machine learning

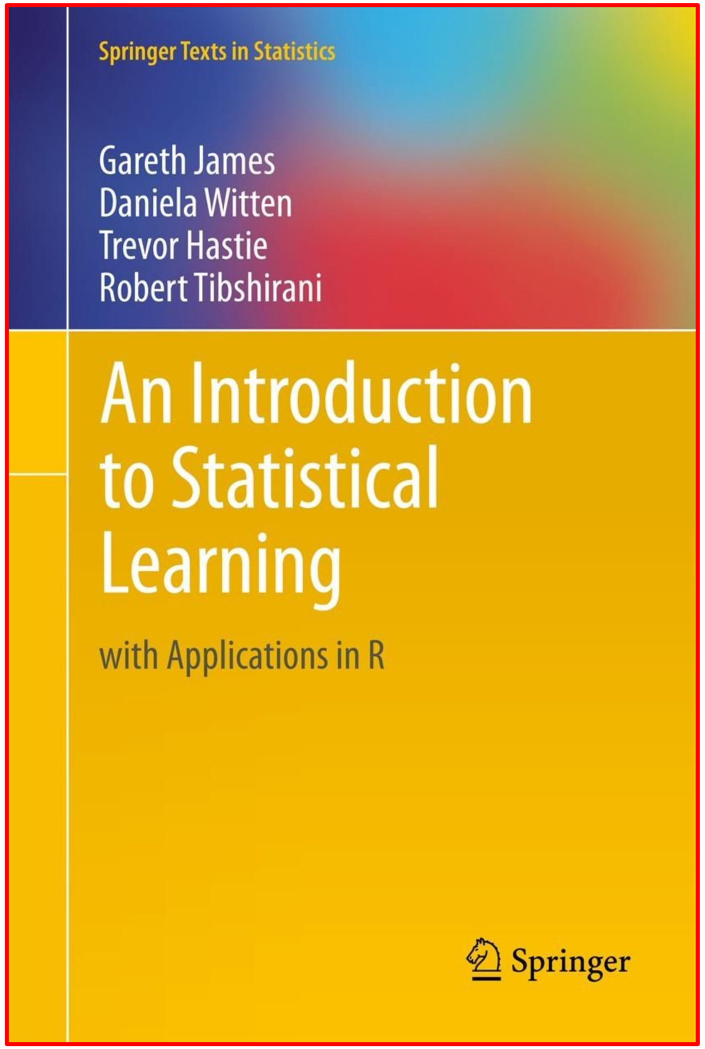
03 - k-means

francisco josé diego acosta

k-means theory

Chapter 10
An Introduction to Statistical Learning
by Gareth James, et al.

https://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf



k-means theory

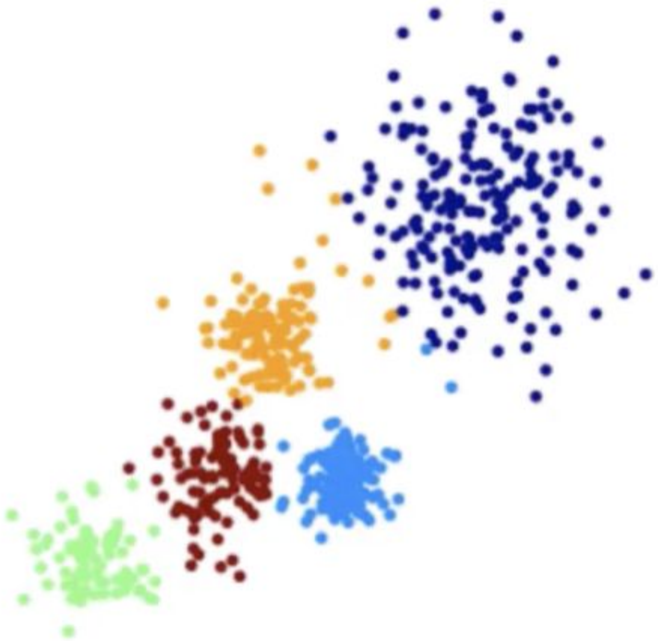
- **Unsupervised** learning algorithm
- It is one of the simplest way to solve a **clustering** problem
- typical clustering problems:
    - cluster similar documents
    - custer customers
    - market segmentation
    - identify similar physical groups

k-means theory

- what is a cluster?
  A cluster refers to a collection of data points aggregated together because of certain similarities.

