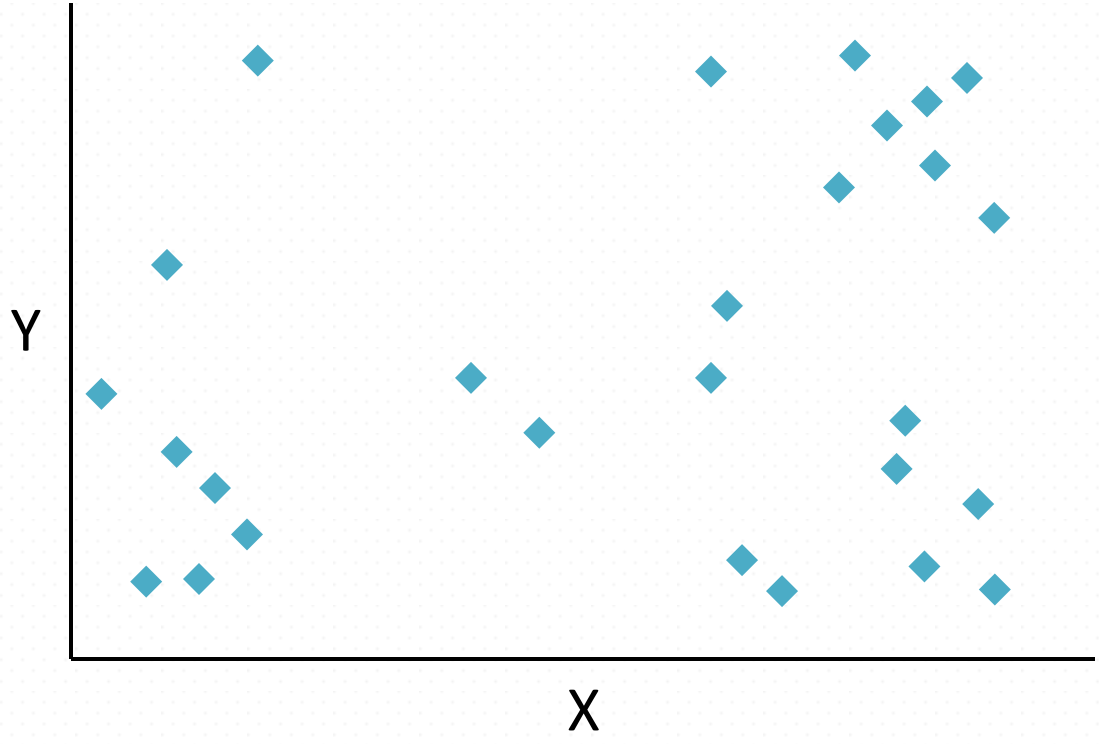
K-means clustering

- Works with numeric data only!
- Algorithm:

1. Pick a number *k* of random cluster centers
2. Assign every item to its nearest cluster center using a distance metric
3. Move each cluster center to the mean of its assigned items
4. Repeat 2-3 until convergence (change in cluster assignment less than a threshold)

