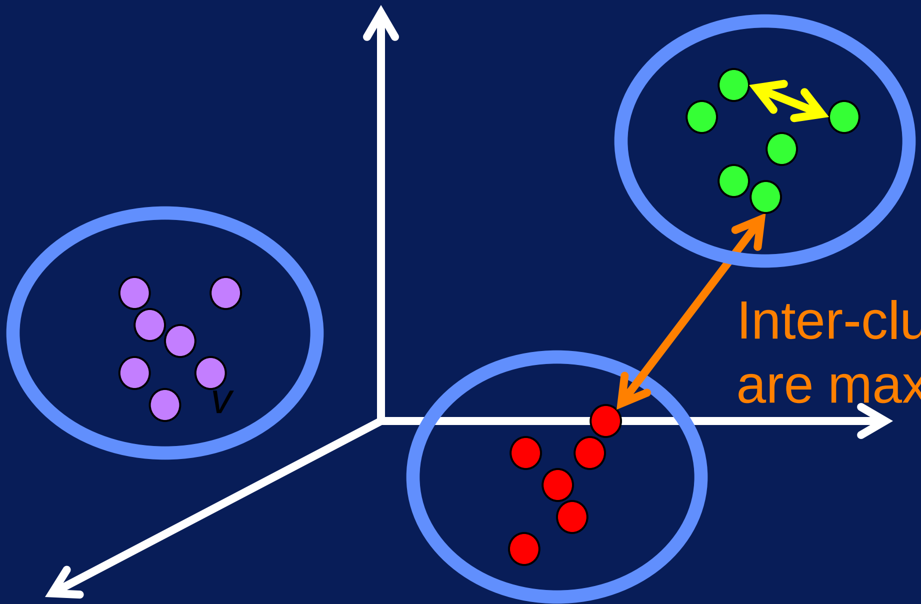
k-Means Clustering

After this video you will be able to..

- Describe the steps in the k-means algorithm
- Explain what the ‘k’ stands for in k-means
- Define what a cluster centroid is