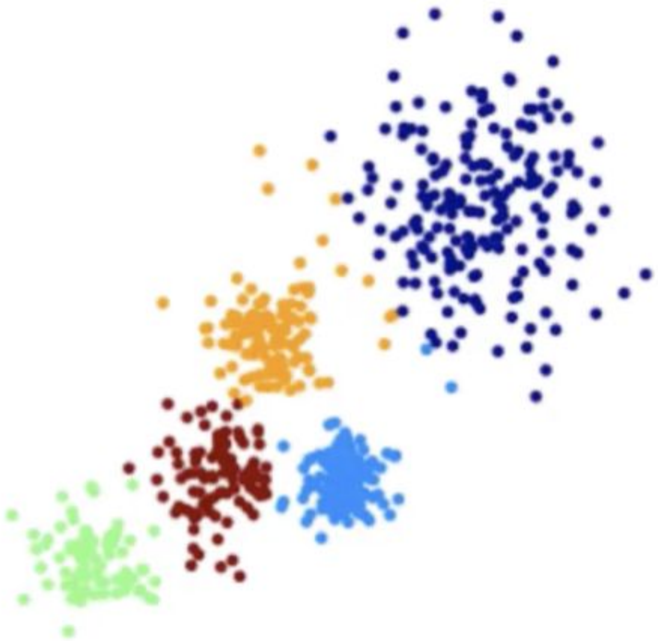
k-means theory

- what is a cluster?
  A cluster refers to a collection of data points aggregated together because of certain similarities.



k-means theory

- The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

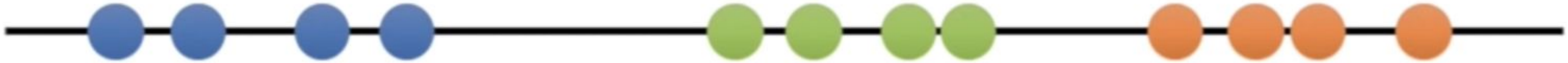


k-means theory

- Imagine you have some data you can plot in a line
- You already know the data is grouped in 3 clusters

k-means theory

- Imagine you have some data you can plot in a line
- You already know the data is grouped in 3 clusters
- In this case, the clusters are easy to see



