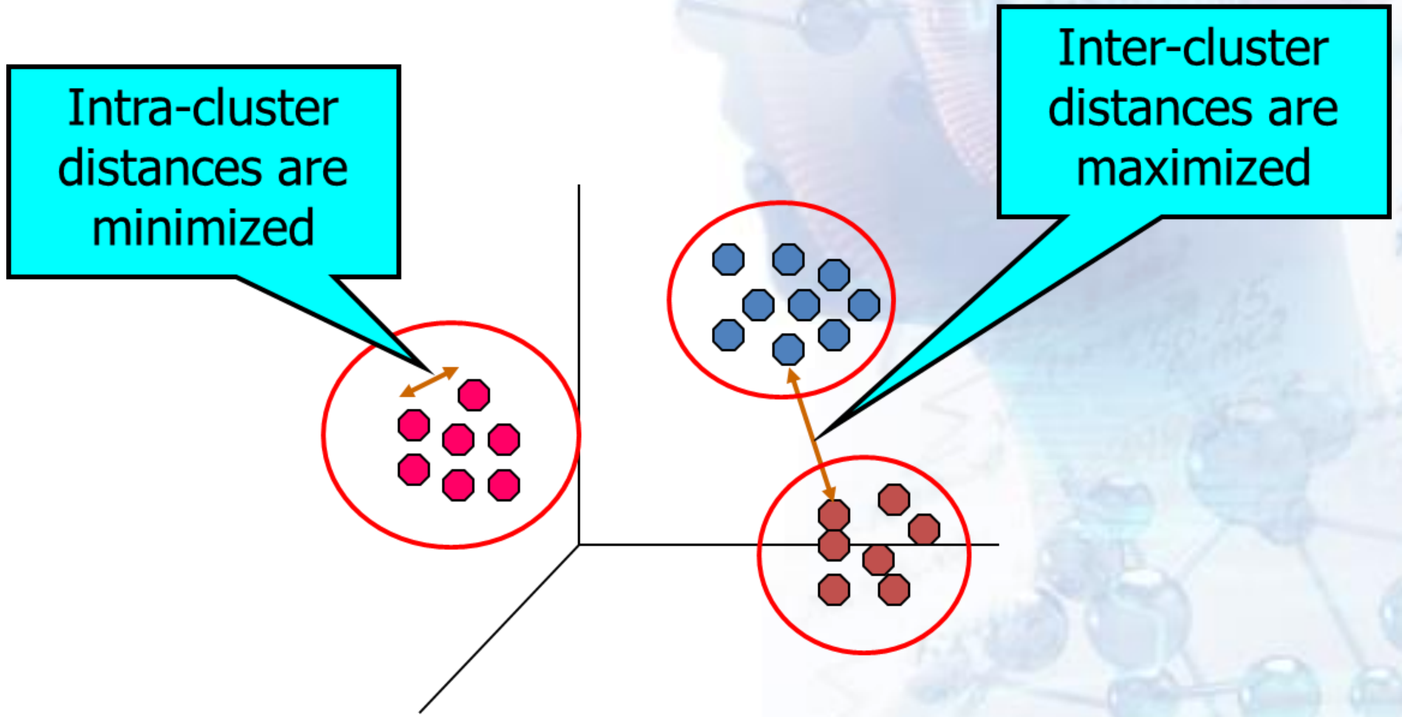
CSC-411 Artificial Intelligence

Introduction to Machine Learning

Clustering

Clustering

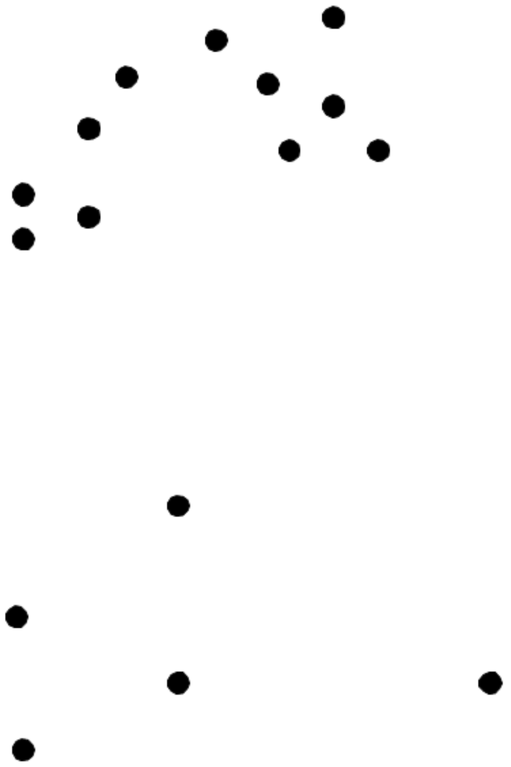
- In general a grouping of objects such that the objects in a group (cluster) are similar (or related) to one another and different from (or unrelated to) the objects in other groups



Types of Clustering

- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
- Partitional Clustering
  - A division data objects into subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree

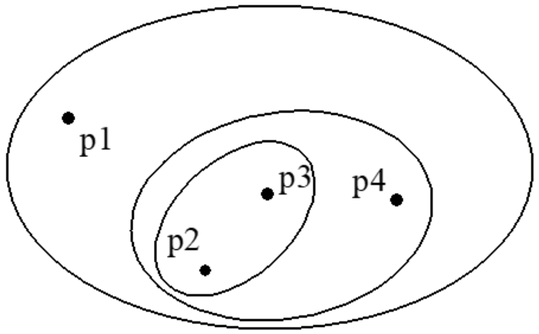
Partitional Clustering



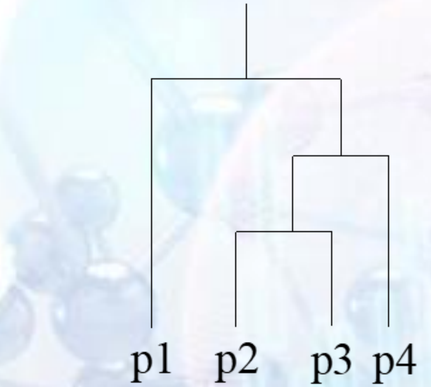
Original Points

A Partitional Clustering

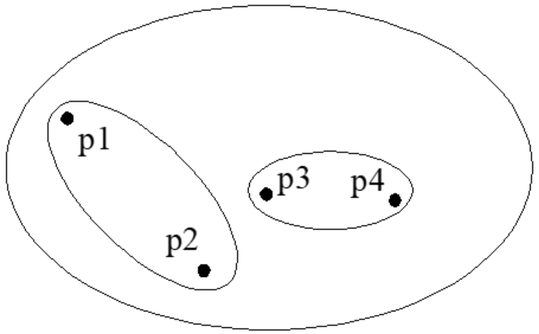
Hierarchical Clustering



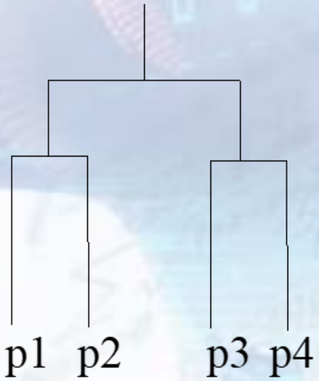
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

K-means Clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The objective is find K centroids and the assignment of points to clusters/centroids so as to minimize the sum of distances of the points to their respective centroid

K-means Algorithm

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-means Calculations

- To find the distance of a data point from the centroid we can use any mathematical distance formula like Manhattan distance or Euclidean distance etc.
- Example: Distance of point A (2,5) from a centroid K1 (6, 8) calculated using Manhattan distance would be:

X1 = 2, X2 = 6

Y1 = 5, Y2 = 8

Manhattan Distance X = |2-6| = |-4| = 4

Manhattan Distance Y = |5-8| = |-3| = 3

Distance of point A from K1 = 4 + 3 = 7

- Use the same formula to calculate the distances of “each” point from “every” centroid.

