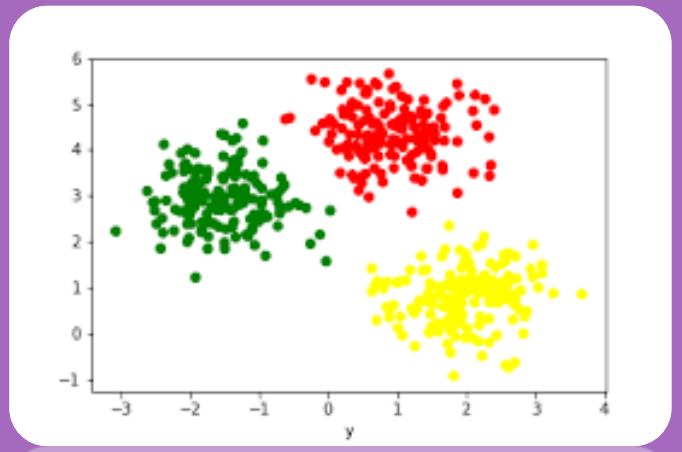
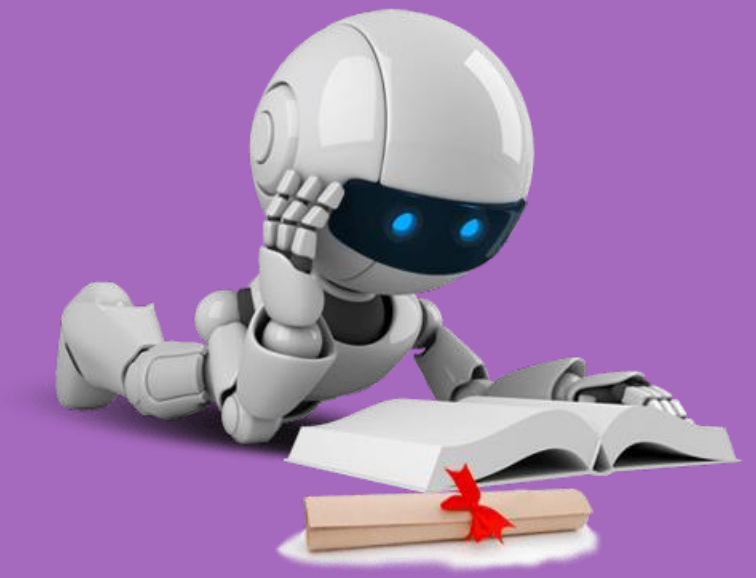
دوره‌ی آموزشی «علم داده»

Data Science Course

جلسه بیست و دوم - (بخش اول)

