BT2101 Tutorial 6

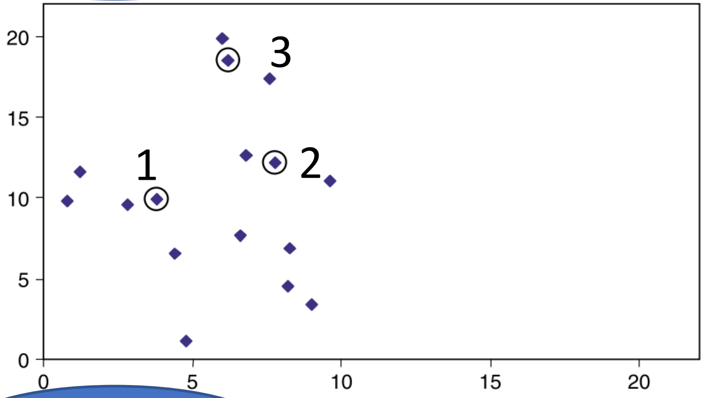
Clustering

Clustering

- K-means (e.g., k=3)

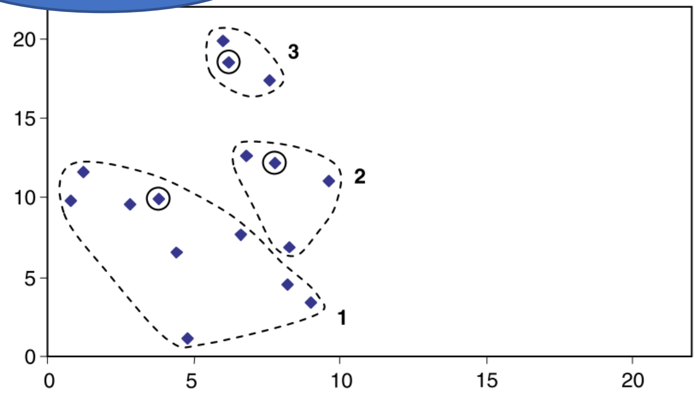
x	y
6.8	12.6
0.8	9.8
1.2	11.6
2.8	9.6
3.8	9.9
4.4	6.5
4.8	1.1
6.0	19.9
6.2	18.5
7.6	17.4
7.8	12.2
6.6	7.7
8.2	4.5
8.4	6.9
9.0	3.4
9.6	11.1

Initial Set Up



	Initial	
	x	y
Centroid 1	3.8	9.9
Centroid 2	7.8	12.2
Centroid 3	6.2	18.5

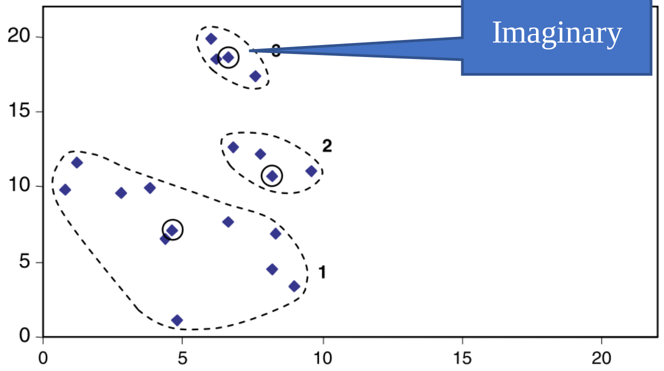
Iteration 1



	Initial		After first iteration	
	x	y	x	y
Centroid 1	3.8	9.9	4.6	7.1
Centroid 2	7.8	12.2	8.2	10.7
Centroid 3	6.2	18.5	6.6	18.6

Iteration 2…n

Repeat…until the centroids no longer move



Imaginary

	Initial		After first iteration		After second iteration	
	x	y	x	y	x	y
Centroid 1	3.8	9.9	4.6	7.1	5.0	7.1
Centroid 2	7.8	12.2	8.2	10.7	8.1	12.0
Centroid 3	6.2	18.5	6.6	18.6	6.6	18.6

Clustering

- Agglomerative Hierarchical Clustering (Bottom-Up Approach)

	a	*b*	*c*	*d*	*e*	*f*
a	0	12	6	3	25	4
b	12	0	19	8	14	15
c	6	19	0	12	5	18
d	3	8	12	0	11	9
e	25	14	5	11	0	7
f	4	15	18	9	7	0

