K-Means Clustering

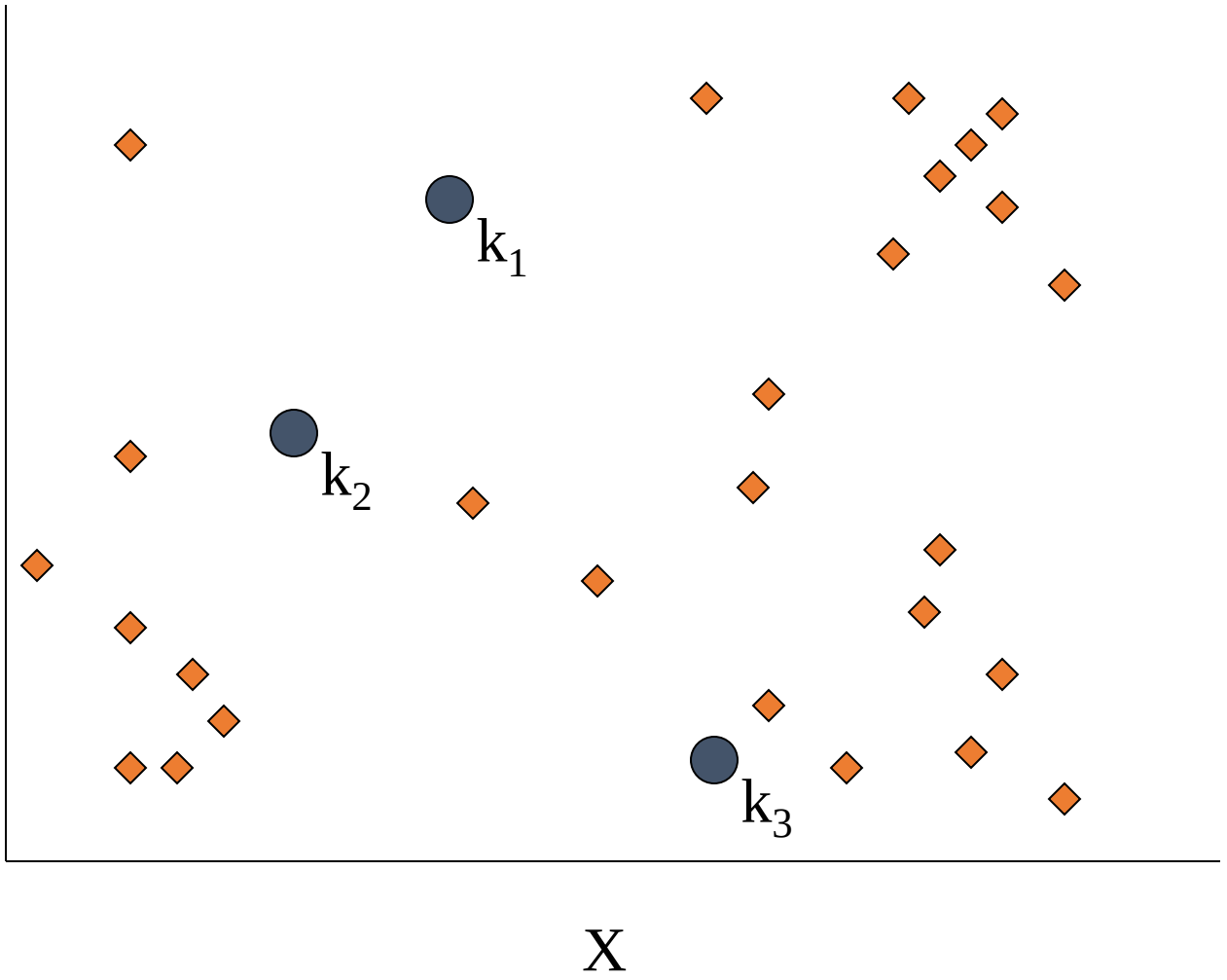
Dr. Muhammad Wasim

K-means Clustering

- Works with numeric data only
  1. Pick a number (K) of cluster centers (at random)
  2. Assign every item to its nearest cluster center (e.g. using Euclidean distance)
  3. Move each cluster center to the mean of its assigned items
  4. Repeat steps 2,3 until convergence (change in cluster assignments less than a threshold)