K-means Calculations

- To find new centroid, find the average of all X values and all Y values to generate a new centroid X_{New}, Y_{New}

- Example: if we have 3 data points in a cluster A, B, C, then the new Centroid will be calculated as:

$$\frac{X_A + X_B + X_C}{3} = X_{new}$$

$$\frac{Y_A + Y_B + Y_C}{3} = Y_{New}$$

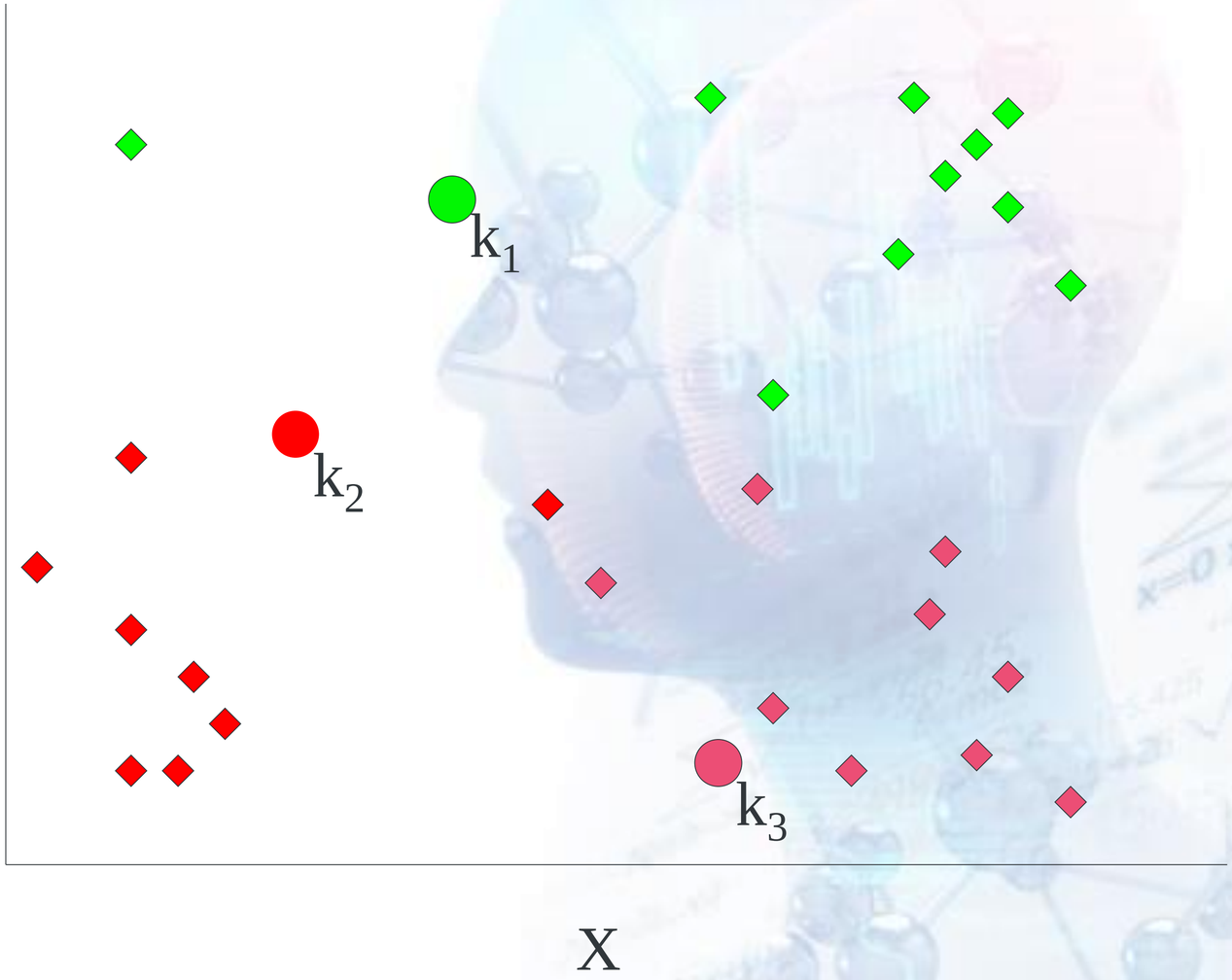
- Do this to calculate the centroids for all K clusters.