Distance matrix

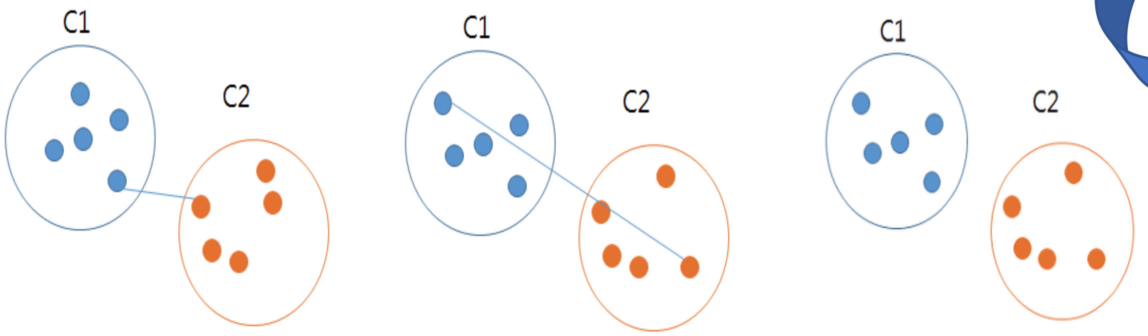
	ad	*b*	*c*	*e*	*f*
ad	0	8	6	11	4
b	8	0	19	14	15
c	6	19	0	5	18
e	11	14	5	0	7
f	4	15	18	7	0

	adf	*b*	*c*	*e*
adf	0	8	6	7
b	8	0	19	14
c	6	19	0	5
e	7	14	5	0

	adf	*b*	*ce*
adf	0	8	6
b	8	0	14
ce	6	14	0

	adfce	*b*
adfce	0	8
b	8	0

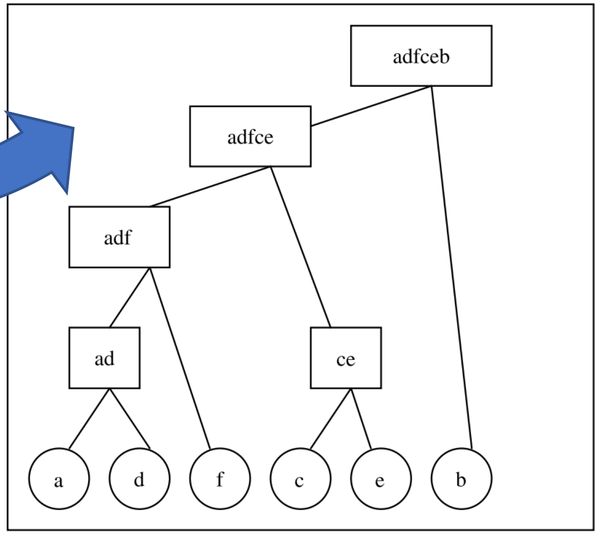
Distance Matrix
1. Single-link (min. distance)
2. Complete-link (max. distance)
3. Average-link (avg. distance)
4. Wald's (ΔΣdistance b/a join)



Single link : min d(c1, c2)

Complete link : max d(c1, c2)

Average link :
average of every distance between c1 and c2

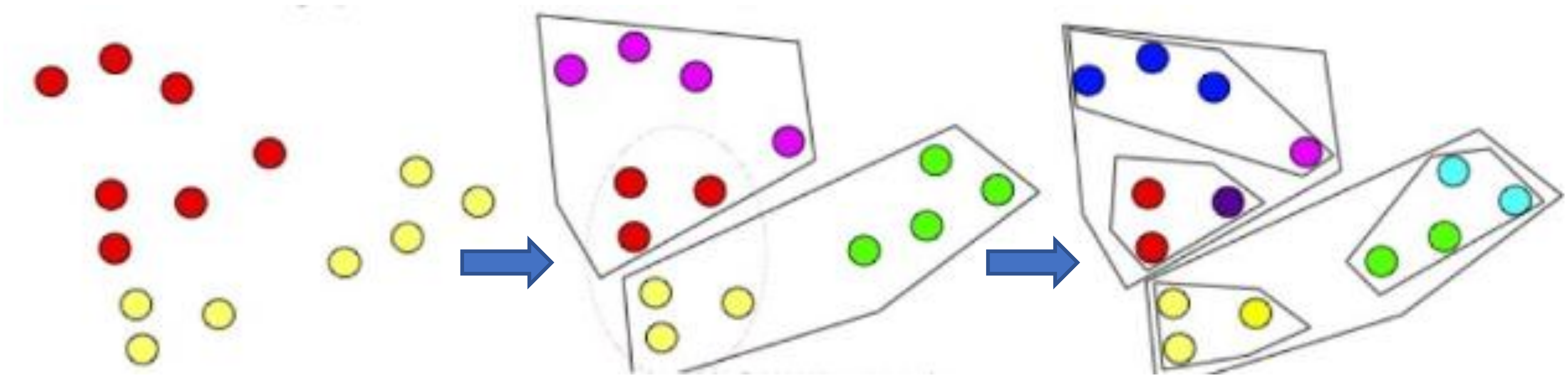


Dendrogram

References: https://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf

Clustering

- Divisive Hierarchical Clustering (Top-Down Approach)
    - K-means in each single cluster when k=2



