Quiz 9: Clustering – 2 (10 points), **15 minutes**
TA handling this quiz: Abhishek Bhatt abhishpb@usc.edu

1. [7 points] Suppose k* = 5, i.e., the best value of k, for running k-means on a data set. Describe the steps that the k* algorithm takes to discover this value. Be sure to indicate the value of k for each k-means clustering that the algorithm performs.

First, it finds the elbow of the curve:

Runs k-means for k = 1, 2, 4, 8

Stop at k = 8 as there must be little change in cohesion from k = 4 to 8

Than it uses binary search in range [4,8]

z = (4+8)/2

z = 6

As there must be little change in cohesion between [6,8] we search for z* in [4,6]

Z = (4+6)/2

Z = 5

k* = 5

2. [3 points] Explain the commonality and key differences between the BFR and CURE algorithms.

Common: [1 point]

Both BFR and CURE assume data to be in euclidean space.

Differences [2 point]

BFR	CURE
Data has to be normally distributed.	Do not assume particular distribution.
Cluster is in form of centroid.	Cluster is in form of set of representatives.