k-means theory

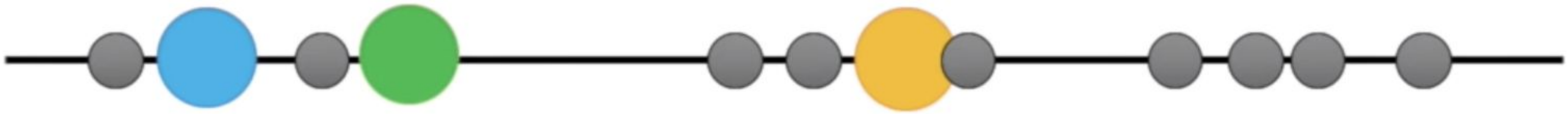
0. Let’s start with the raw data

k-means theory

1. Select the number of clusters you want to identify.
   This is the "K" in K-Means clustering
   Let's select k = 3

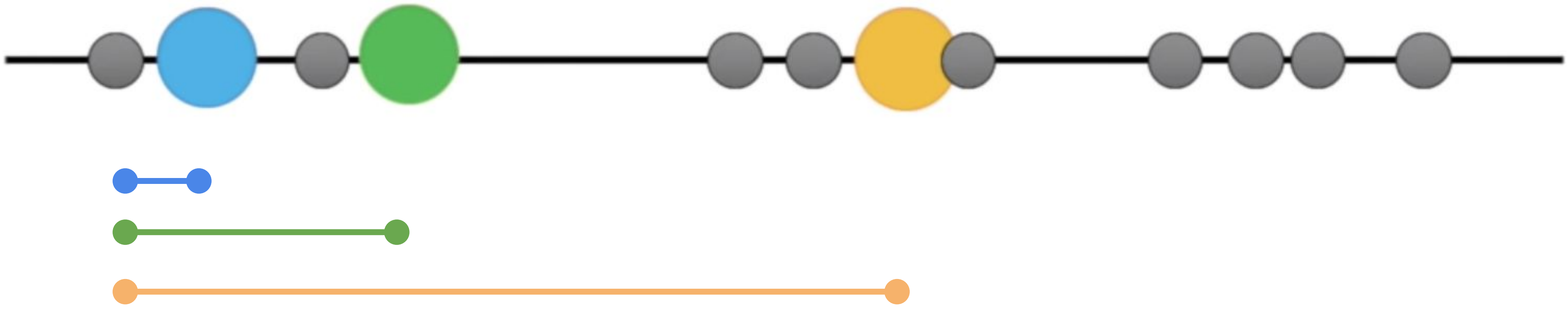
k-means theory

2. Randomly select 3 distinct data points
These are the initial clusters centroids

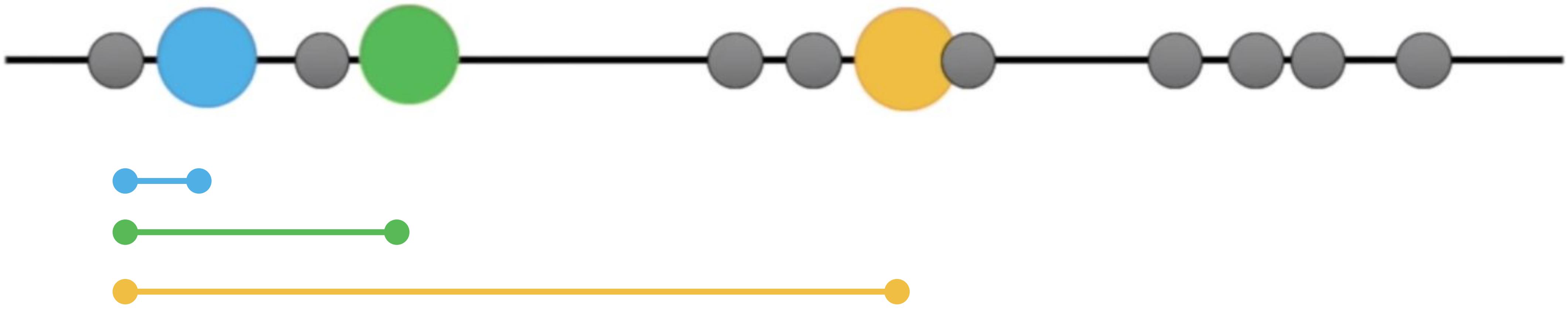


k-means theory

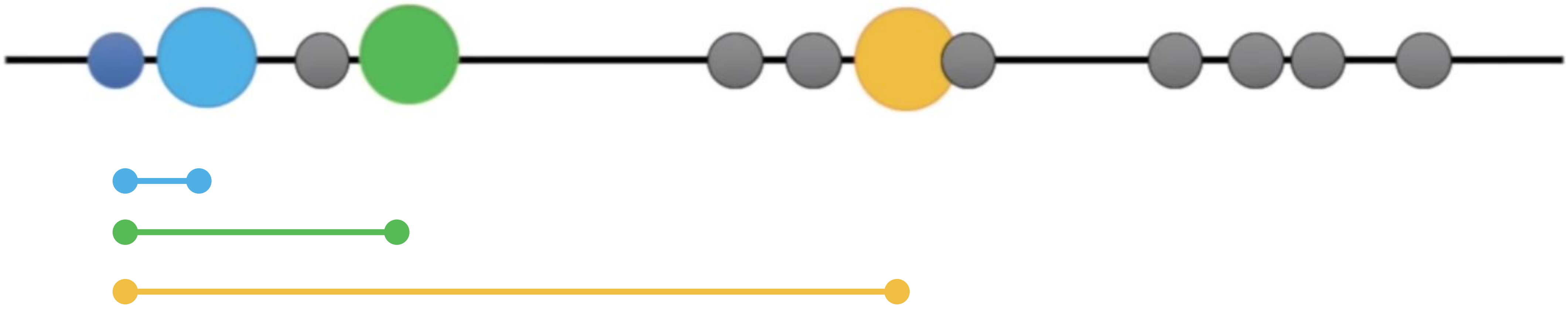
3. Measure the distance between the first point and the three centroids