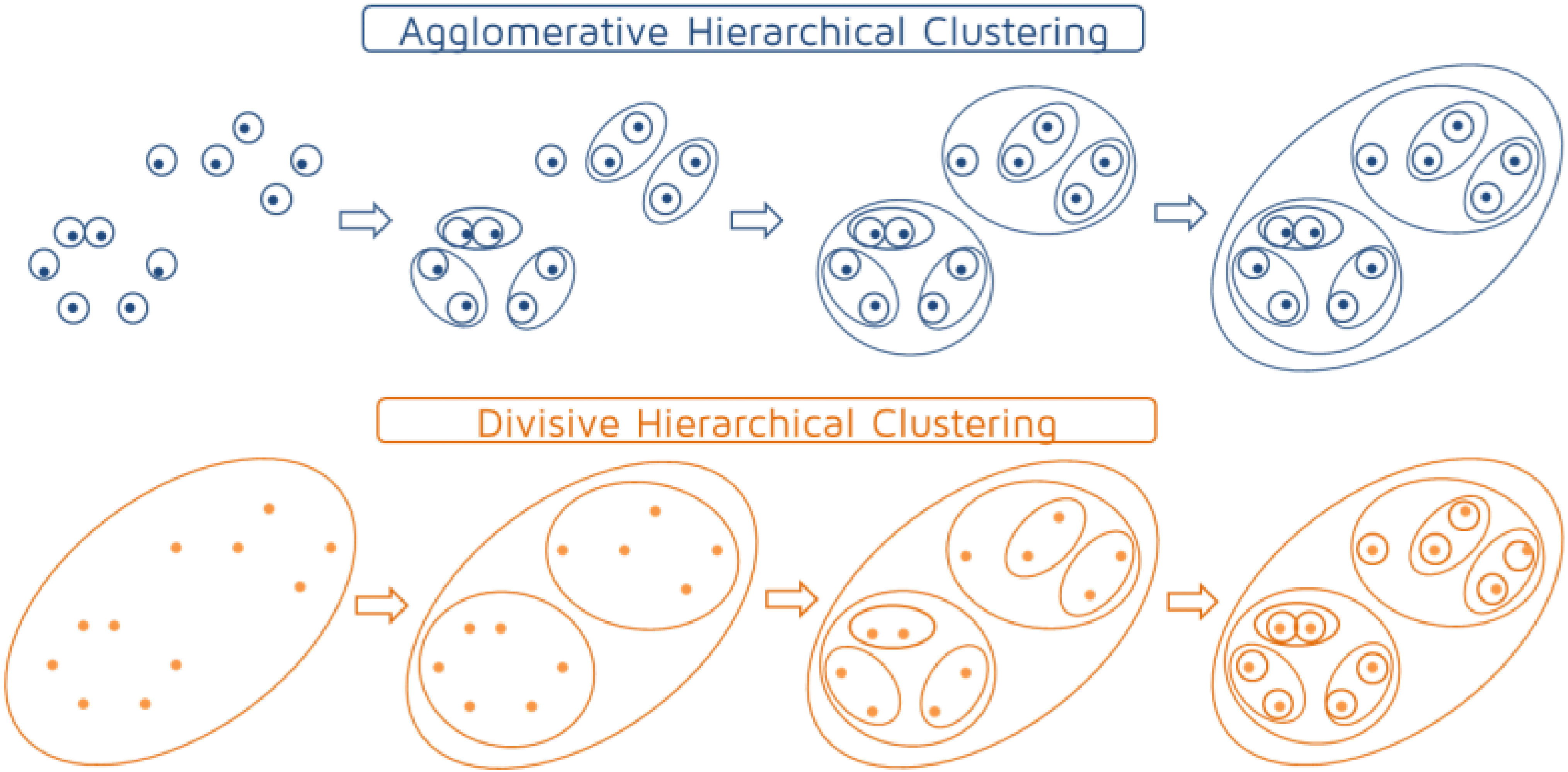
State 1

State 2

State 3

Clustering

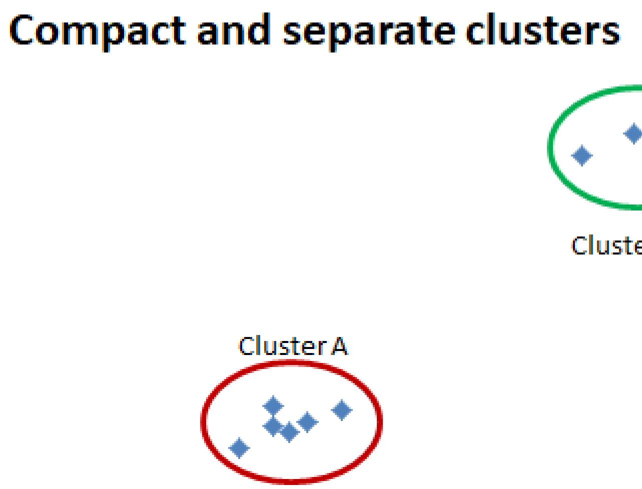
- Agglomerative Hierarchical Clustering & Divisive Hierarchical Clustering

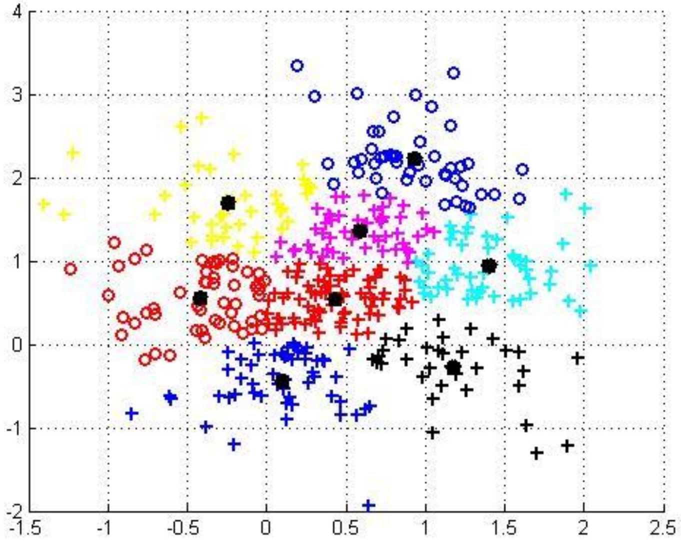


Clustering

Evaluation:

- Compactness (e.g., within-groups/clusters sum of squares)
- Seperation (e.g., group-average euclidean distance between cluster centroids)





Clustering Evaluation: Alternative Metrics

- Silhouette index
- Davies-Bouldin
- Calinski-Harabasz
- Dunn index
- R-squared index
- Hubert-Levin (C-index)
- Krzanowski-Lai index
- Hartigan index
- Root-mean-square standard deviation (RMSSTD) index
- Semi-partial R-squared (SPR) index
- Distance between two clusters (CD) index
- weighted inter-intra index
- Homogeneity index
- Separation index

Data Standardization

- When do we need **Data Standardization/Scaling**?
  - Balancing the dimensions
  - Easy to calculate distance

Example: Person=(age, marathon distance)
A. (22, 10000m)
B. (22, 20000m)
C. (80, 5000m)

Question: Based on your reasonable intuition, who is more similar to A? B or C?

Data Standardization

- **Decimal scaling** $$x'_{ij} = \frac{x_{ij}}{10^h},$$

- **Min-max** $$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- ***z*-index**

$$x'_{ij} = \frac{x_{ij} - \bar{\mu}_j}{\bar{\sigma}_j},$$

Clustering

How to choose k?

Choose k based on the **how results will be used**

- e.g., "How many market segments do we want?"

Also experiment with slightly different k's

- Initial partition into clusters can be random, or based on **domain knowledge**
- If random partition, repeat the process with different random partitions