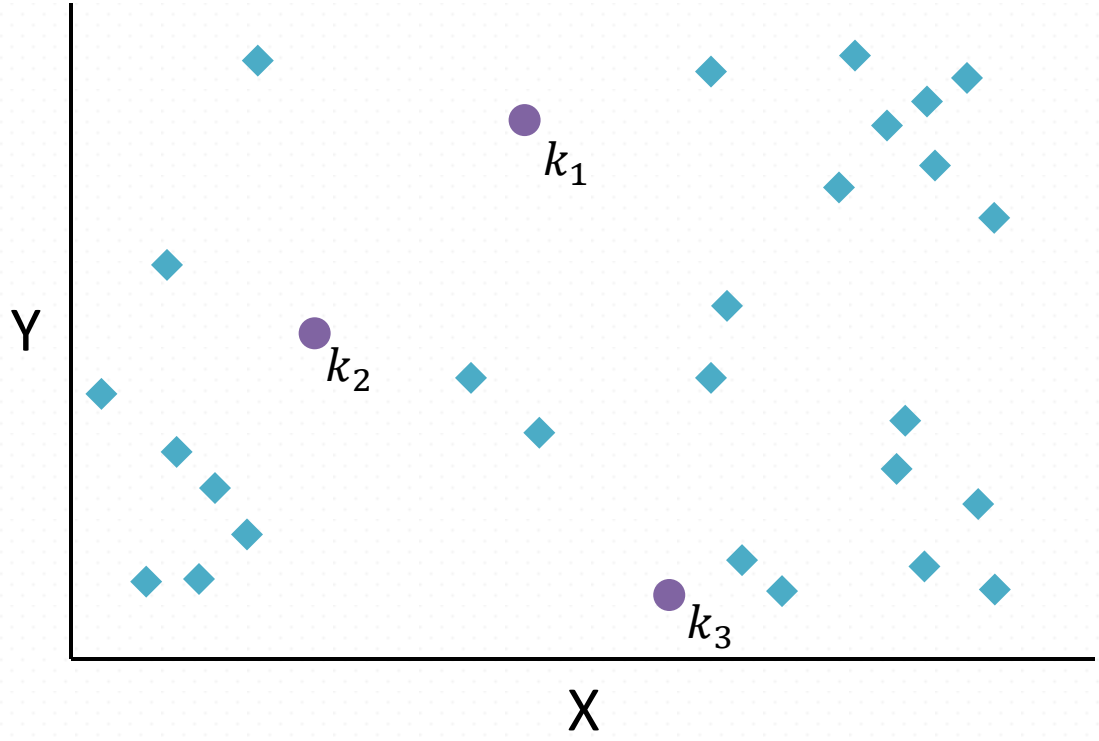
K-means clustering – Example

Suppose we wish to cluster these items



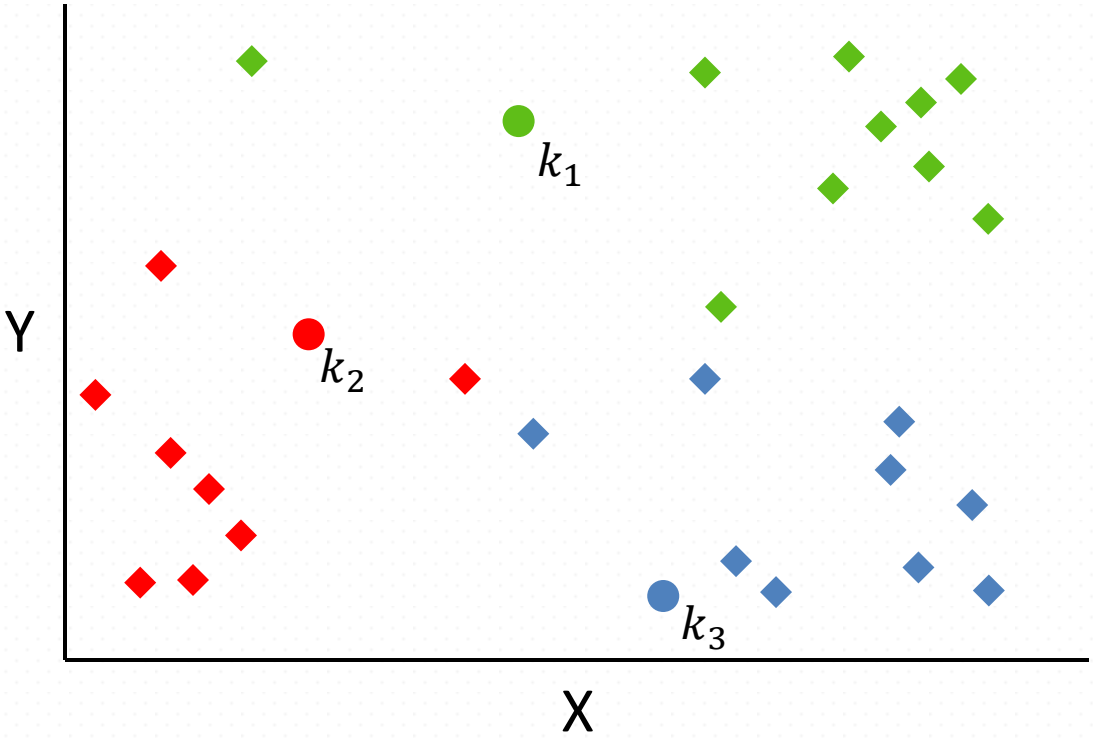
K-means clustering – Example

Pick 3 initial cluster centers at random



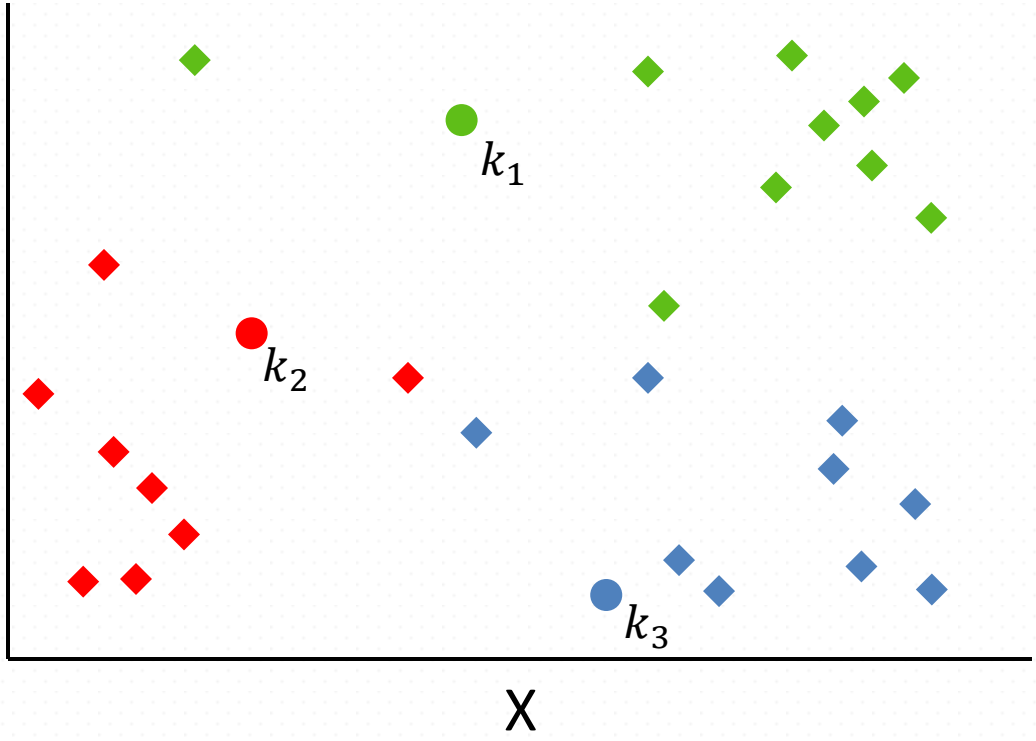
K-means clustering – Example

Assign each instance to the closest cluster center



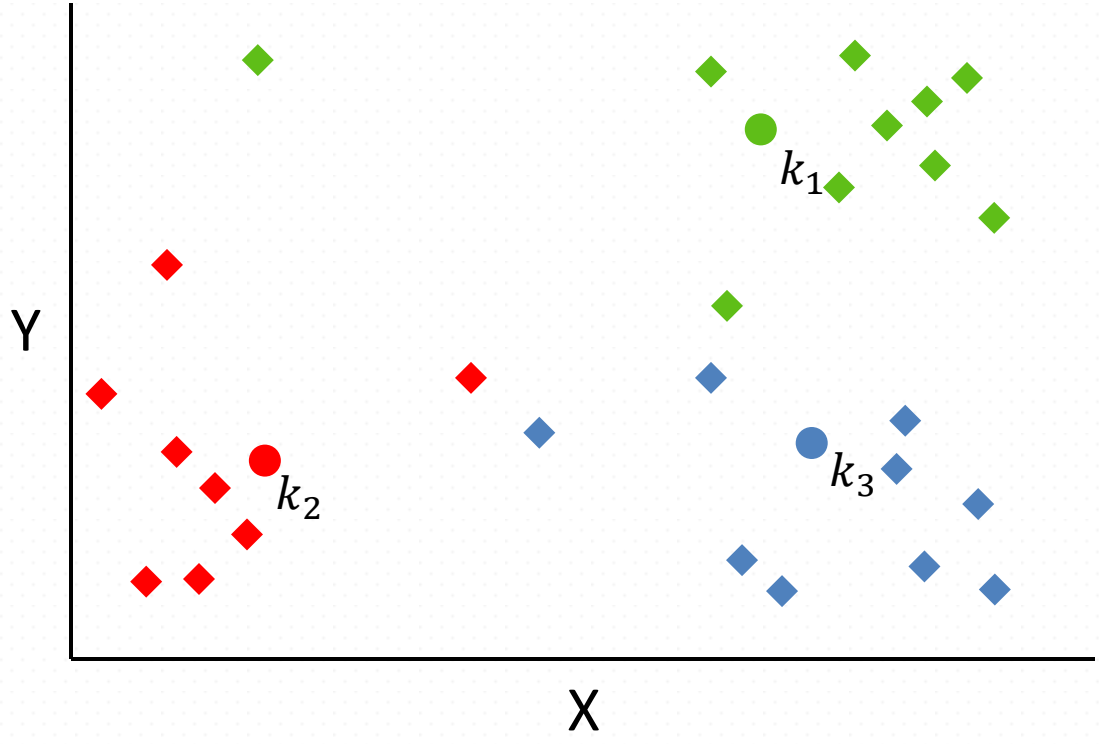
K-means clustering – Example

Move each cluster center to the mean of each cluster



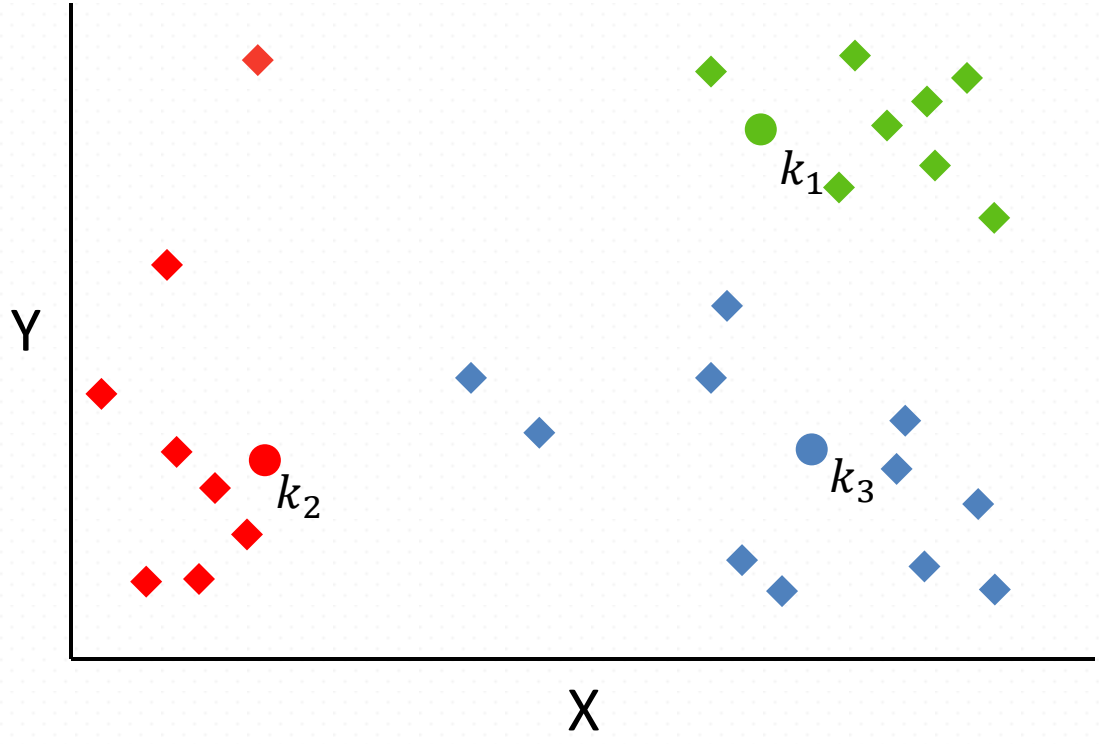