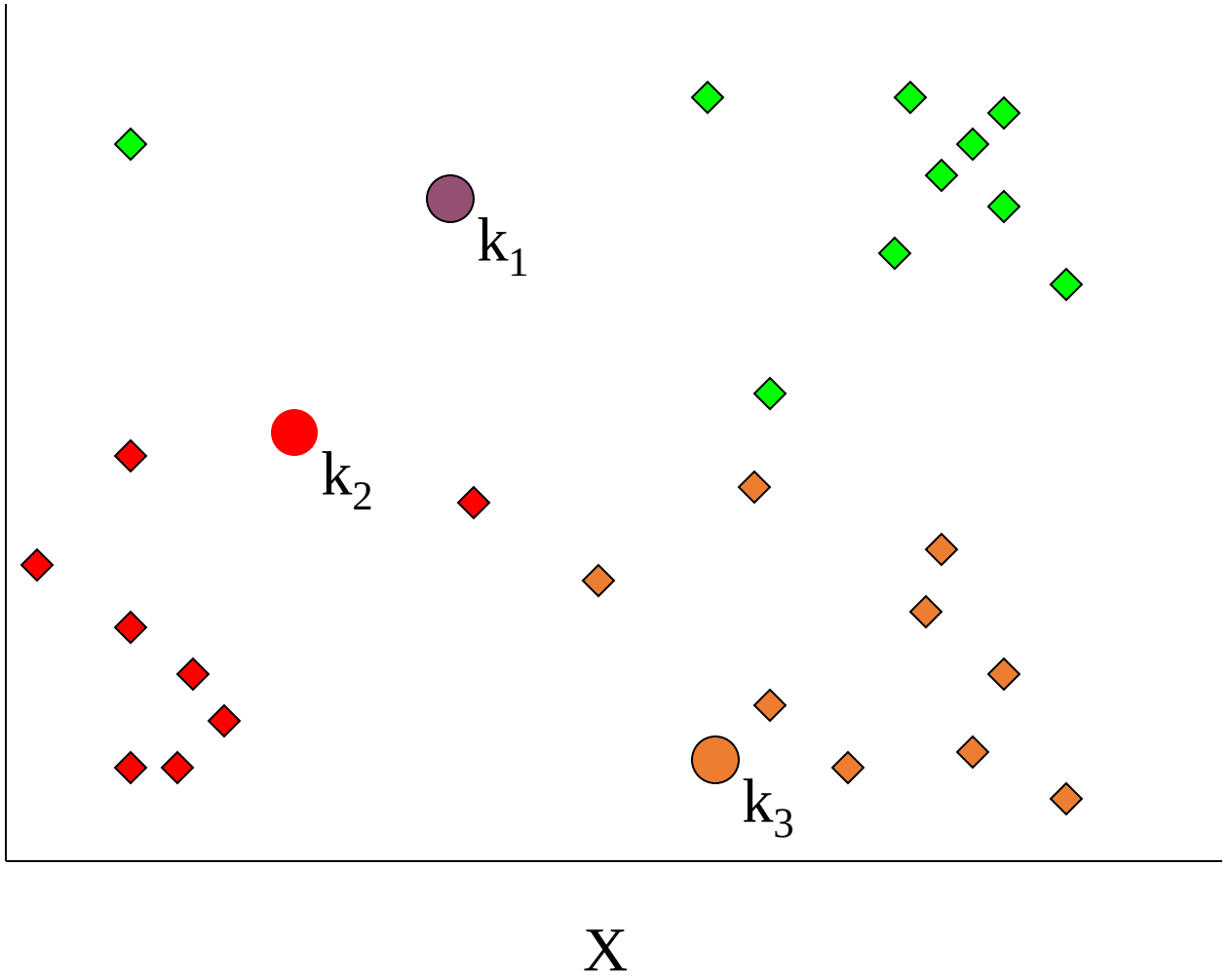
K-means example, step 1

Pick 3 initial cluster centers (randomly)