K-Means Example, Step 1

Pick 3 initial cluster centers (randomly)

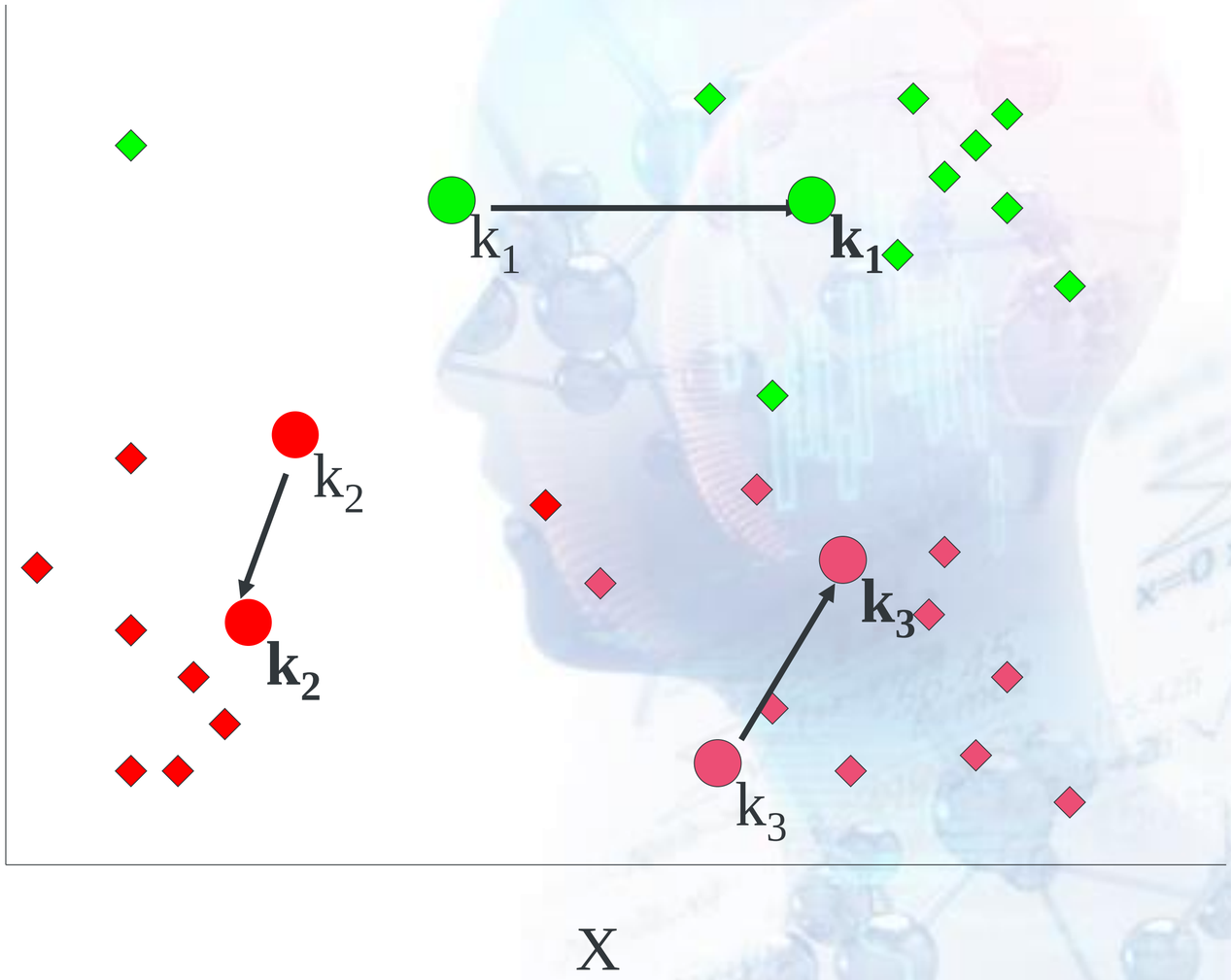
K-Means Example, Step 2

