

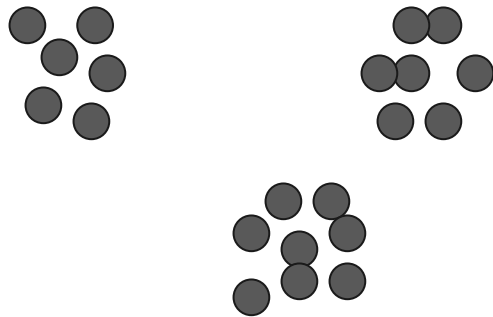


# **CIS 635 Knowledge Discovery & Data Mining**

Optimal Number of Clusters

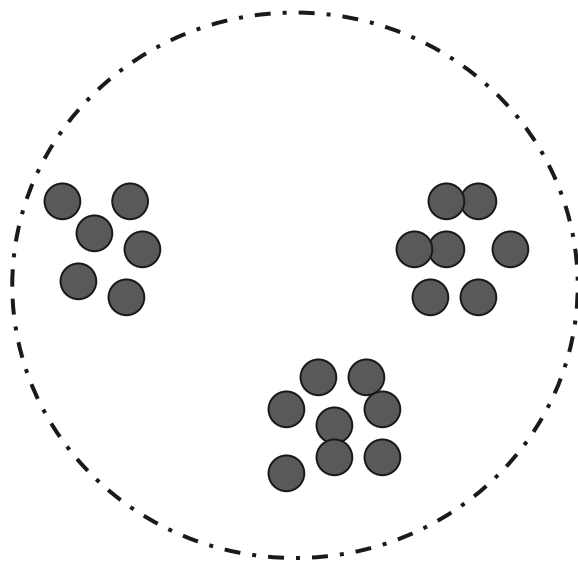


## Number of clusters



*How many clusters better describe these data points?*

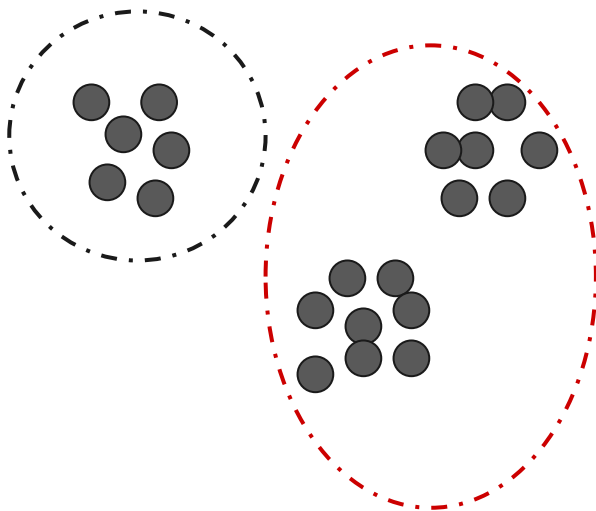
## Number of clusters



*How many clusters better describe these data points?*

1?

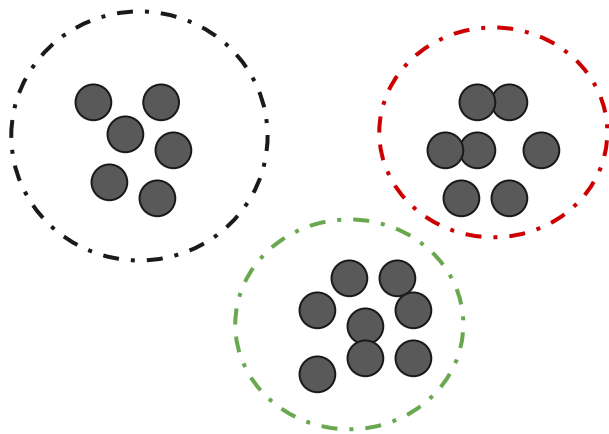
## Number of clusters



*How many clusters better describe these data points?*

2?