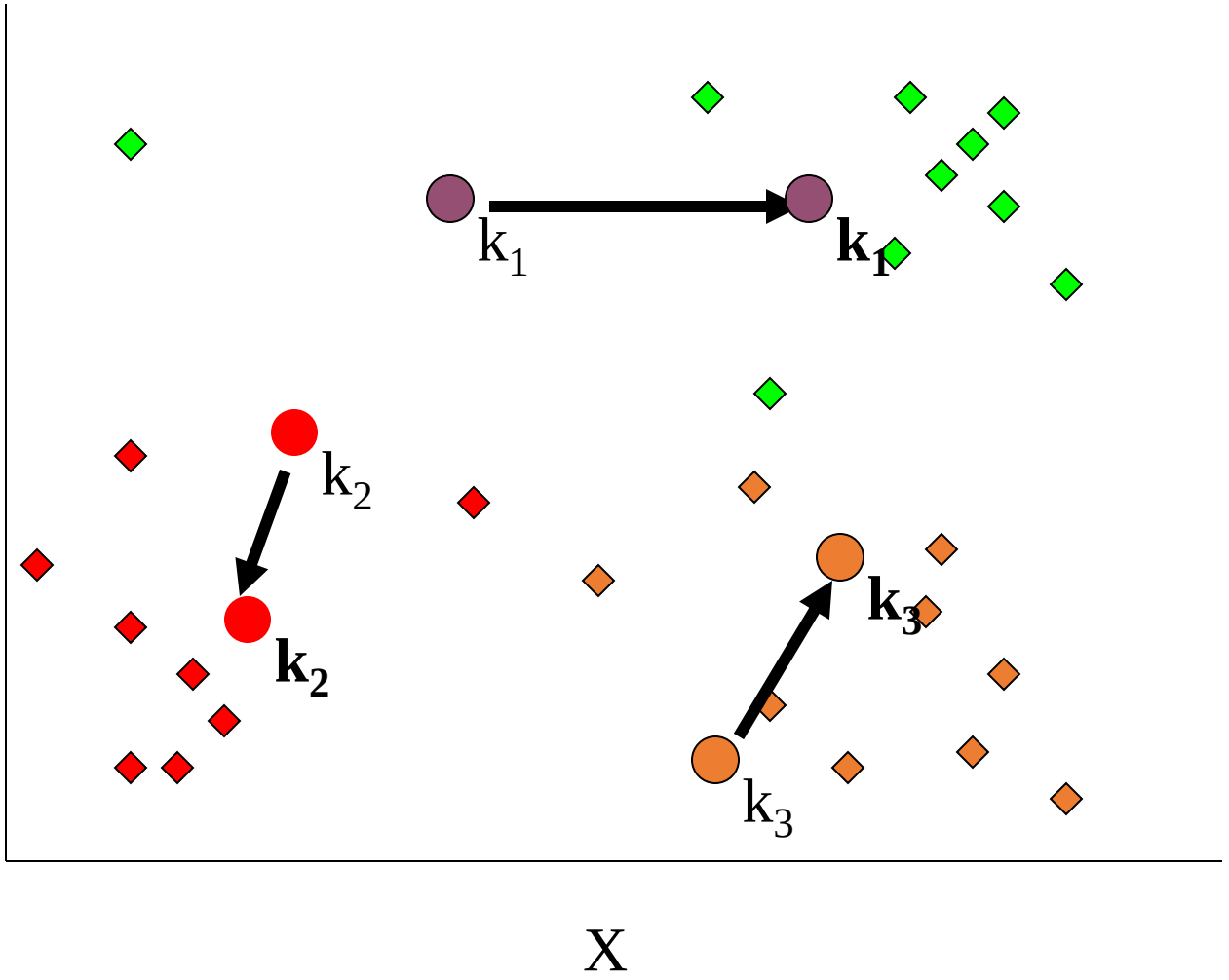
K-means example, step 2

Assign each point to the closest cluster center