Assign each point to the closest cluster center



K-Means Example, Step 3

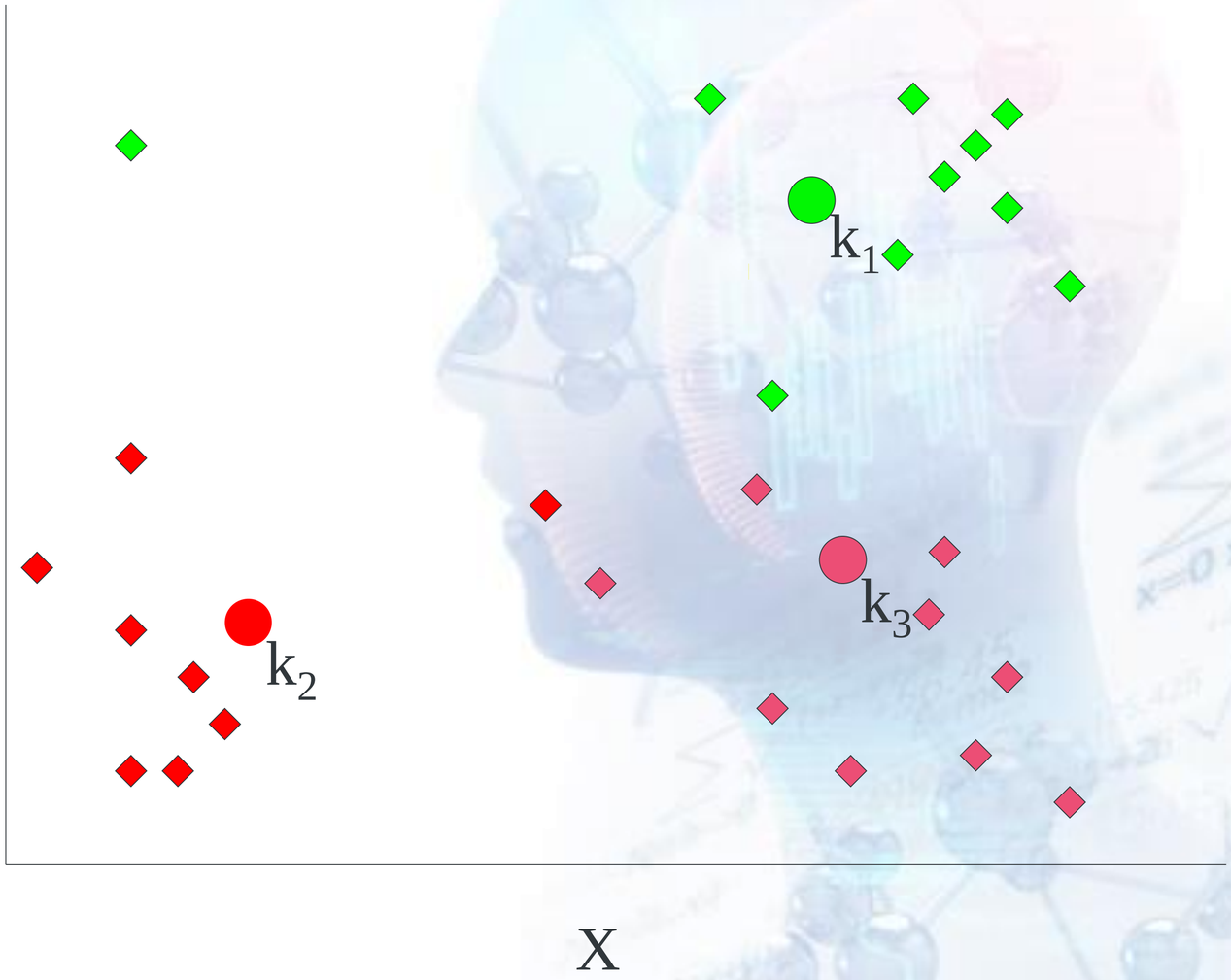
Move
each cluster
center
to the mean
of each cluster



