

# Clustering

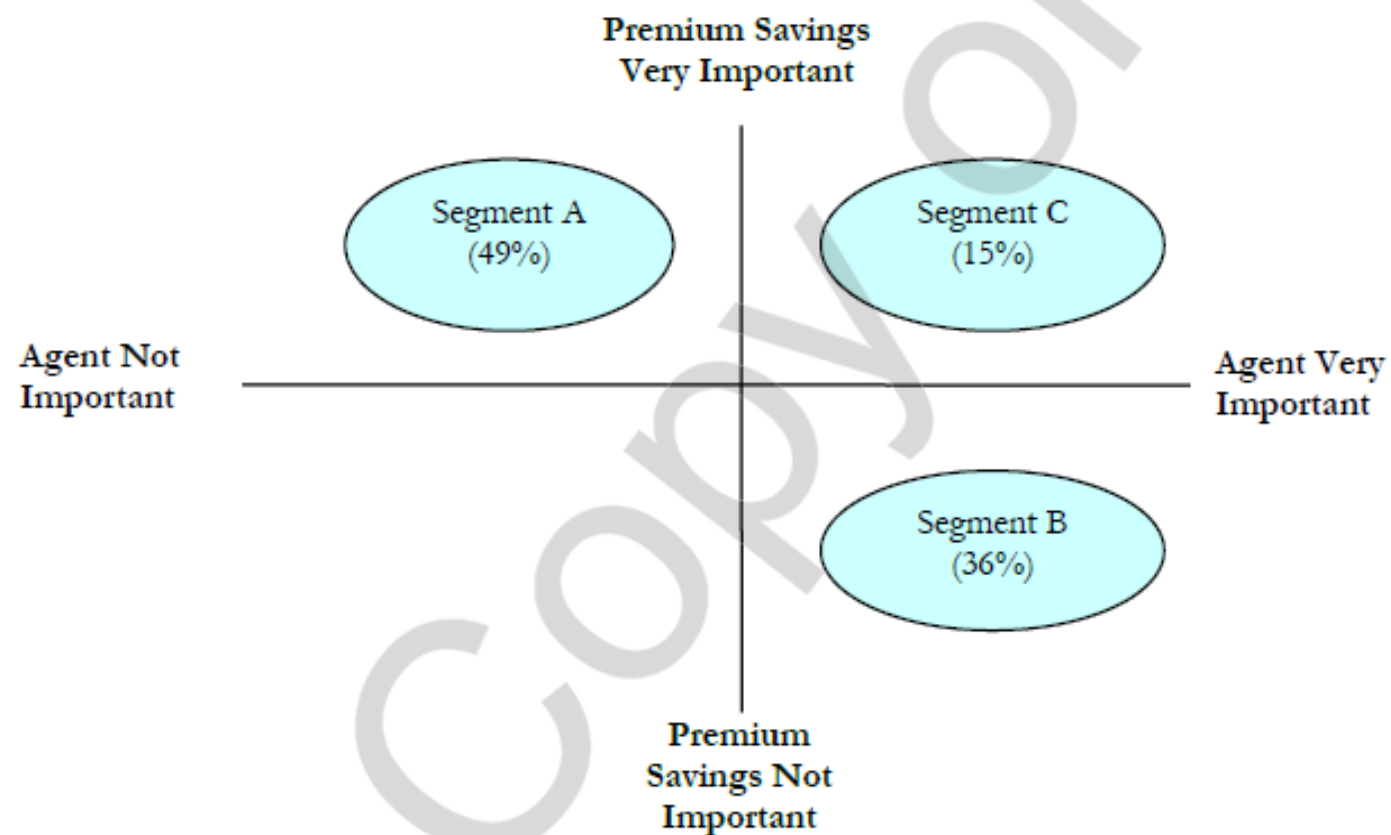
# What is clustering?