خوشه‌بندی به روش K-Means





مدرس: محمد فزونی

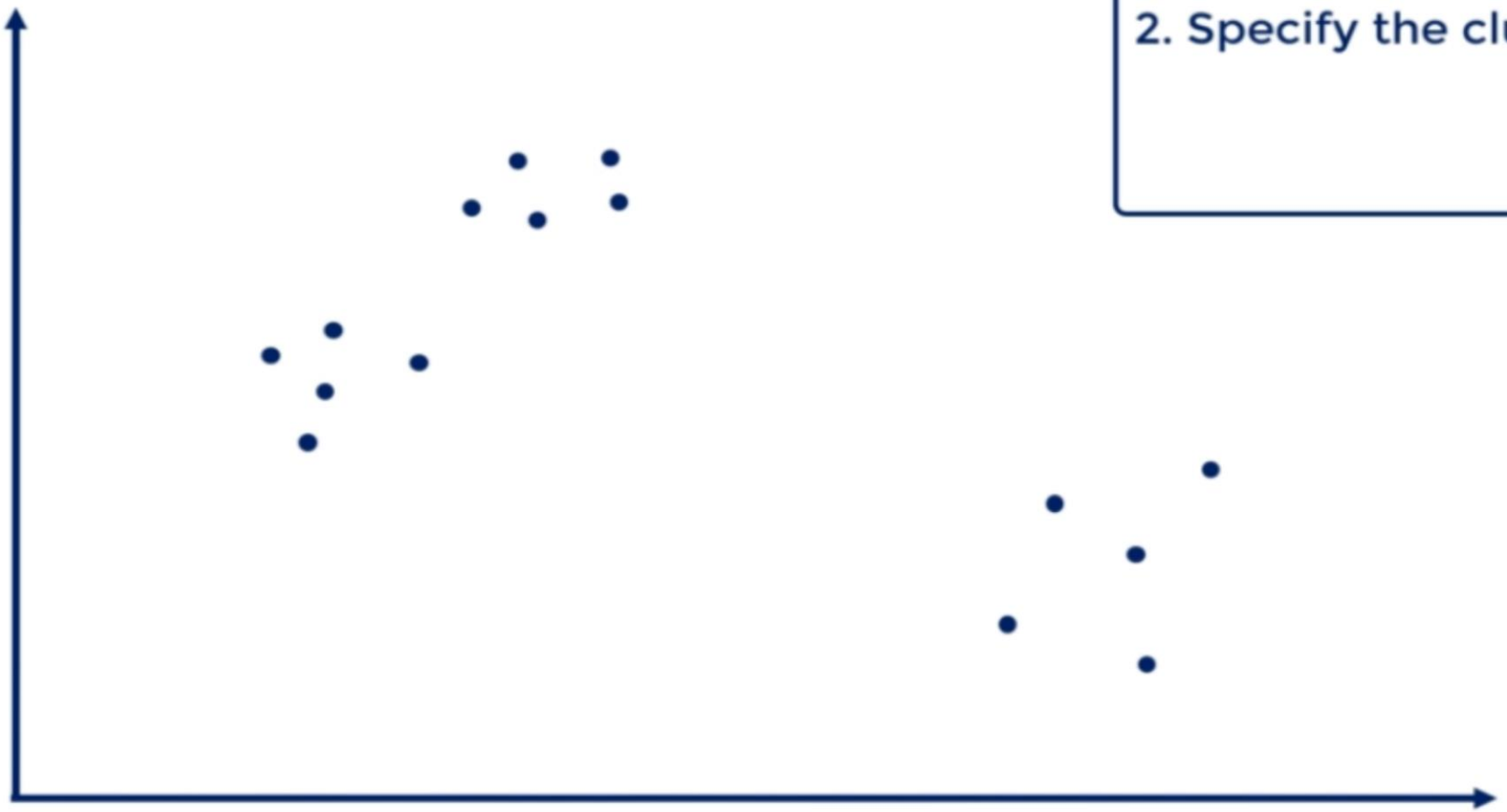
عضو هیئت علمی دانشگاه گنبدکاووس

K-means clustering

1. Choose number of clusters	2. Specify the number of seeds	3. Assign each point to a centroid	4. Adjust the centroids
Number – K, chosen by the person performing the clustering	A seed is a starting centroid (can be chosen at random, with an algorithm or according to some prior knowledge)	Based on proximity (measured by Euclidian distance)	Repeat 2. and 3. until there is you can no longer find a better clustering solution

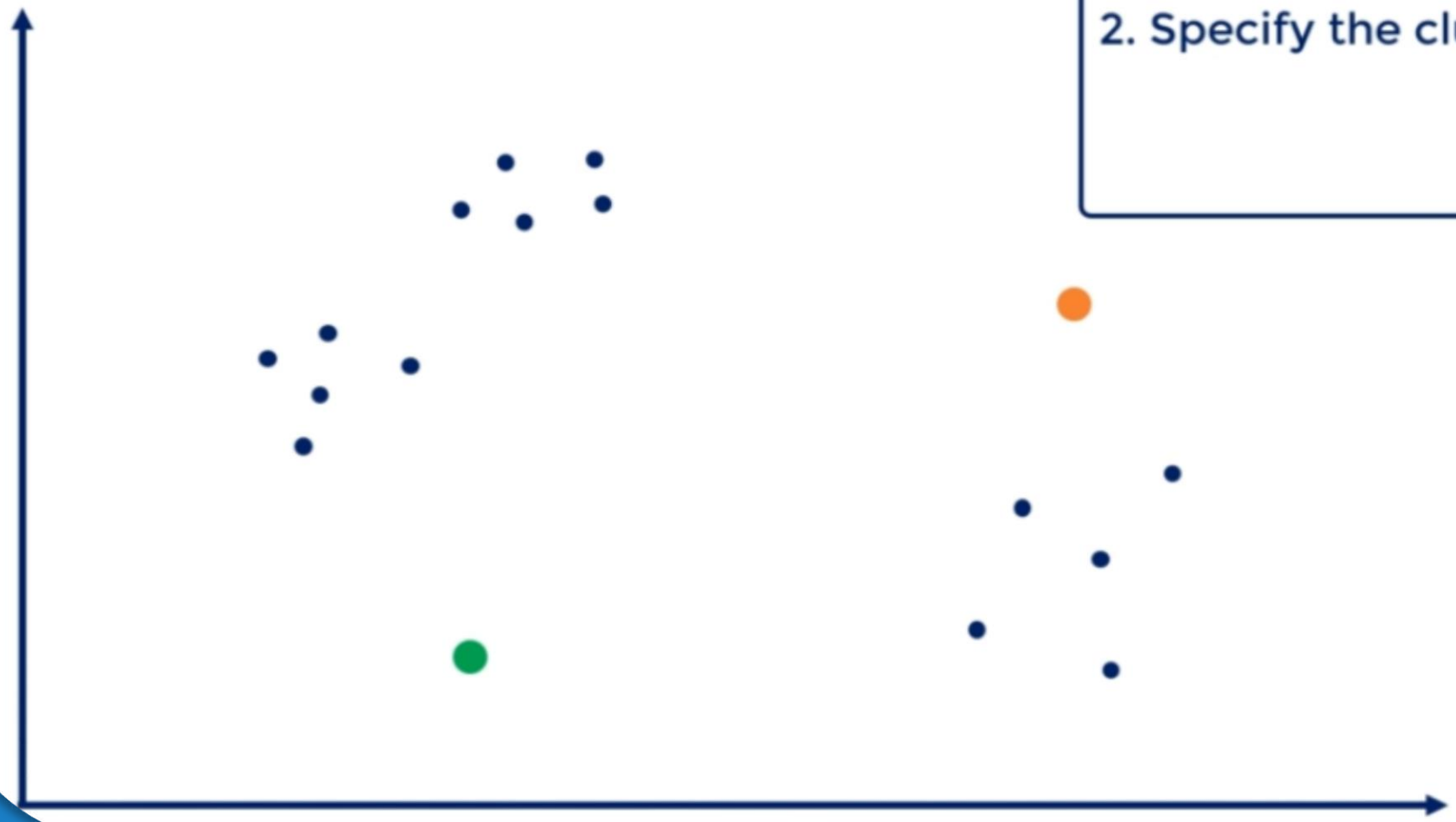
K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds



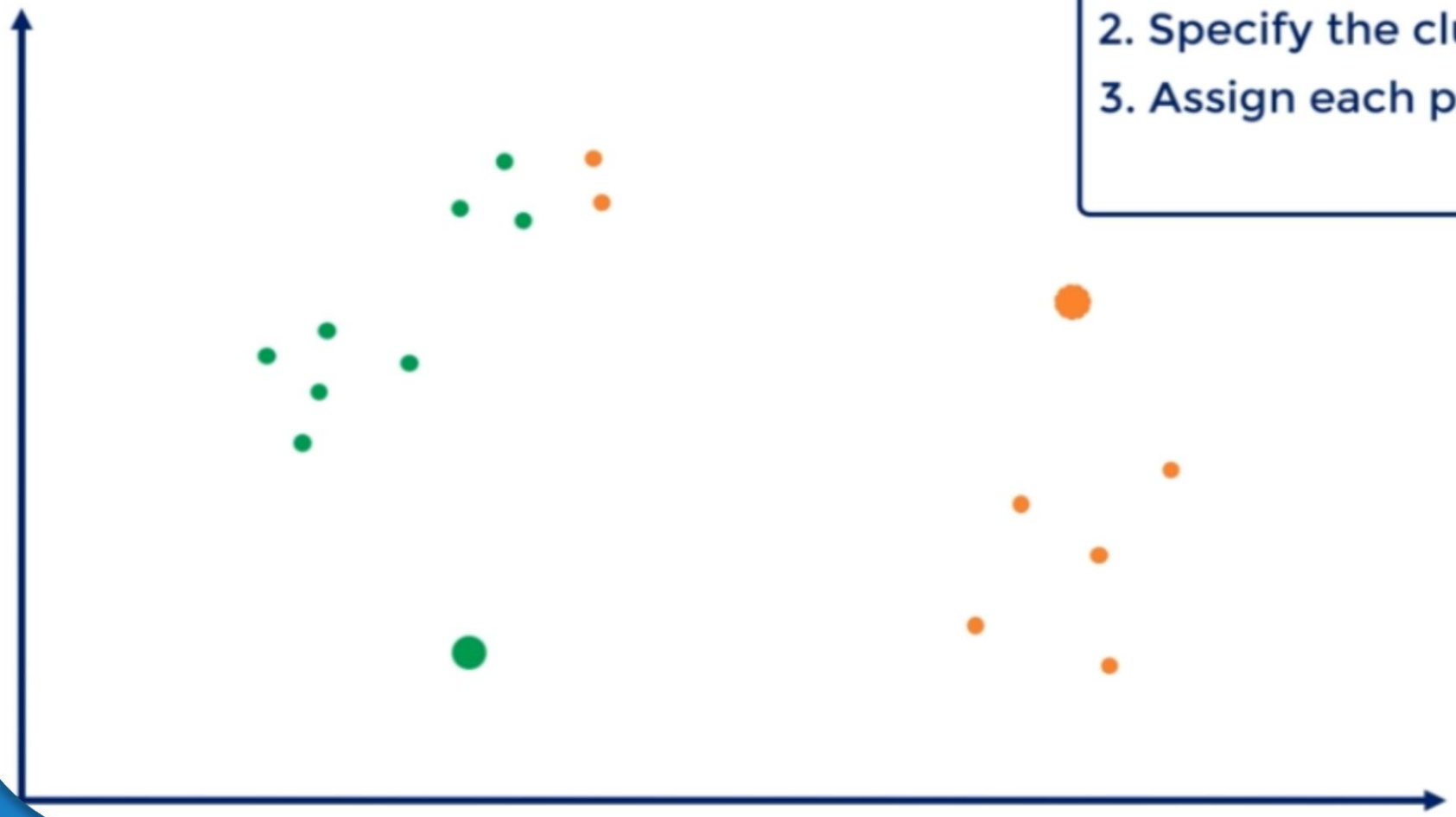
K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds



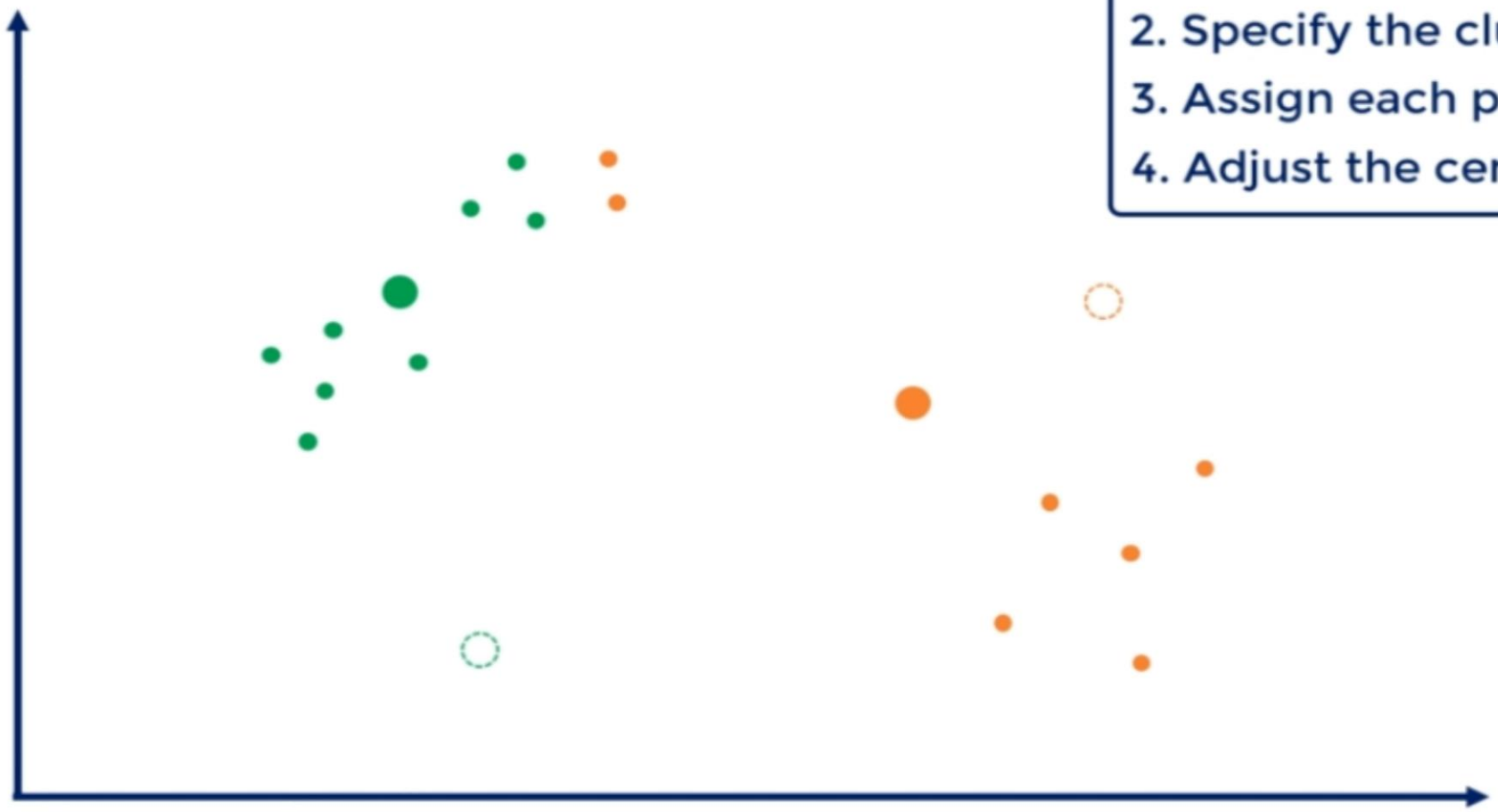
K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid