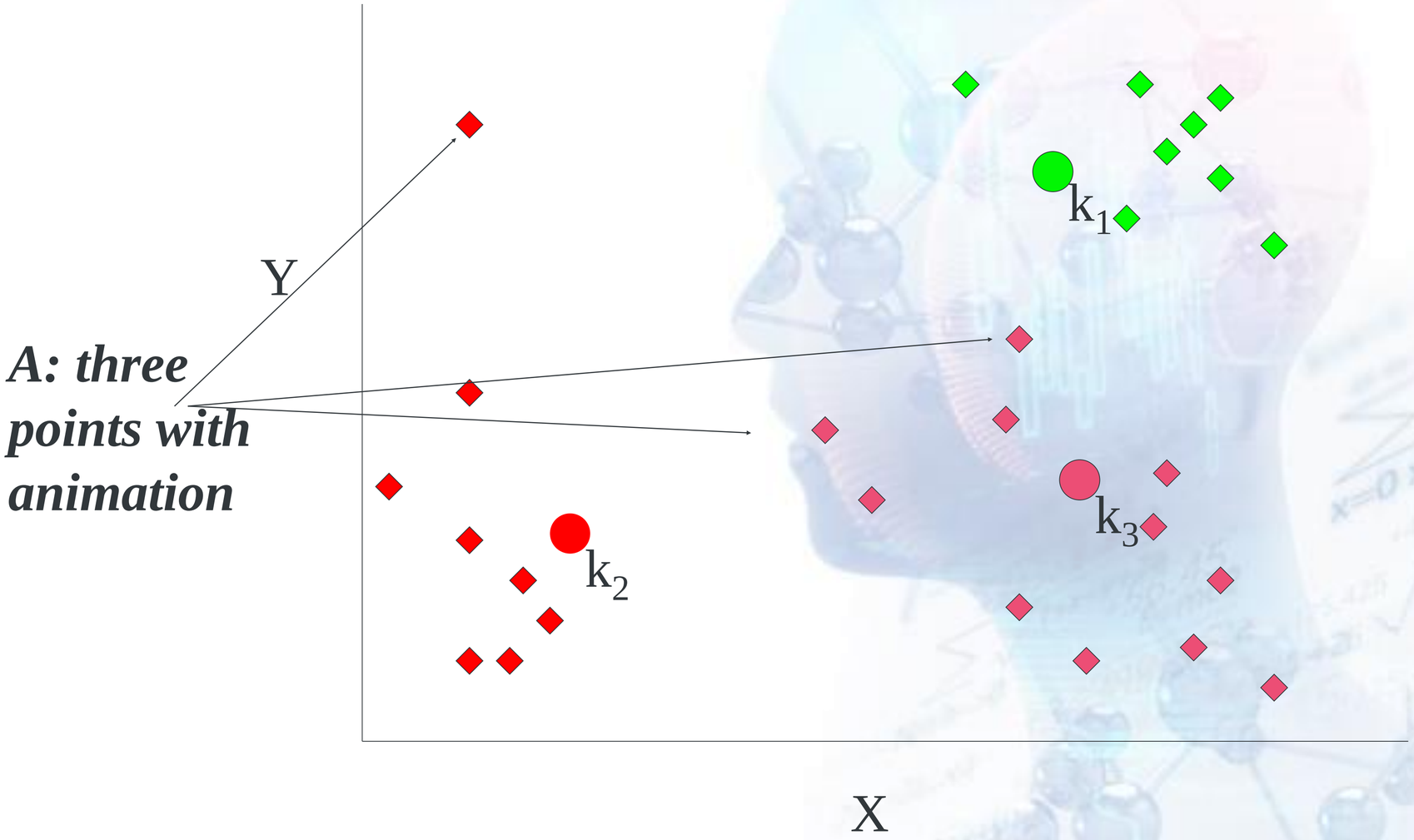
K-Means Example, Step 4

Reassign points closest to a different new cluster center

Q: Which points are reassigned?