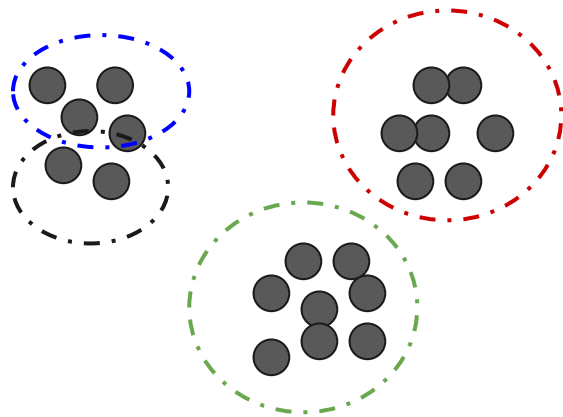
## Number of clusters



*How many clusters better describe these data points?*

3?

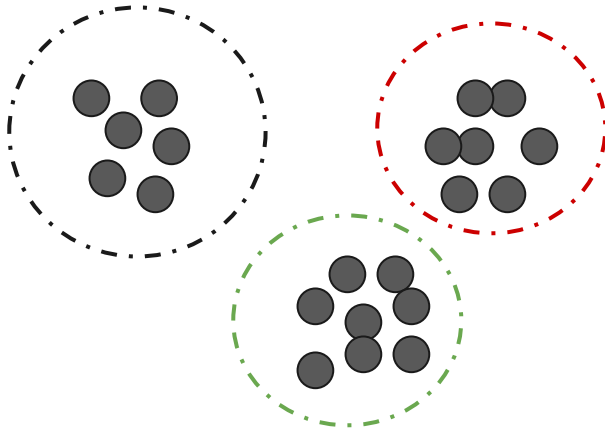
## Number of clusters



*How many clusters better describe these data points?*

4?

## Number of clusters



*How many clusters better describe these data points?*

*3 could be most likely answer; right?*



## Number of clusters

- Elbow method
- Silhouette score based





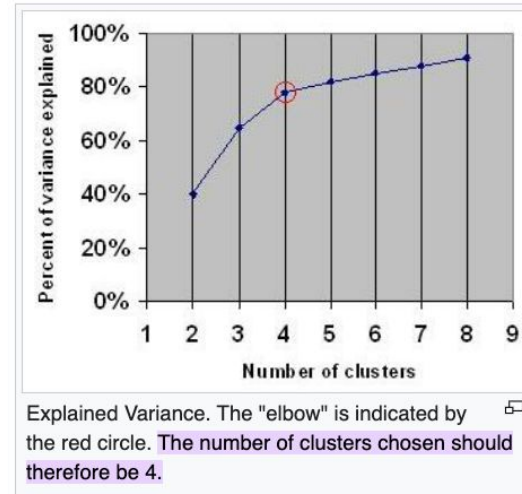
## Number of clusters

- Elbow method ([notebook presentation](#))

# Elbow method

- *Explained Variance* method can be used for number of clusters selections, and also for some other applications such as number of PCA components etc.

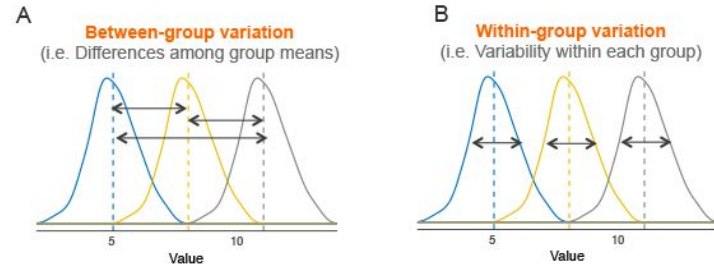
**Explained Variance:** Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an F-test.



# Elbow method

- *Explained Variance* method can be used for number of clusters selections, and also for some other applications such as number of PCA components etc.

**Explained Variance:** Percentage of variance explained is the ratio of the between-group variance to the total variance, also known as an F-test.



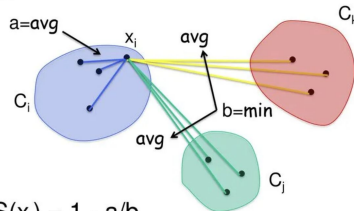
# Silhouette score based

For each sample:

a : mean intra-cluster distance

b : mean nearest-cluster distance

□ The idea...



□ Usually,  $S(x_i) = 1 - a/b$

Silhouette Score is the **mean** of the (Silhouette Coefficient for each data point)

Silhouette Coefficient:

$$SC = \frac{(b-a)}{\max(a,b)}$$