3. Measure the distance between the first point and the three centroids

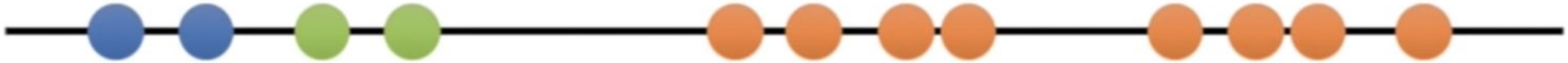


4. Assigns the point to the cluster to which the centroid is closest



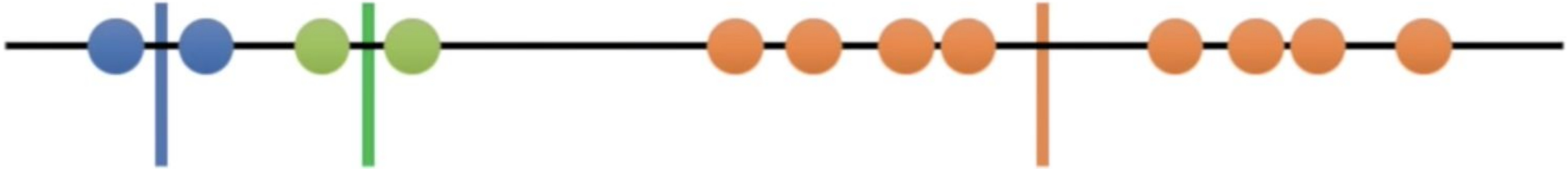
k-means theory

4. Assigns the point to the cluster to which the centroid is closest
Do the same thing for the rest of points



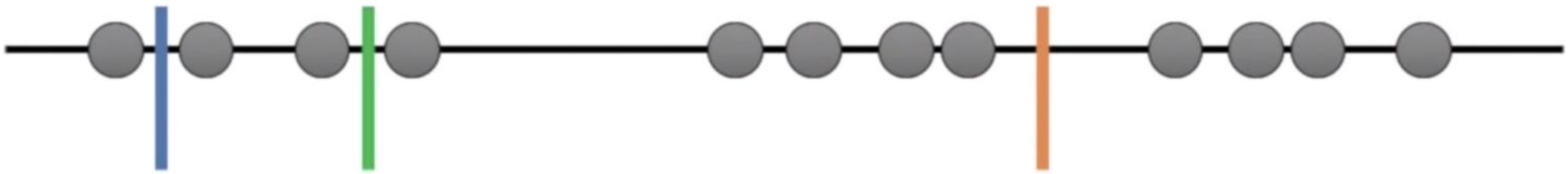
k-means theory

5. Calculate the mean of each cluster