K-Means Example, Step 4 …

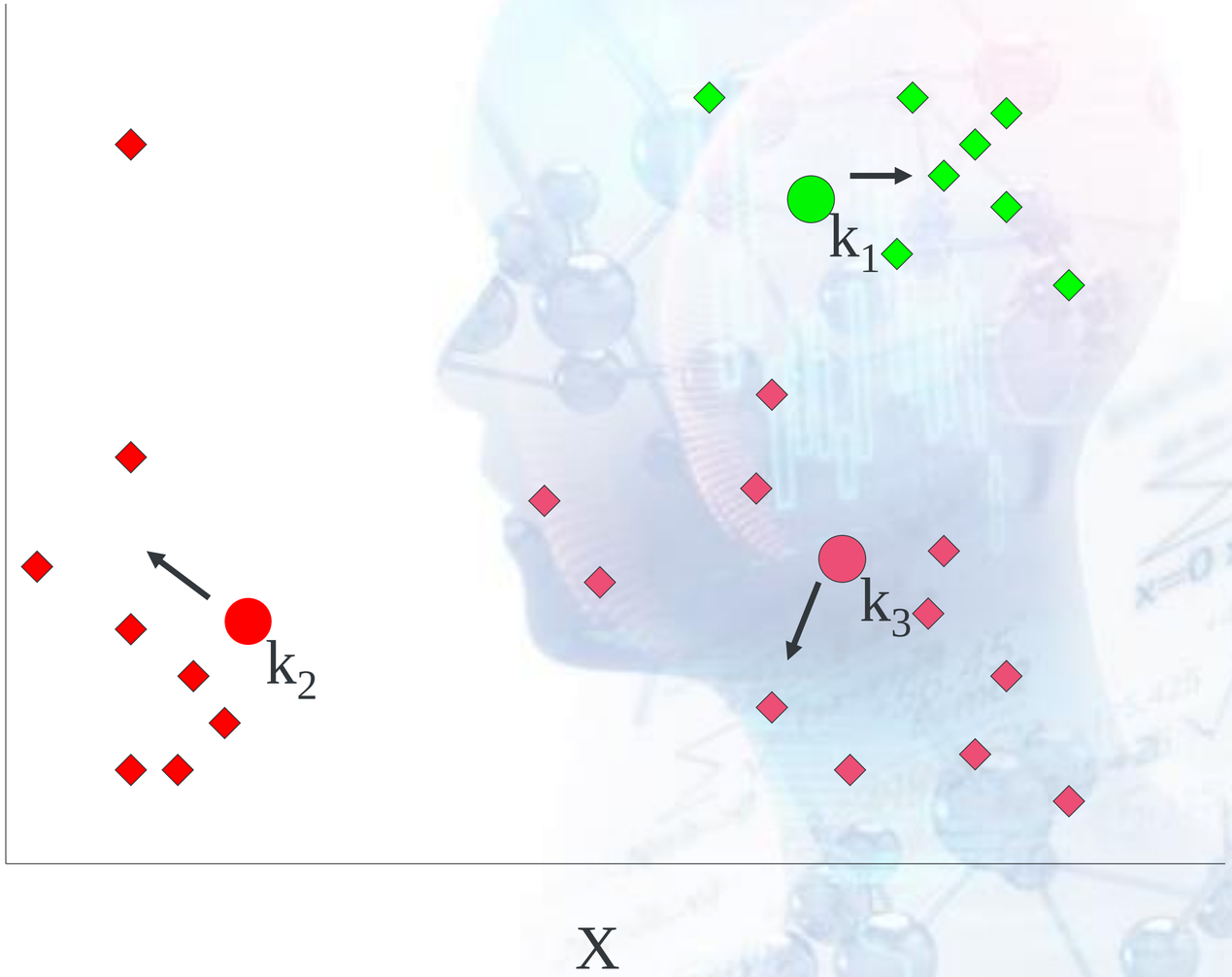


A: three points with animation

K-Means Example, Step 4b

Y

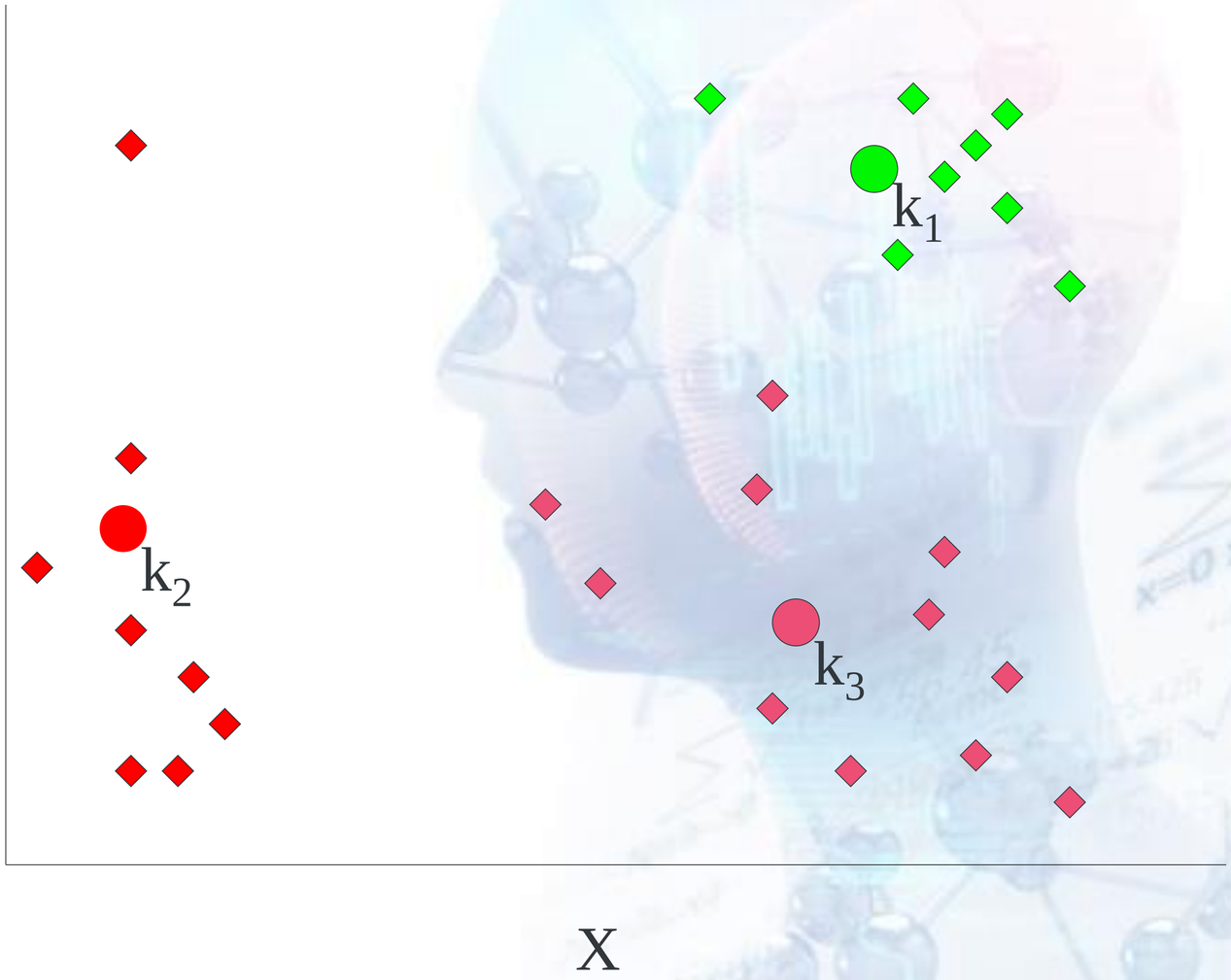
re-compute
cluster means



X

K-Means Example, Step 5

move cluster centers to cluster means



Limitations of K-means

- K-means has problems when clusters are of different:
  - sizes
  - densities
  - non-globular shapes

- K-means has problems when the data contains outliers.