K-means clustering – Example

Reassign points closest to a different new cluster center



K-means clustering – Example

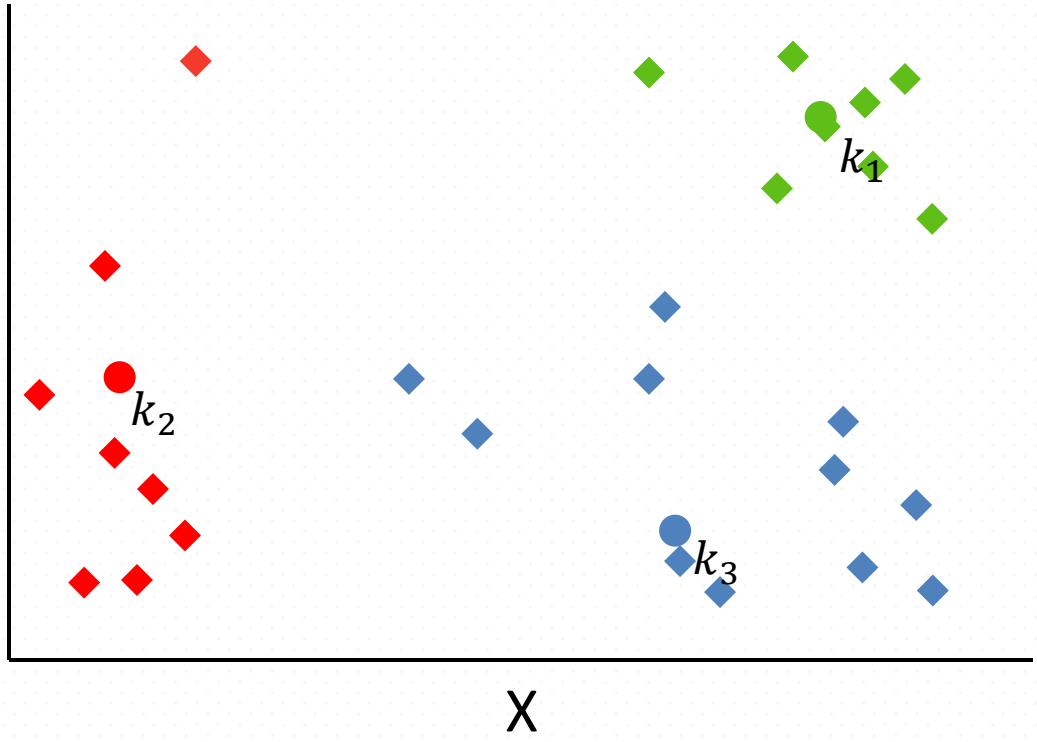
Recompute cluster means



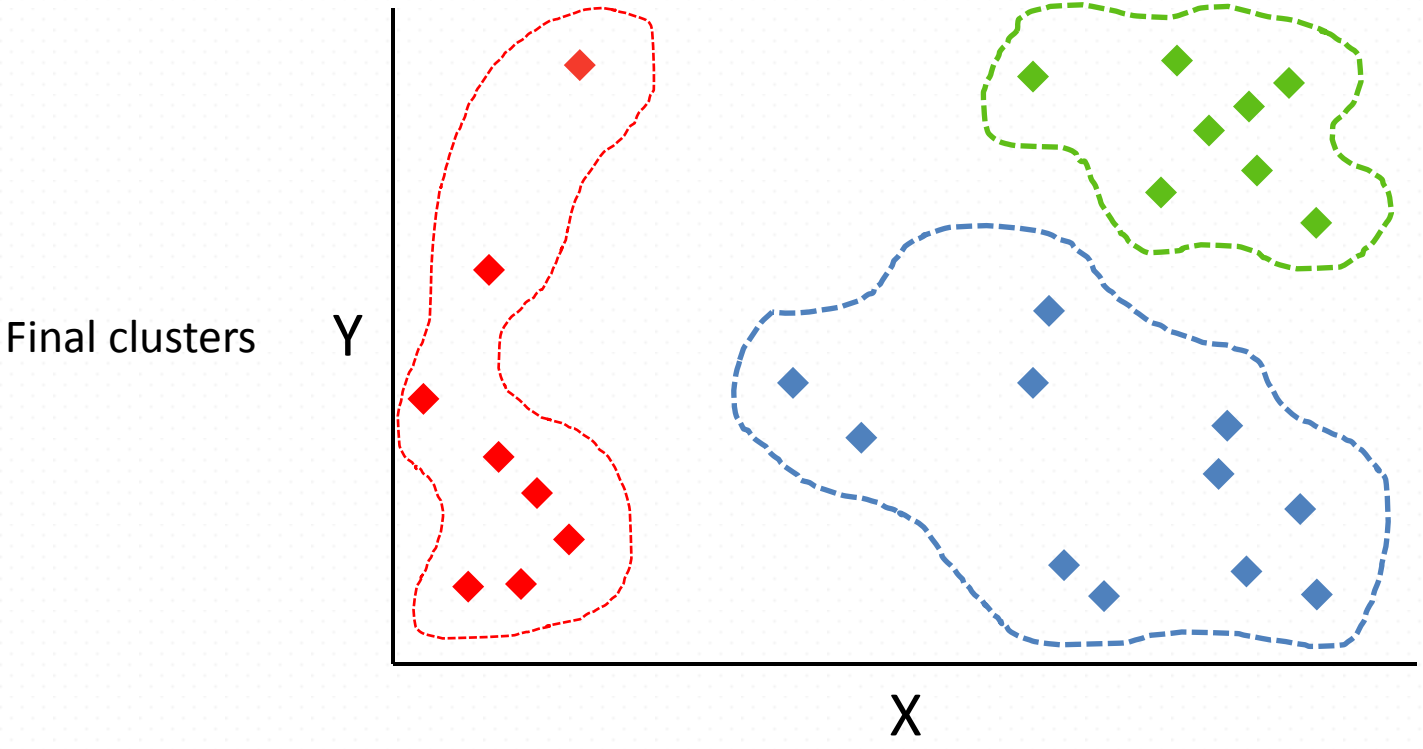
K-means clustering – Example

Reassign labels

No change. Done!



K-means clustering – Example



K-means – Evaluating clusters

- If Euclidian distance is utilized as the proximity measure, the quality of clusters can be evaluated by the sum of squared error (SSE)