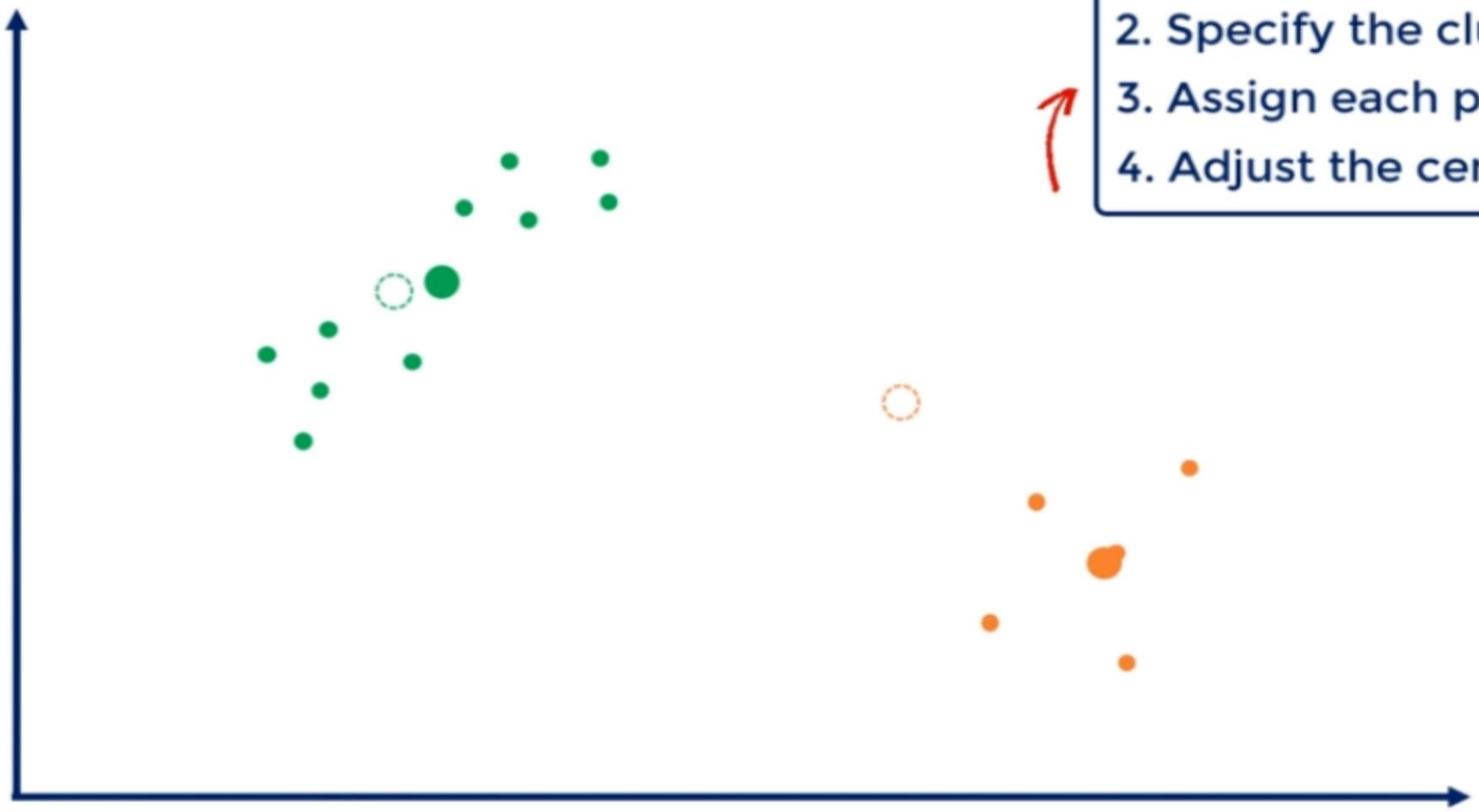
K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids



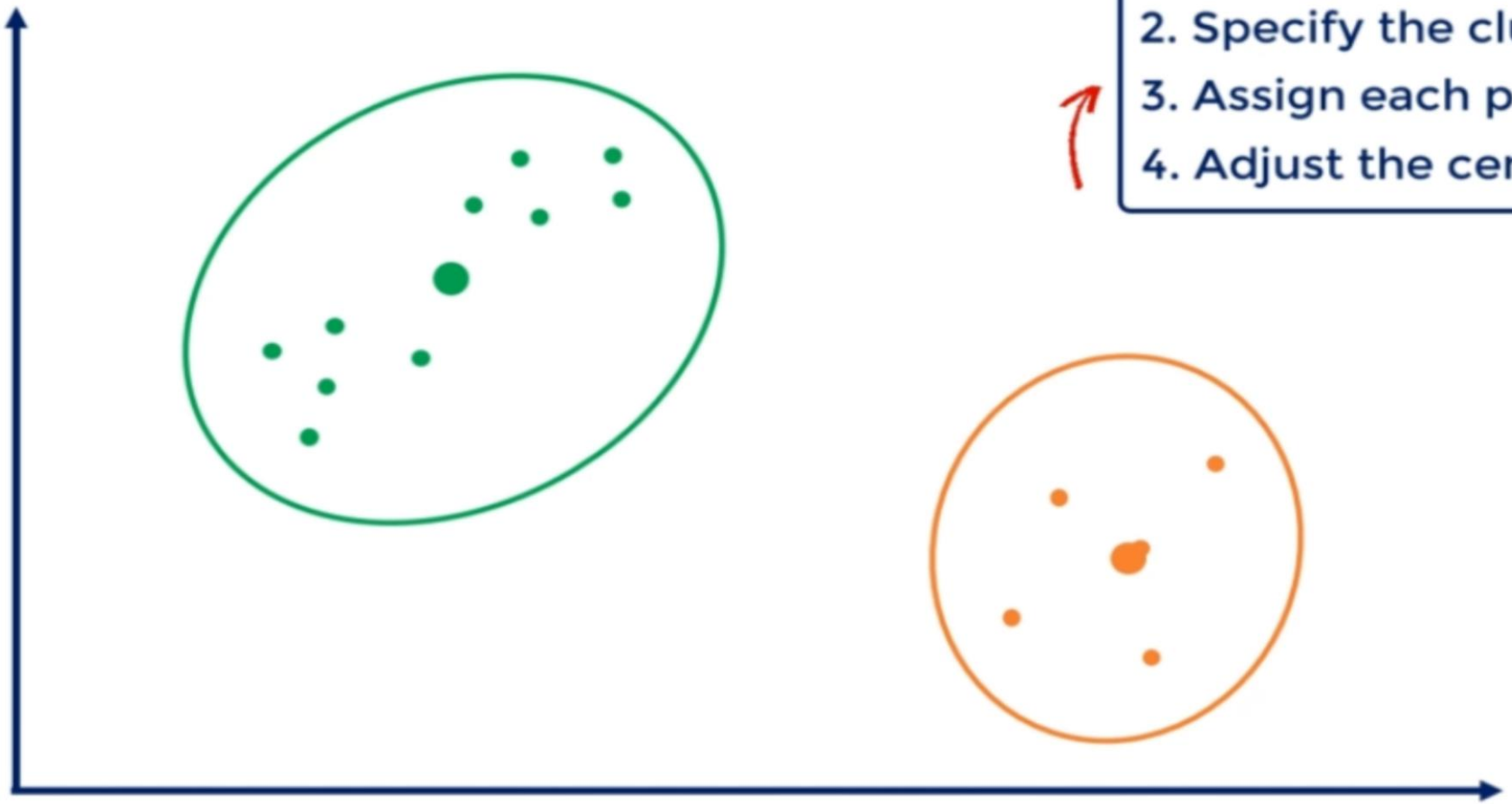
K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids

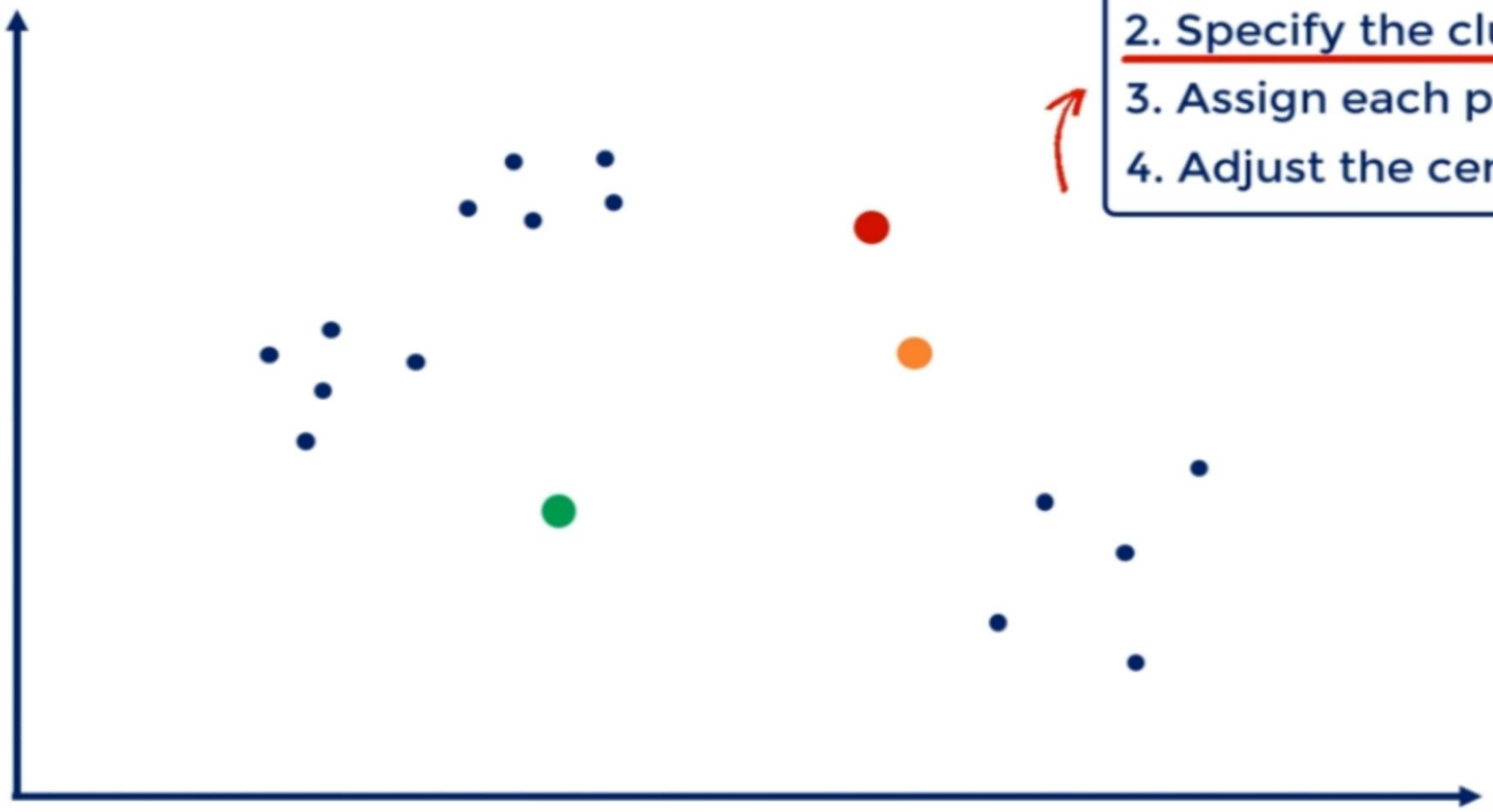


K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids



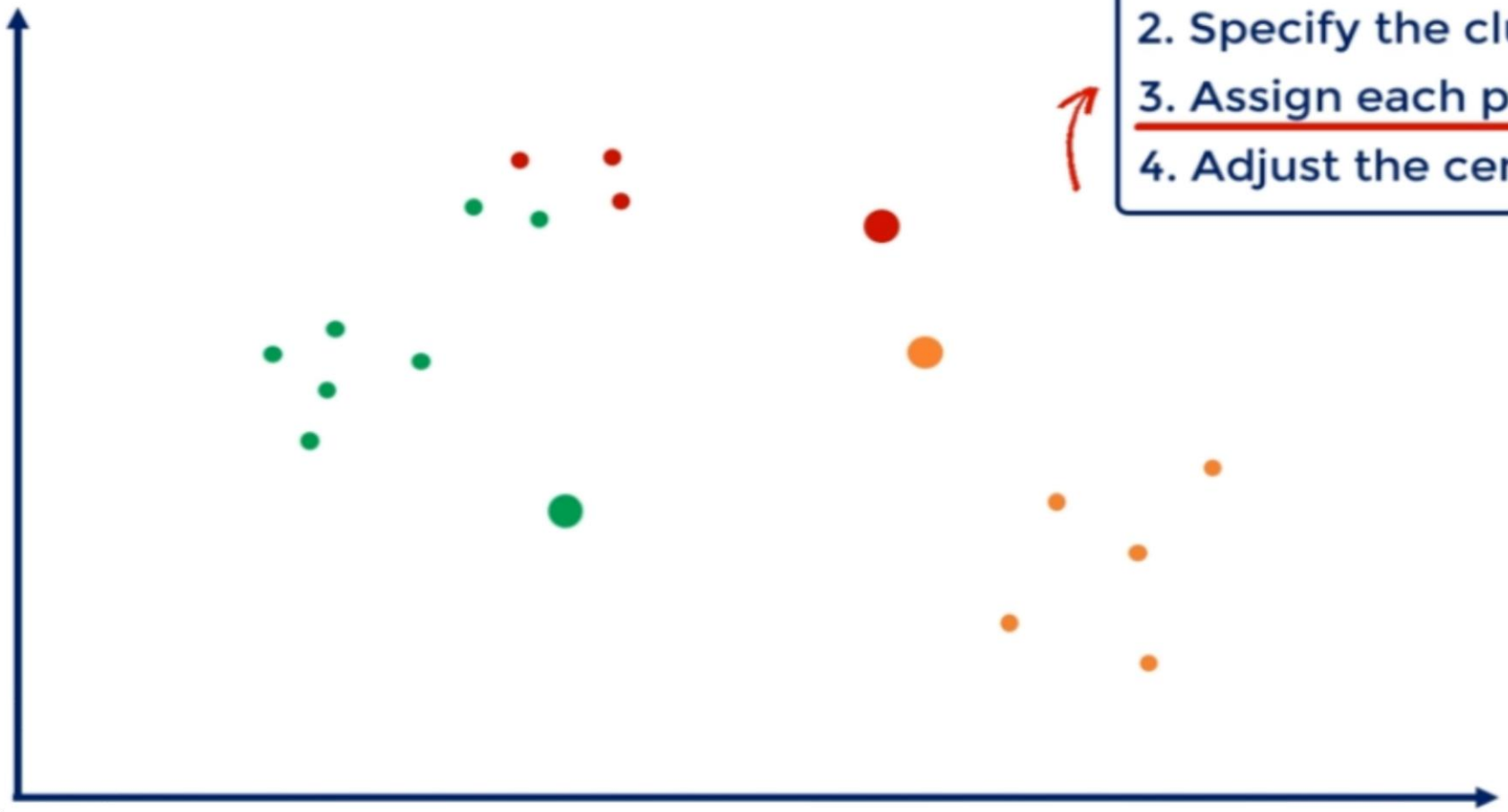
K-means clustering



1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids

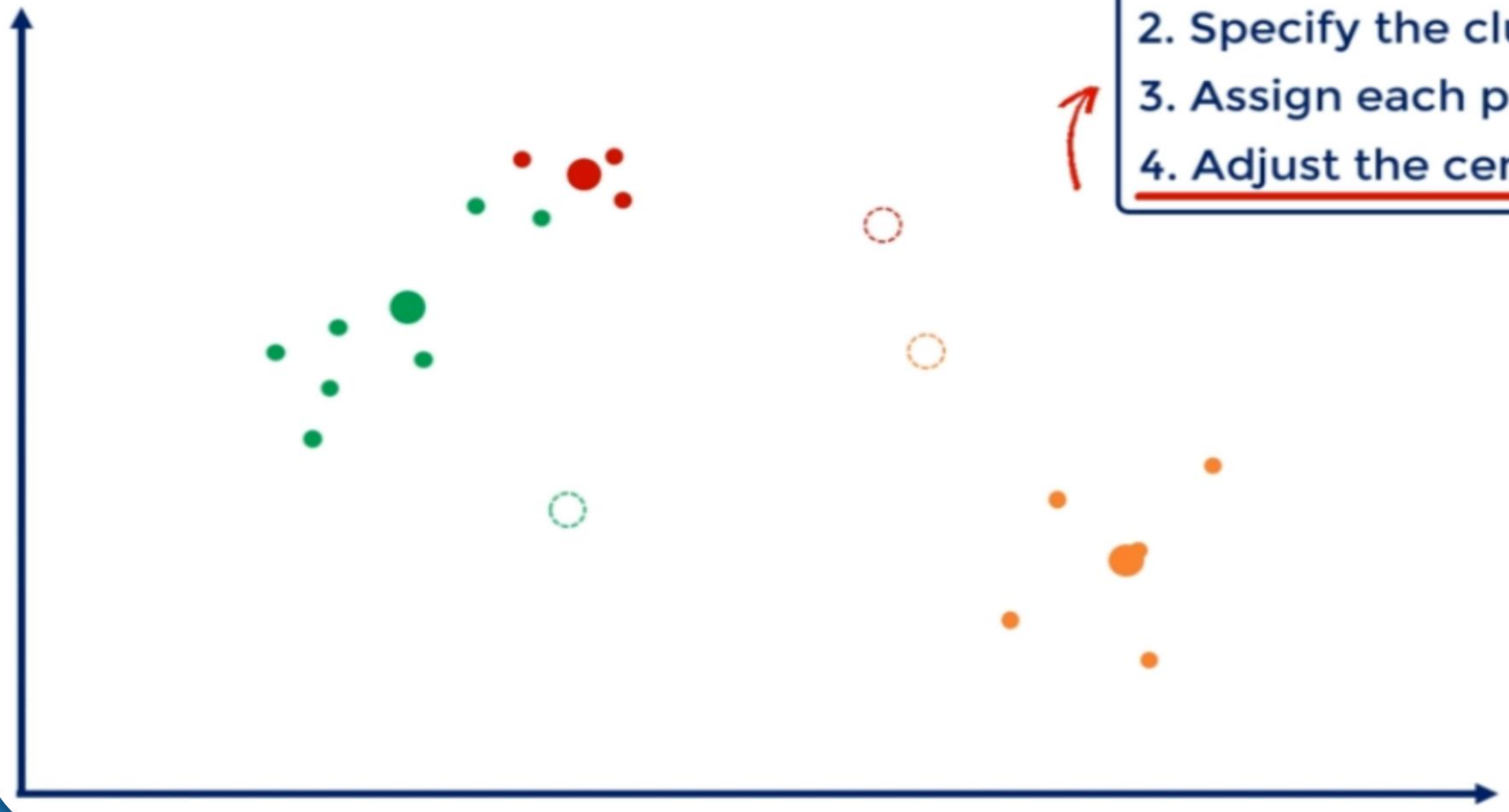
K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids



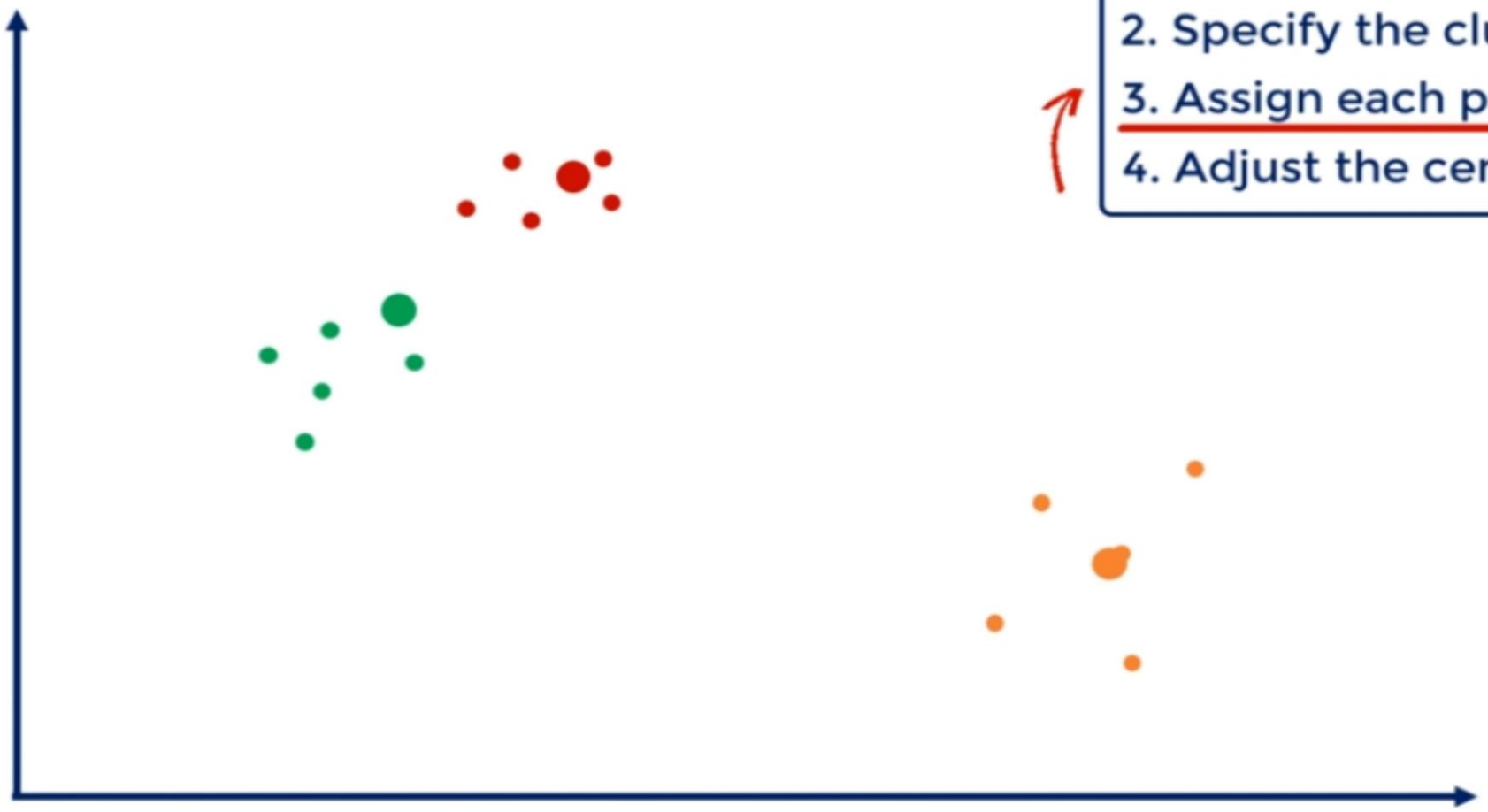
K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids

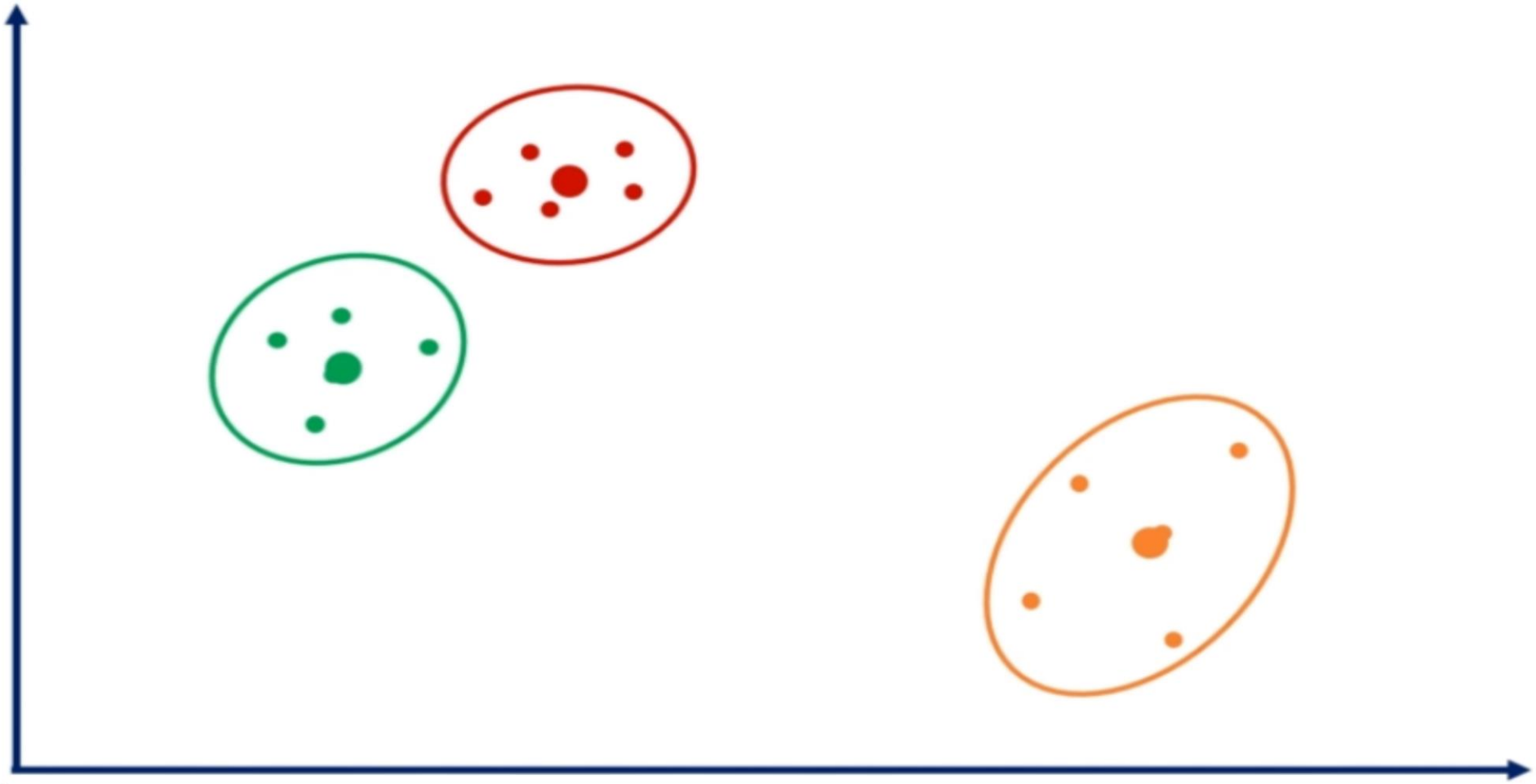


K-means clustering

1. Choose the number of clusters
2. Specify the cluster seeds
3. Assign each point to a centroid
4. Adjust the centroids



K-means clustering



K-means clustering - pros and cons

PROS

- Simple to implement
  (so many people can use it)
- Computationally efficient
  (it takes considerably less time than any hierarchical clustering model)
- Widely used
  (popular, therefore, in demand)
- Always yields a result
  (also a con as it may be deceiving)

CONS

- We need to pick K
  (often, we don't know how many clusters we need)
- Sensitive to initialization
  (but we can use methods such as kmeans++ to determine the seeds)
- Sensitive to outliers
  (by far the biggest downside of k-means)
- Produces spherical solutions
  (thus, not as generalizable)

Never Give up. You Can Nail it.