- A statistical method to identify groups/clusters/segments of customers based on differences in some variables
- Variables: (1) demographic, (2) psychographics, (3) desired benefits from offerings, and (4) past-purchase and past-use behaviors
- Cluster = a group of relatively homogeneous/similar members

# Example

An auto insurance company asked its customers two questions (on 1 to 7 Likert scale; 1 = Not important, 7 = Very important)

1. How important is savings on premium for you?
2. How importance is the existence of a neighborhood agent for you?



# Steps

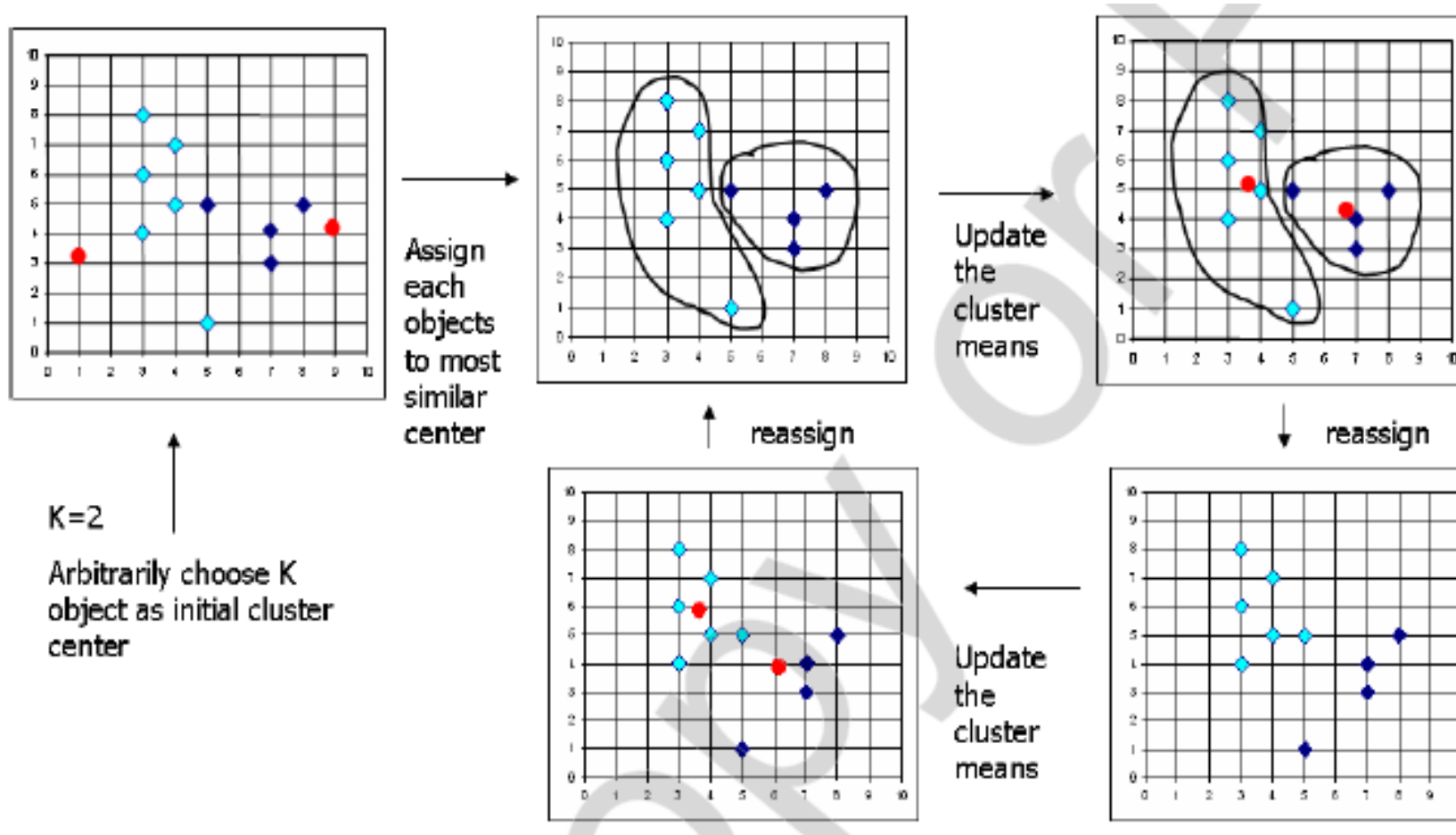
1. Select variables you will use for clustering
2. Decide #clusters (if K-means clustering), say, 2
3. Compute distance between each pair of customers on each variable.  
For example, Euclidean distance
4. Apply clustering “method” on the distances. Do the clusters look distinct? If no, change K.
5. Profile each cluster in terms of demographic vars