  - Calculate the distance between each point and its closest centroid (error)
  - Sum the square of all errors

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

- It is important to choose values of k that minimize SSE

K-means – Discussion

- Results can vary significantly depending on initial choice of seeds
- How do you tackle this?
- How do you choose k?

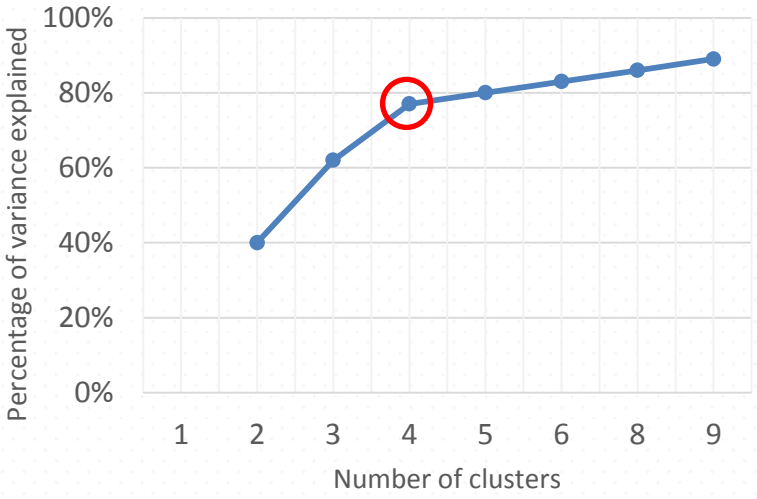
How do you choose k?

- Determining the number of clusters in a data set
  - The choice of k is often ambiguous and dependent on scale and distribution of your dataset
  - However, some generic methods for doing that do exist

- Rule of thumb:
  - $k \approx \sqrt{n/2}$, where n is the number of instances
  - This is a good starting point, but not a very reliable approach

How do you choose *k*?

- The *Elbow Method*:
  - Choose a number of clusters that covers most of the variance



How do you choose *k*?

- Other methods:
  - Information Criterion Approach
  - Silhouette method
  - Jump method
  - Gap statistic

K-means variations

- K-medoids – instead of mean, use medians of each cluster
  - Mean of 1, 3, 5, 7, 1009 is **205**
  - Median of 1, 3, 5, 7, 205 is **5**
  - Advantage: not affected by extreme values
- For large datasets, use sampling

K-means pros and cons

- **Pros:** very efficient (even if multiple runs are performed), can be used for a large variety of data types.
- **Cons:** Not suitable for all types of data, susceptible to initialization problems and outliers, restricted to data in which there is a notion of a center

And now...

Lets see some k-meaning!