K-means example, step 3

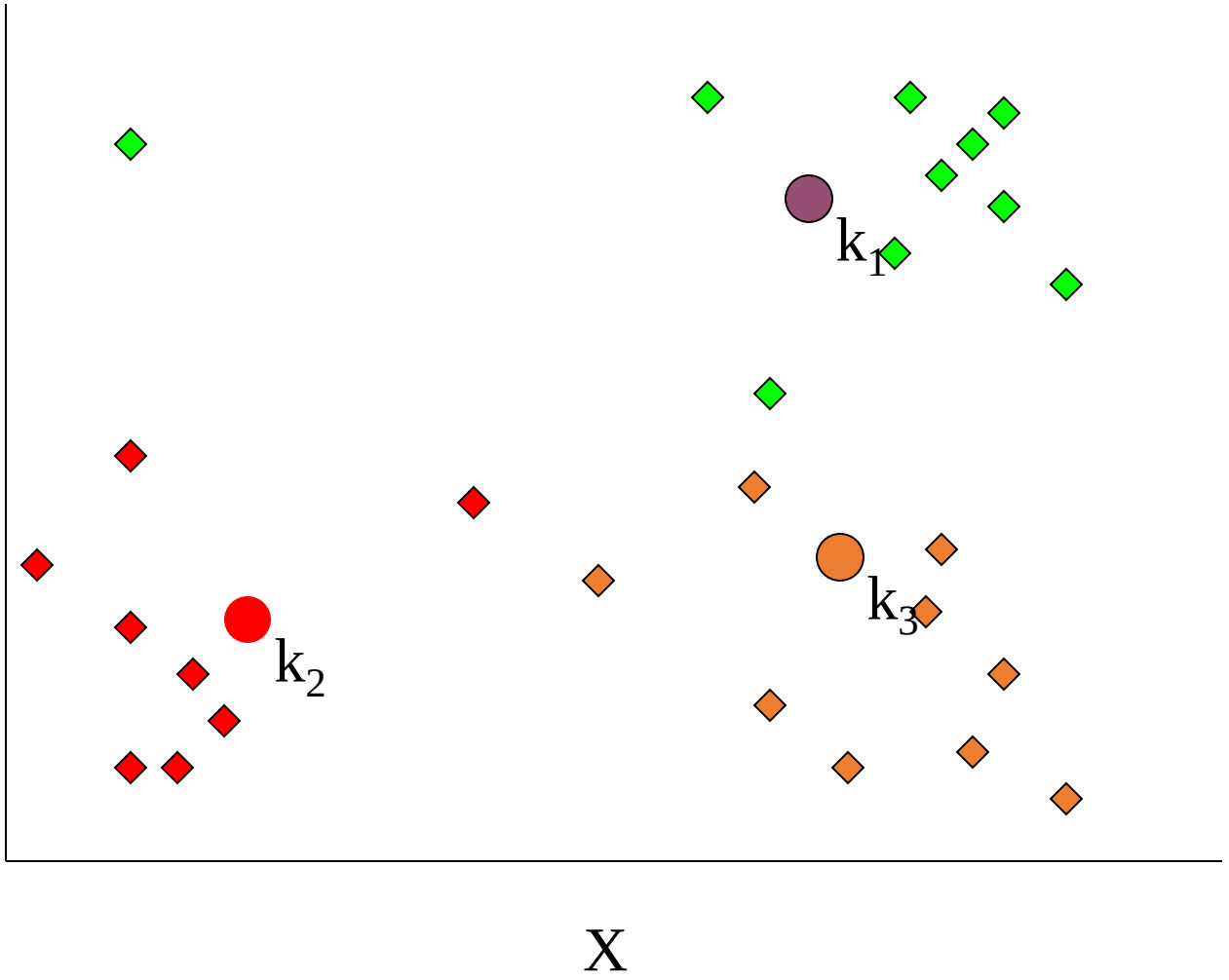
Move
each cluster
center
to the mean
of each cluster