Cluster Analysis

- Divides data into clusters
- Similar items are in same cluster

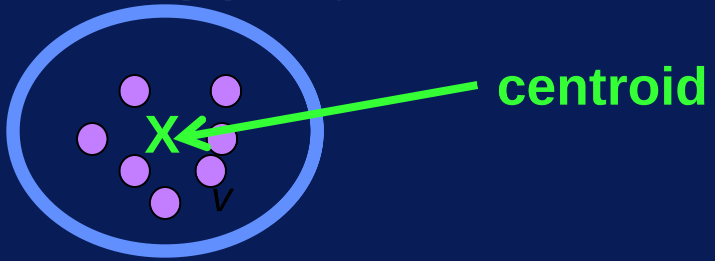


Intra-cluster differences are minimized

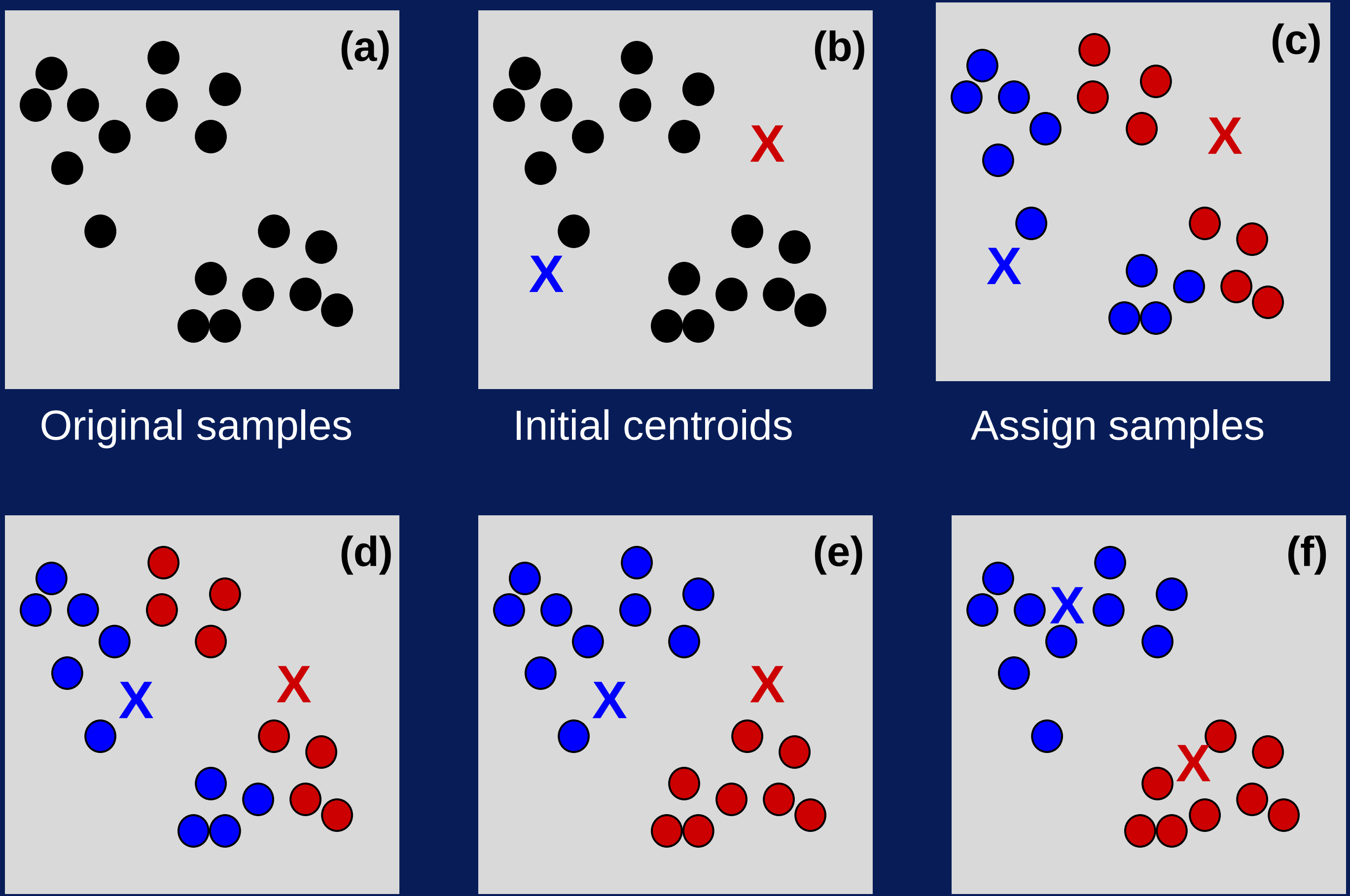
Inter-cluster differences are maximized

k-Means Algorithm

- Select k initial centroids (cluster centers)
- Repeat
  - Assign each sample to closest centroid
  - Calculate mean of cluster to determine new centroid
- Until some stopping criterion is reached



k-Means



Original samples

Initial centroids

Assign samples

Re-calculate centroids

Assign samples

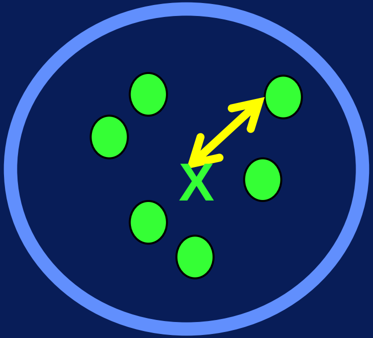
Re-calculate centroids

Choosing Initial Centroids

Issue:
Final clusters are sensitive to initial centroids

Solution:
Run k-means multiple times with
different random initial centroids,
and choose best results

Evaluating Cluster Results



error = distance between sample & centroid

squared error = error^2

Sum of squared errors
between all samples & centroid

Sum over all clusters → **WSSE**

**Within-Cluster Sum
of Squared Error**

Using WSSE

$WSSE_1 < WSSE_2$ → WSSE1 is better *numerically*

Caveats:

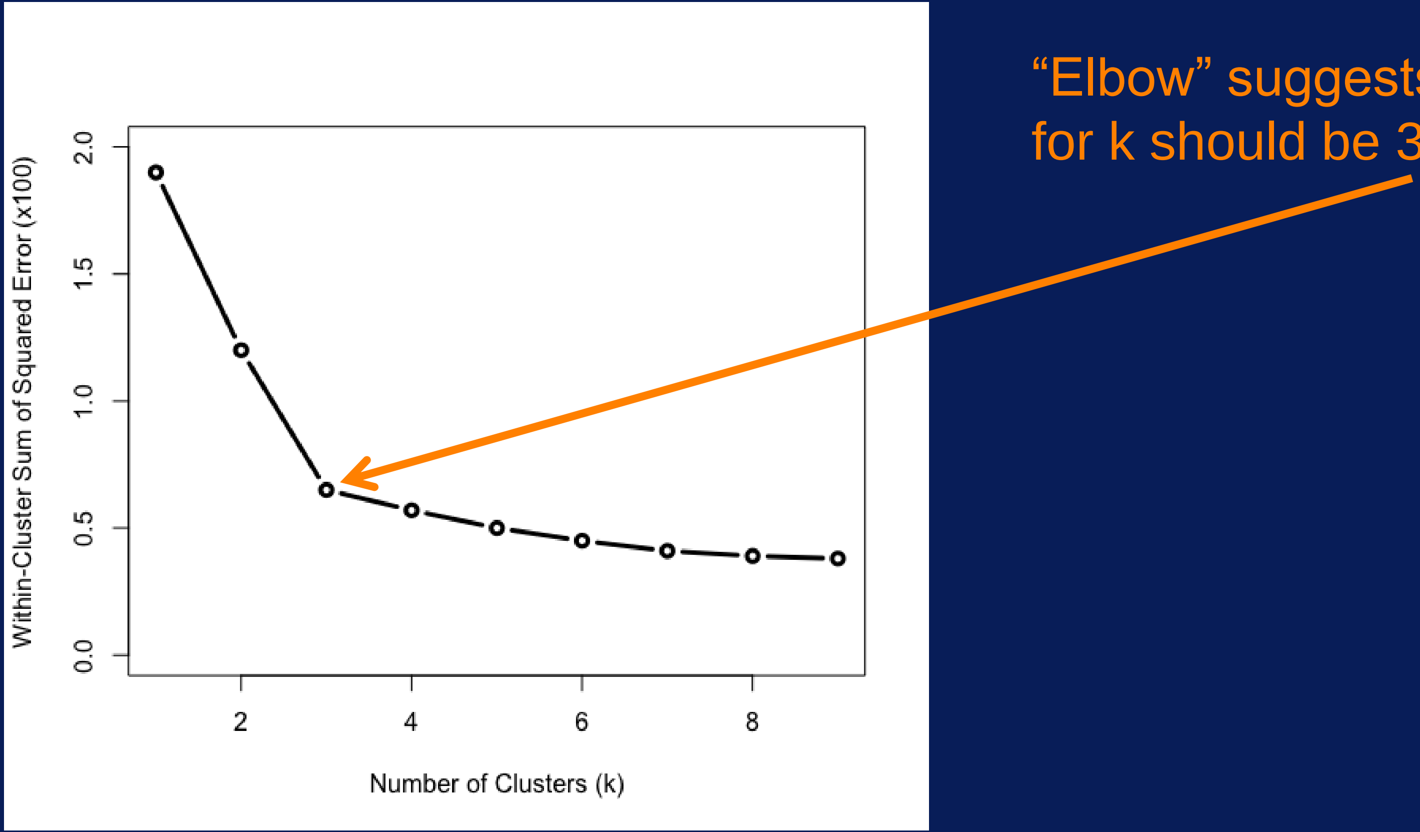
- Does not mean that cluster set 1 is more ‘correct’ than cluster set 2
- Larger values for k will always reduce WSSE

Choosing Value for k

- **Approaches:**
  - Visualization
  - Application-Dependent
  - Data-Driven

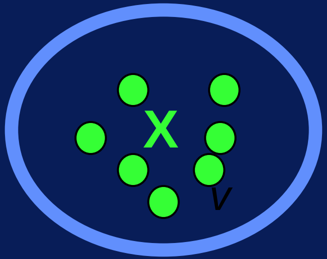
k = ?

Elbow Method for Choosing k



“Elbow” suggests value for k should be 3

Stopping Criteria

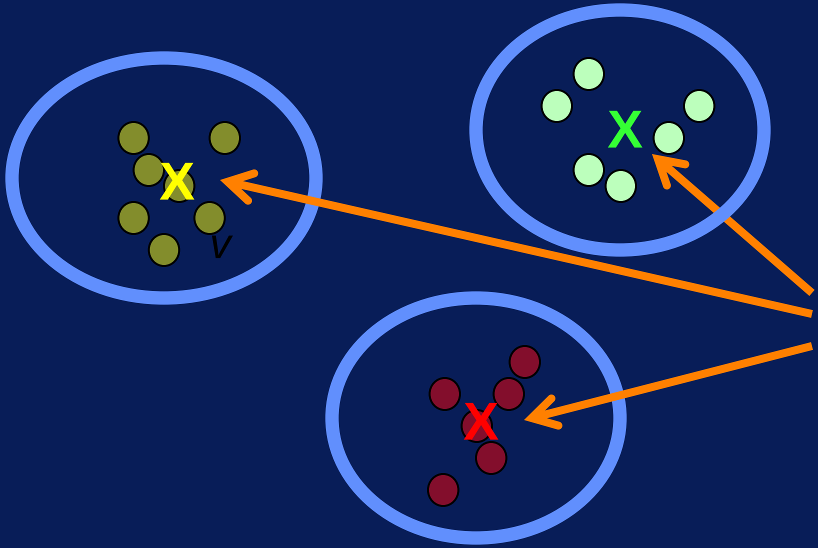


When to stop iterating?

- No changes to centroids
- Number of samples changing clusters is below threshold

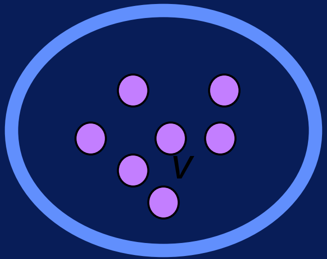
Interpreting Results

- **Examine cluster centroids**
  - How are clusters different?



Compare centroids to see how clusters are different

K-Means Summary



- Classic algorithm for cluster analysis
- Simple to understand and implement and is efficient
- Value of k must be specified
- Final clusters are sensitive to initial centroids