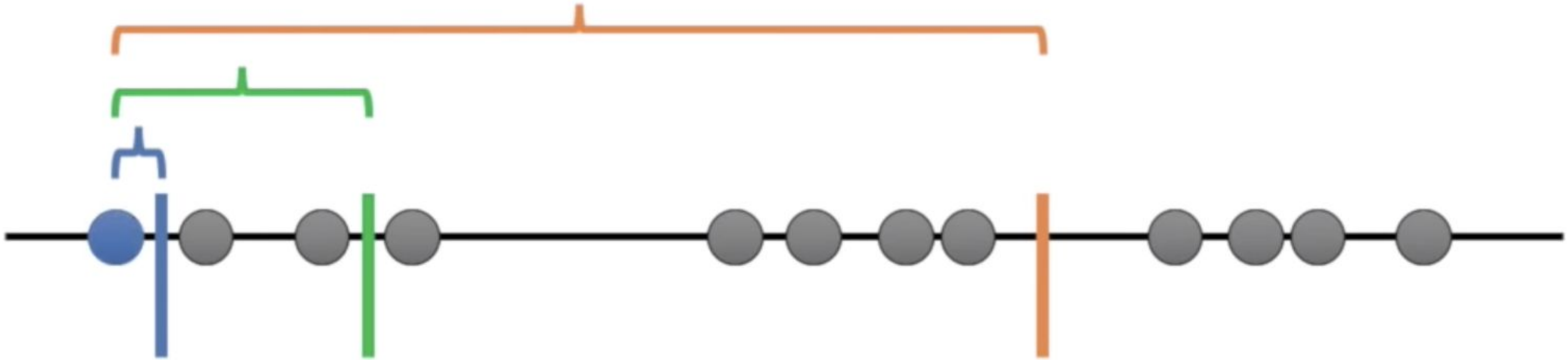
k-means theory

6. Repeat the process but using the mean instead of the centroids



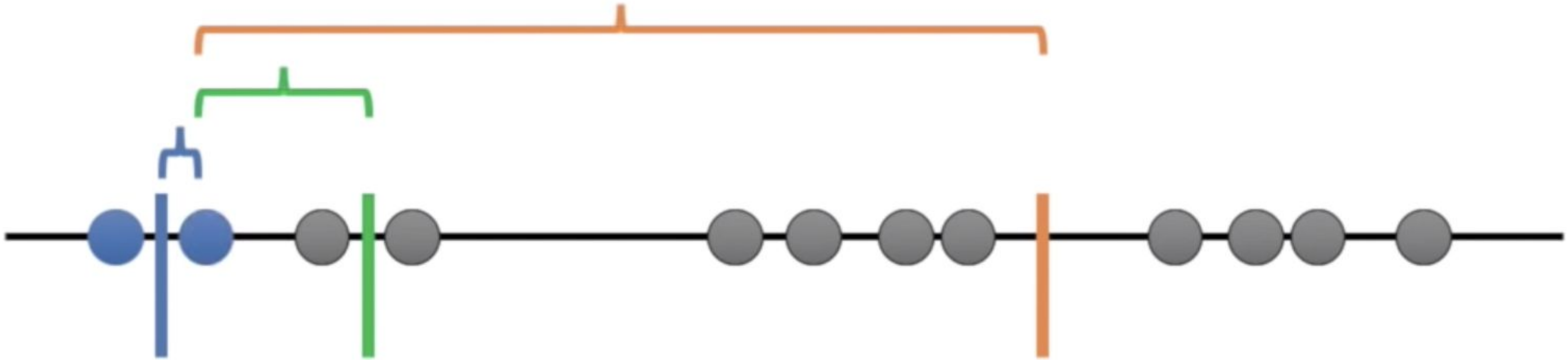
k-means theory

6. Repeat the process but using the mean instead of the centroids



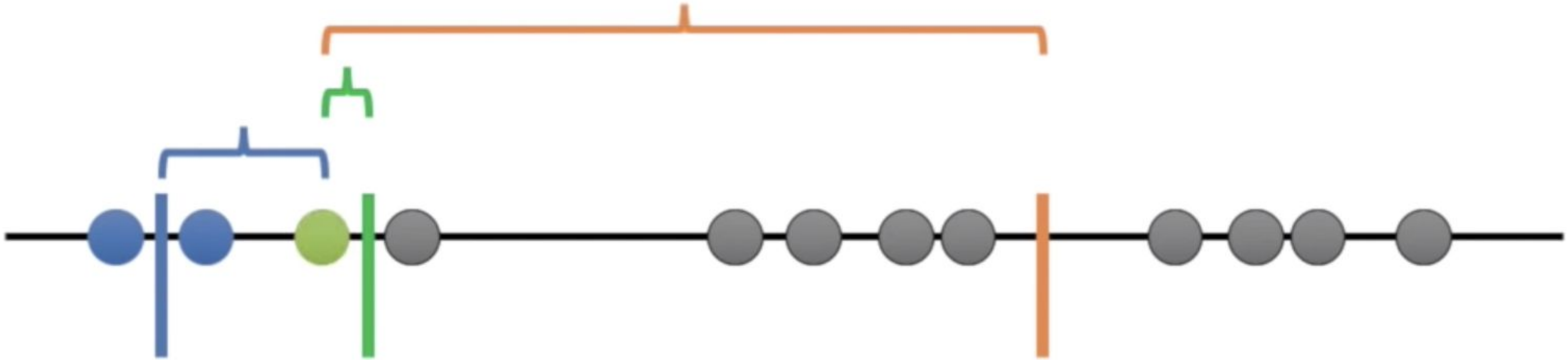
k-means theory

6. Repeat the process but using the mean instead of the centroids



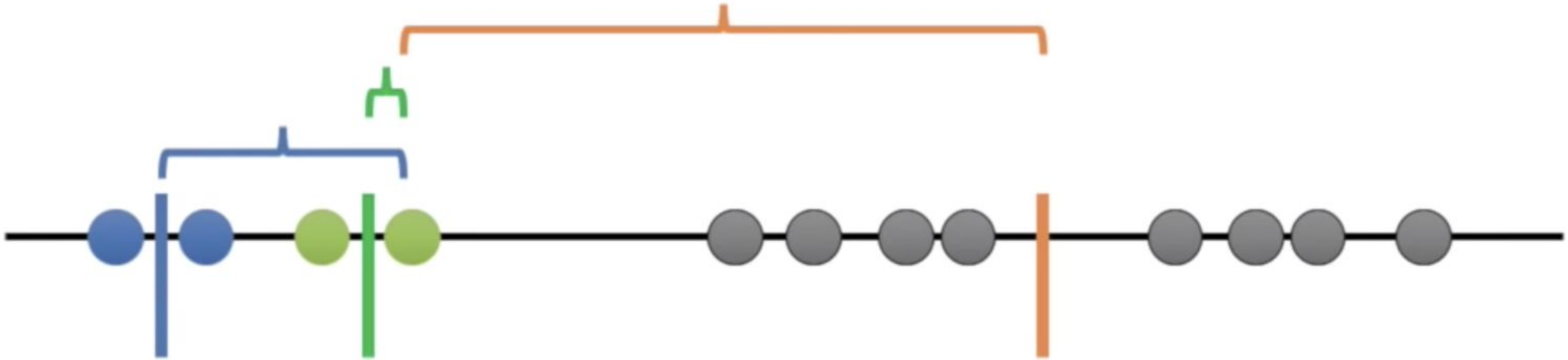
k-means theory

6. Repeat the process but using the mean instead of the centroids