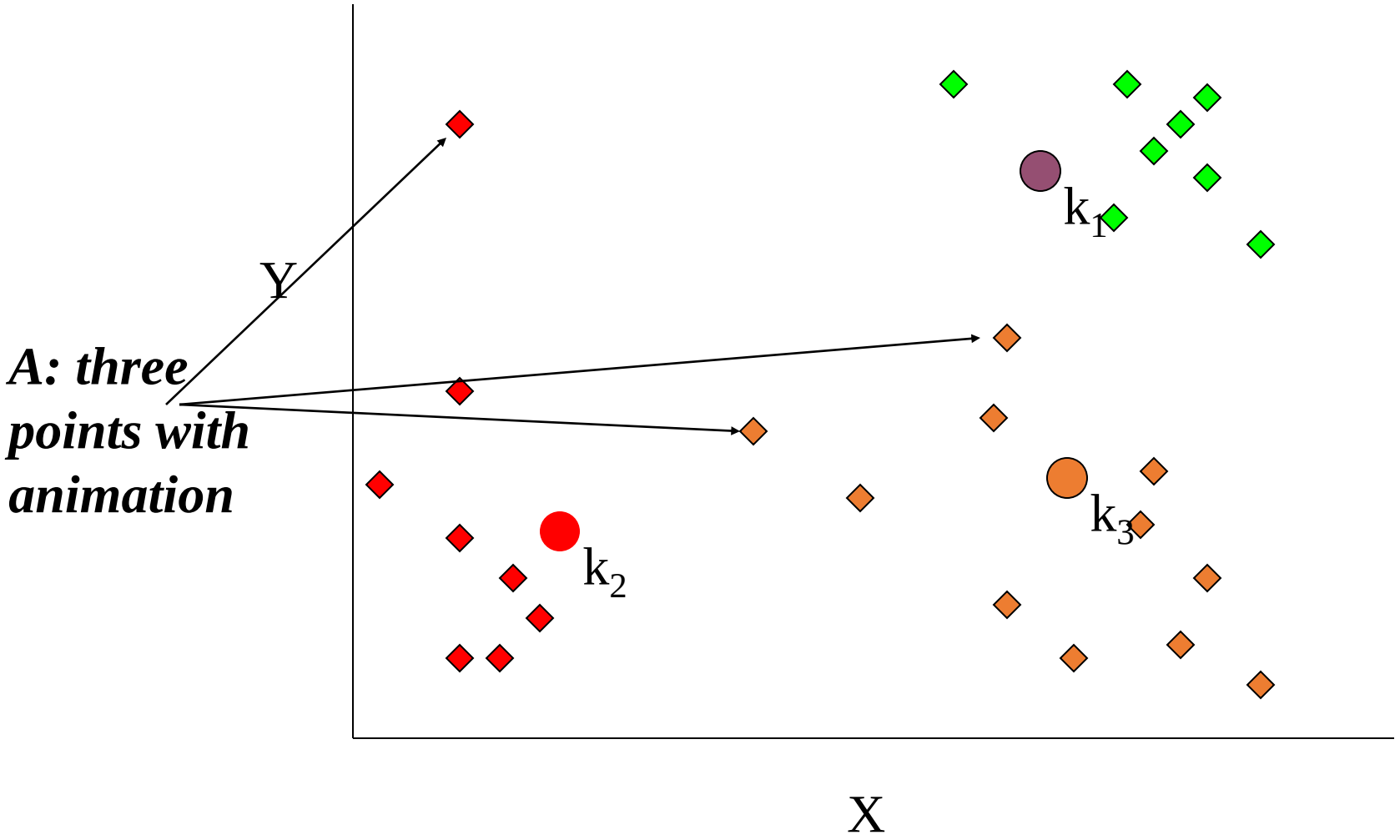
K-means example, step 4

Reassign points closest to a different new cluster center

Q: Which points are reassigned?