# Distance measure

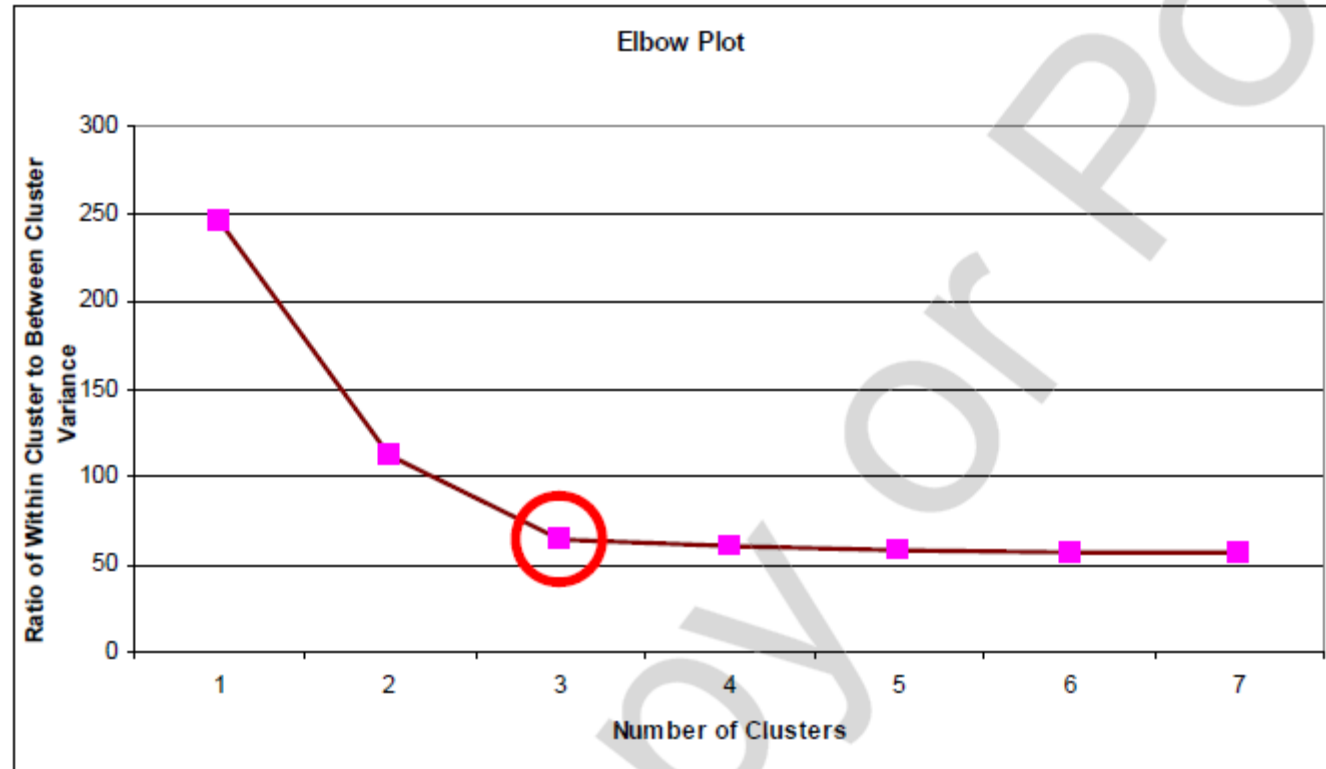
Individual Name	Importance Score	
	Premium Savings	Neighborhood Agent
Joe	4	7
Sam	3	4
Sara	5	3

$$\text{Euclidean distance (Joe, Sam)} = \sqrt{(4-3)^2 + (7-4)^2} = 3.2.$$

# K-Means Clustering



# Elbow plot to determine K





# Profiling

Segment	Mean		Range	
	Age	Income (\$)	Age	Income (\$)
A	21	15,000	16–25	0–25,000
B	45	120,000	33–55	75,000–215,000
C	39	40,000	39–54	24,000–60,000