Elbow Method

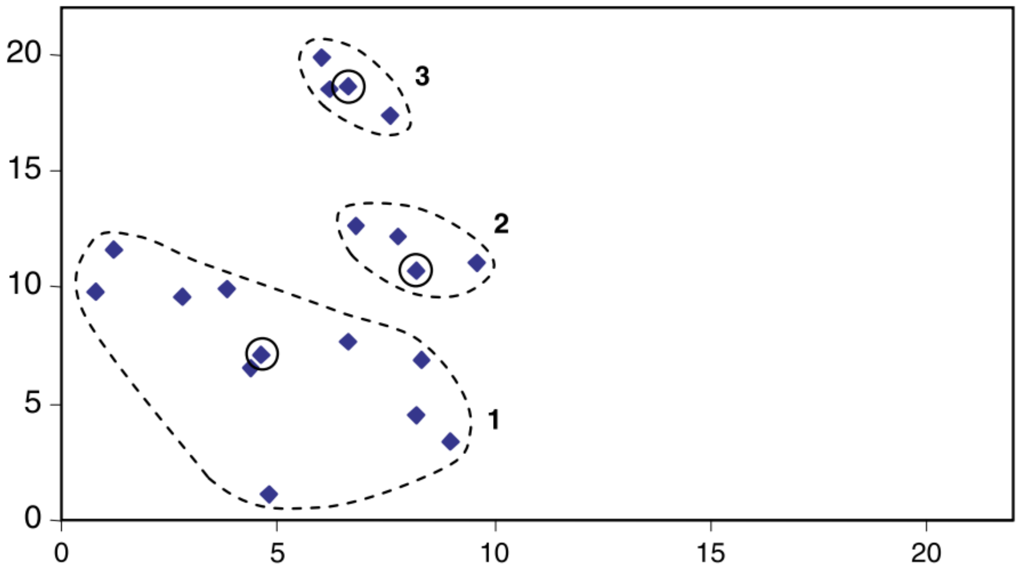
- (Average) Within-groups/clusters sum of squares (WSS)

Clustering

The within-groups/clusters sum of squares (WSS):

$$WSS(k) = \sum_{i=1}^{n} \sum_{j=0}^{p} (x_{ij} - mean(x_{kj}))^2$$

where, k is the cluster, x_{ij} is the value of the j^{th} variable for the i^{th} observation, and $mean(x_{kj})$ is the mean of the j^{th} variable for the k^{th} cluster.



Clustering

How to choose k?

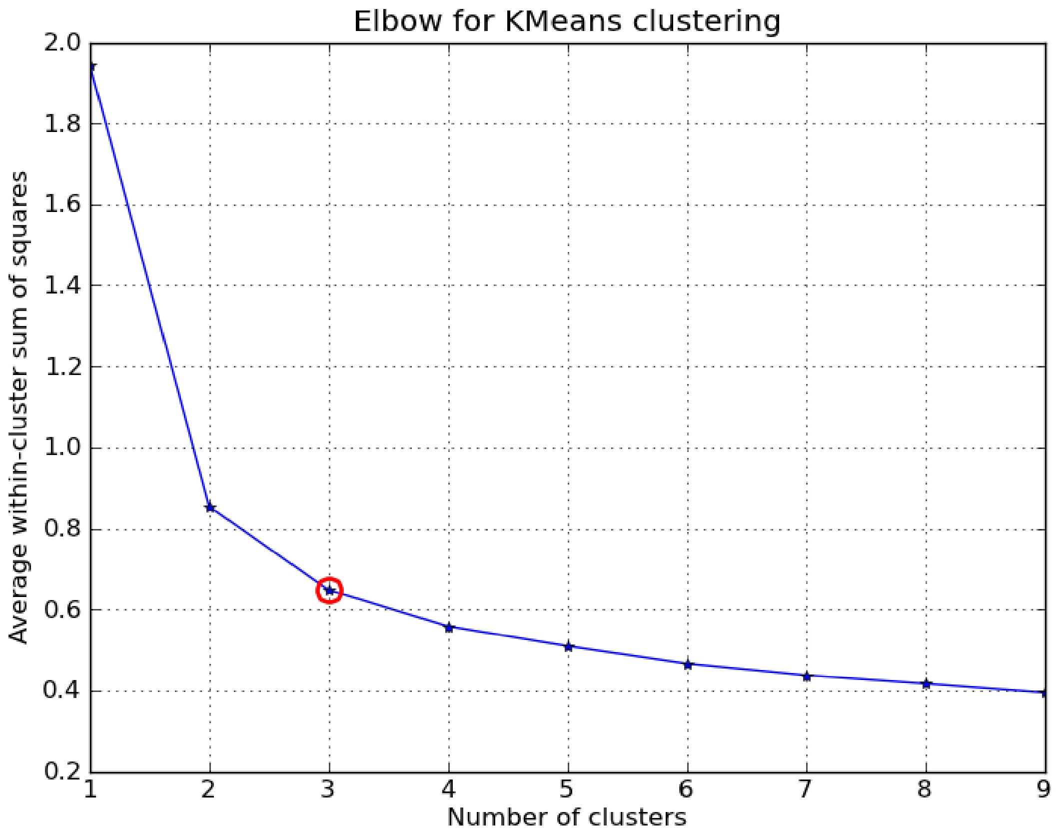
Elbow method

– Gauge how the heterogeneity within clusters changes for various of k.

• The heterogeneity within clusters is expected to **decreases** with more clusters.

• The heterogeneity is measured by within-clusters/groups sum of squares (WSS)

• Is this a measure of compactness or separation?



Thank you!