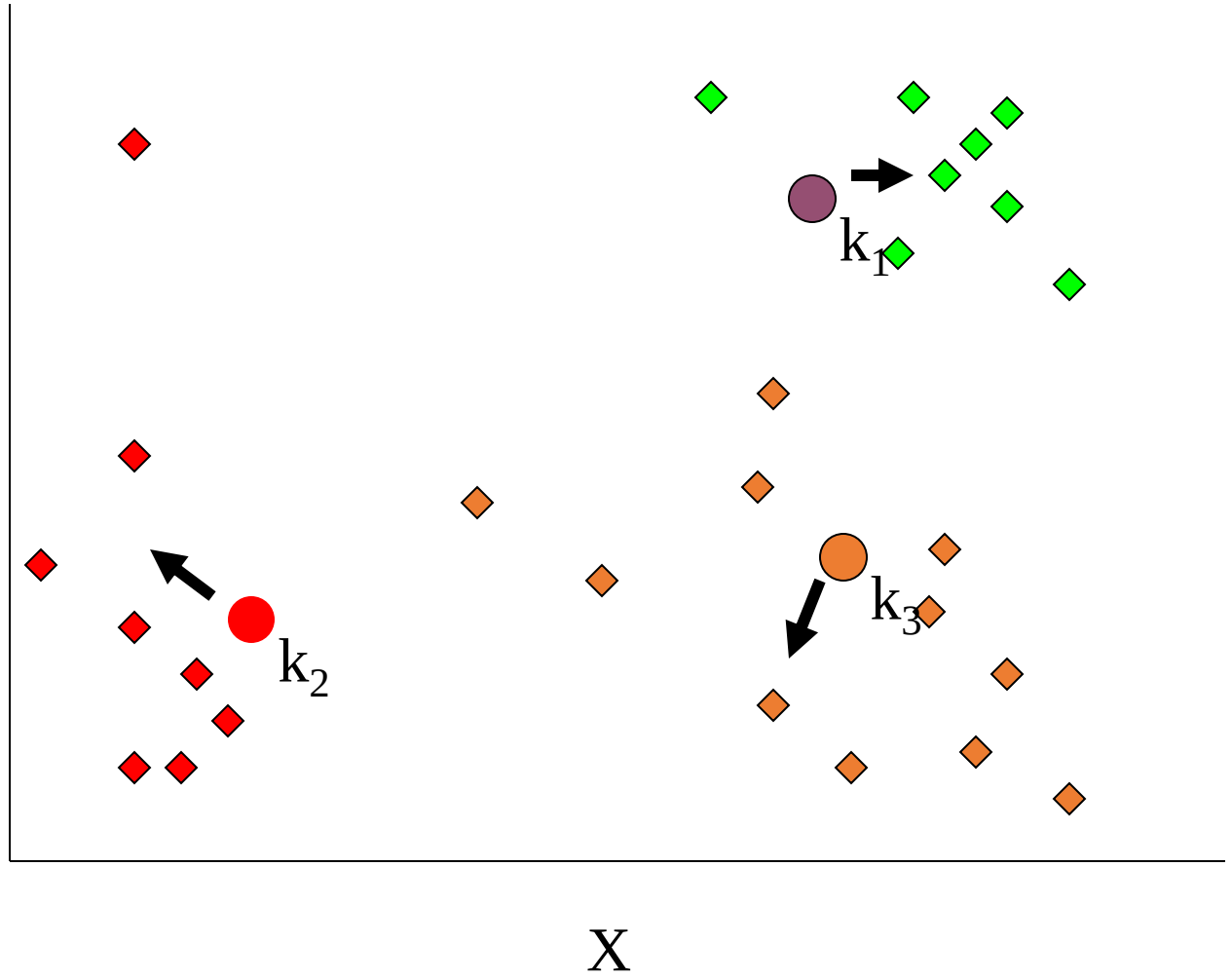


K-means example, step 4 …

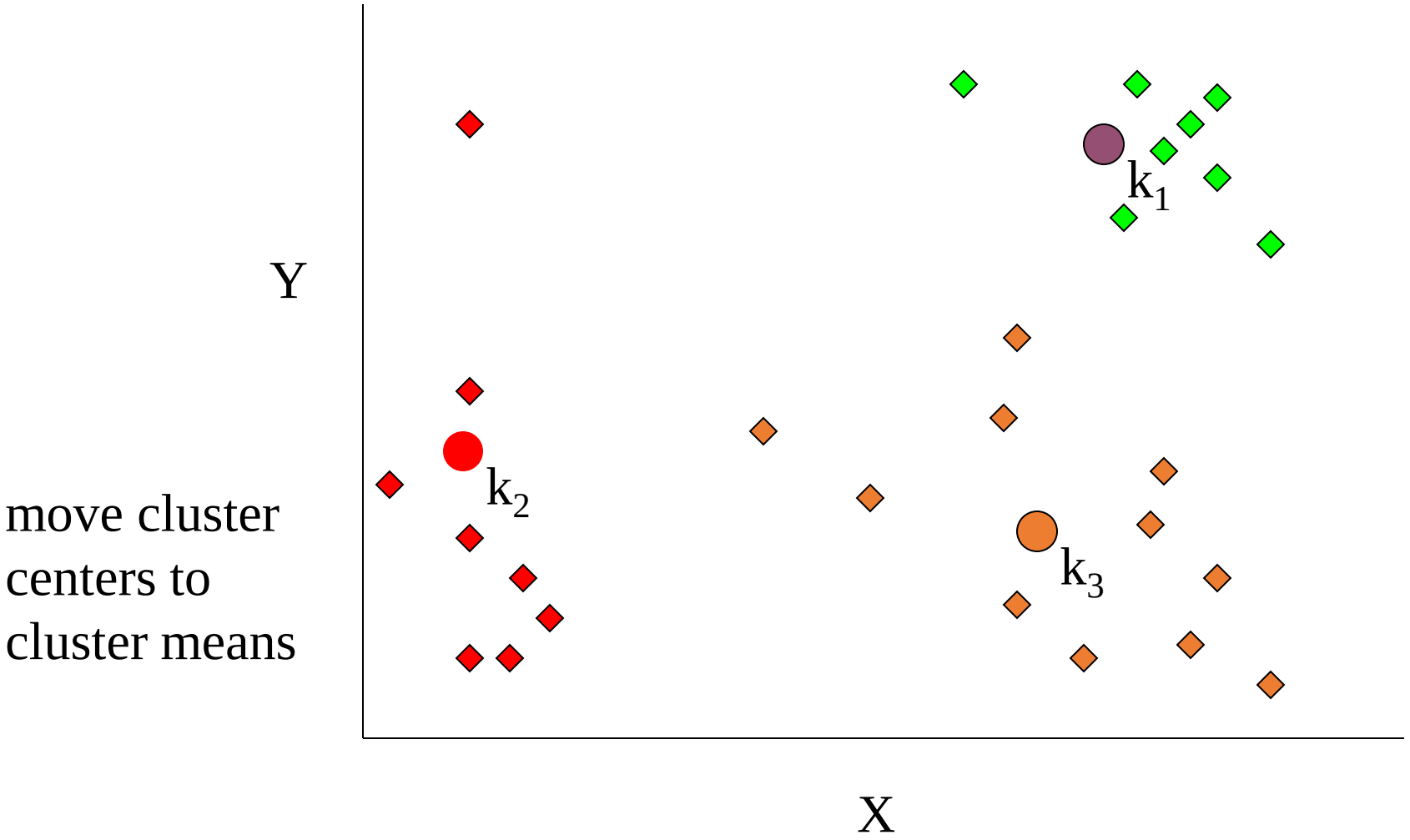


K-means example, step 4b

re-compute cluster means



K-means example, step 5



Pros and Cons of k-means clustering

- Advantages: Scalability
  - K-means is fast. In K-means only the distance between points and cluster center is calculated. It has a linear complexity $O(n)$
- Disadvantage: Flexibility
  - The number of clusters must be predetermined.
  - Using K-means with mixed variables (categorical, continuous) isn't trivial.
  - K-means is not reproducible.