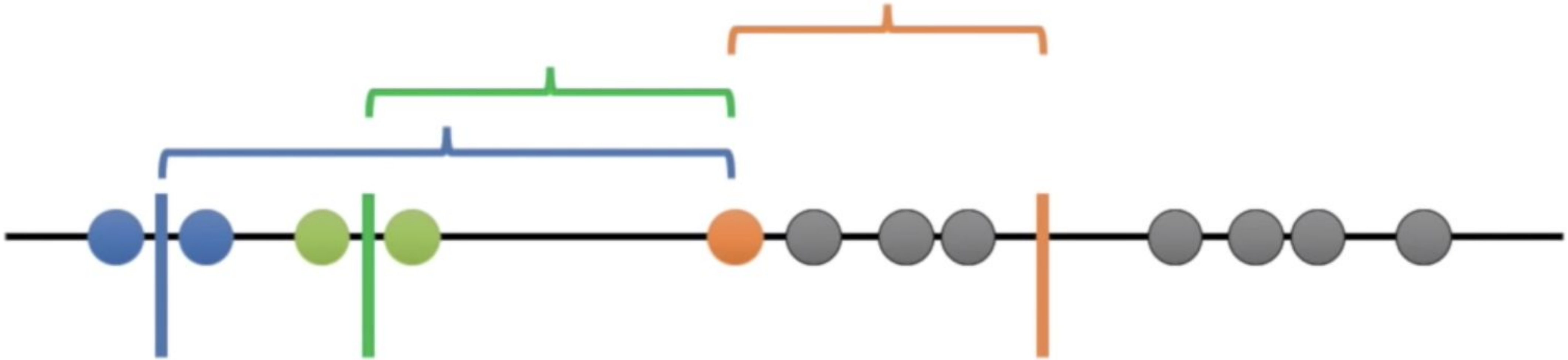
k-means theory

6. Repeat the process but using the mean instead of the centroids



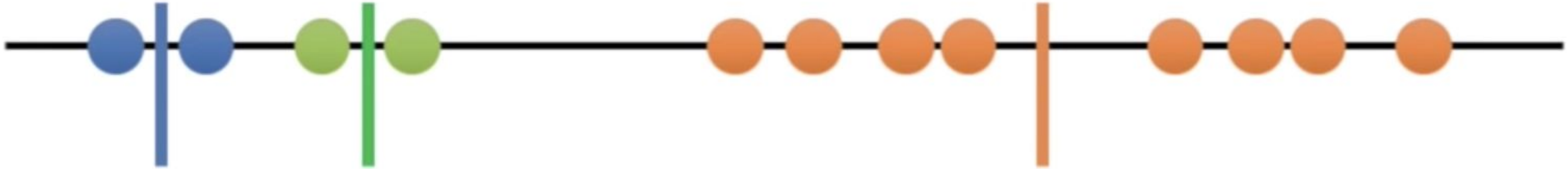
k-means theory

6. Repeat the process but using the mean instead of the centroids

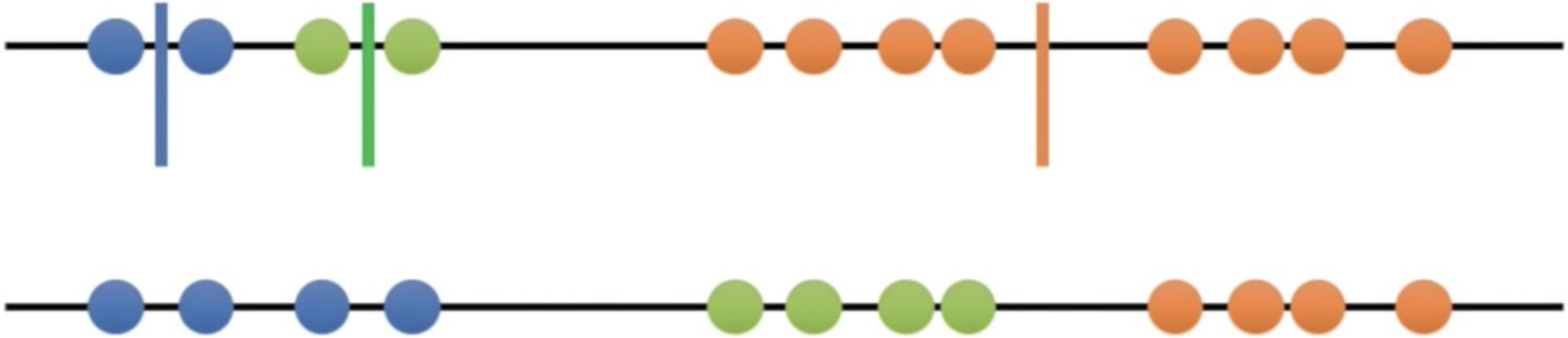


k-means theory

6. Repeat the process but using the mean instead of the centroids

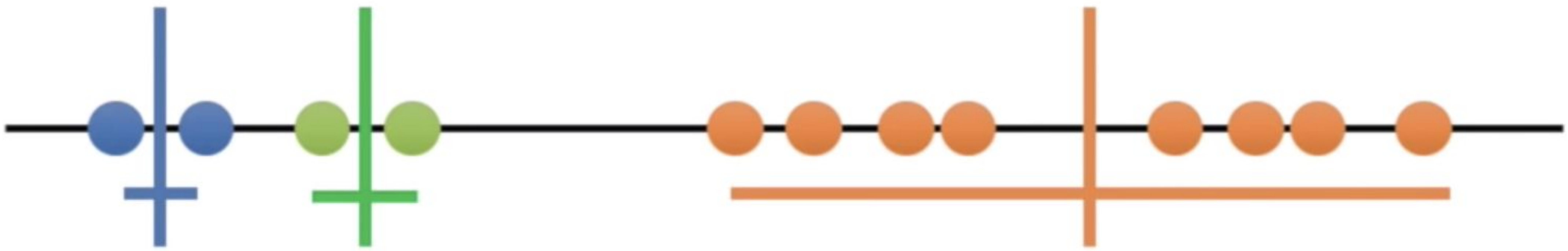


But this is not what we expected to have!



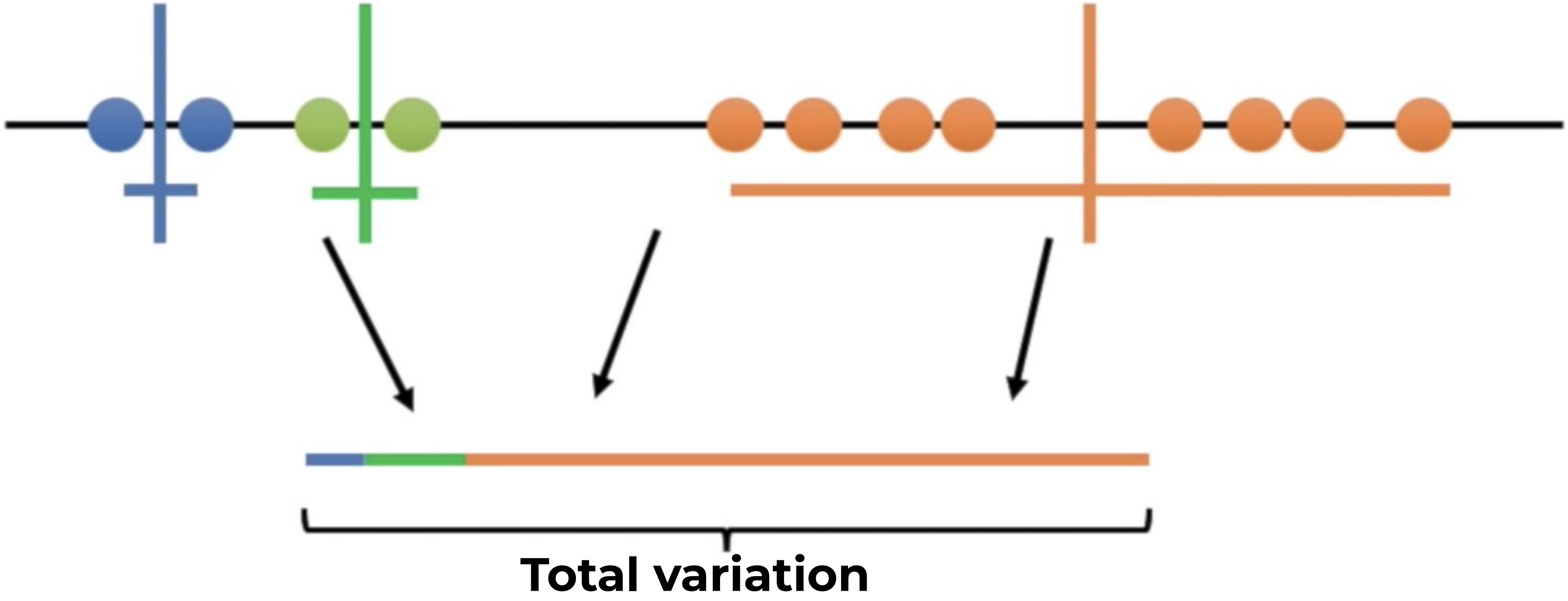
k-means theory

7. Add the variation within cluster



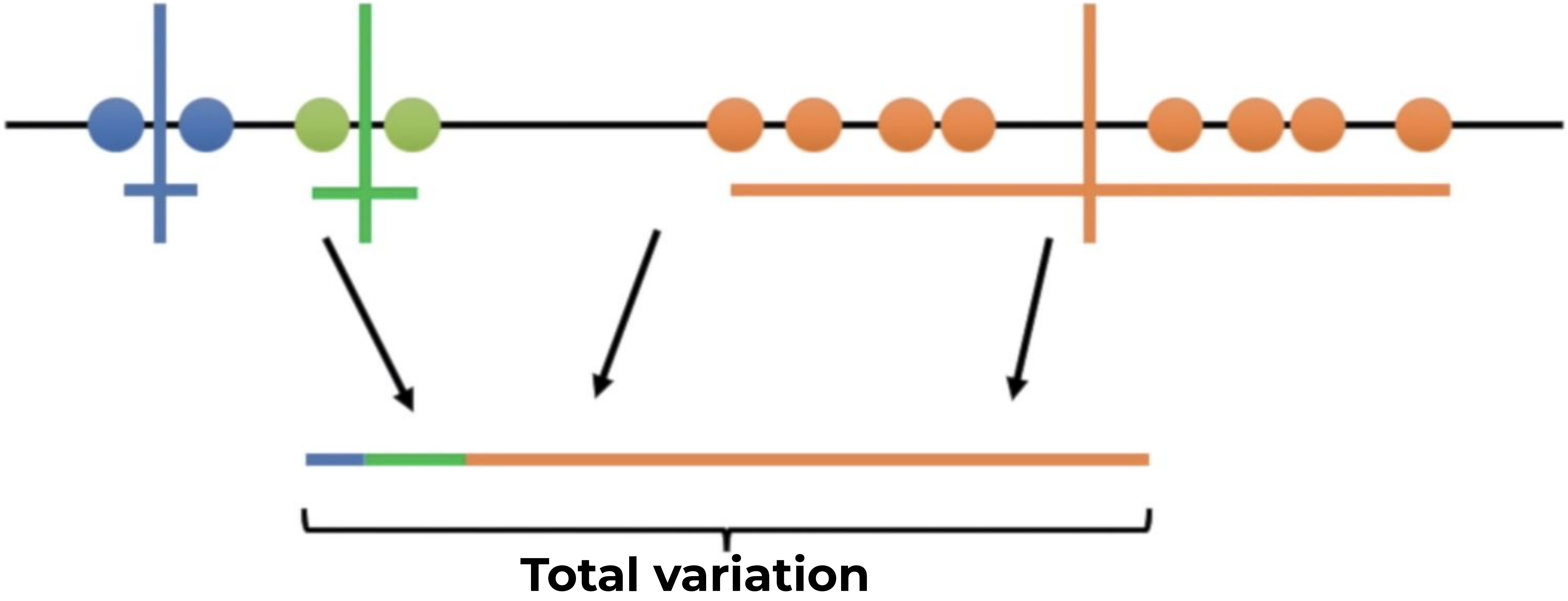
k-means theory

7. Add the variation within cluster



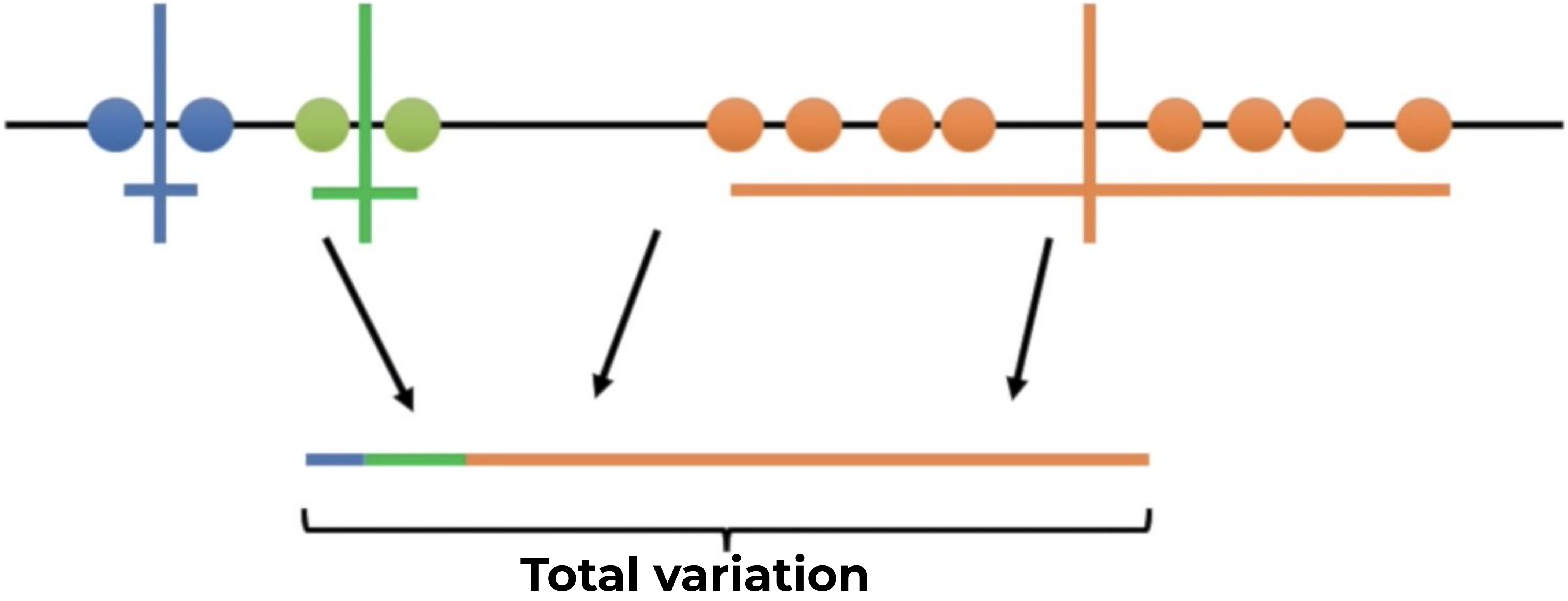
k-means theory

7. Add the variation within cluster



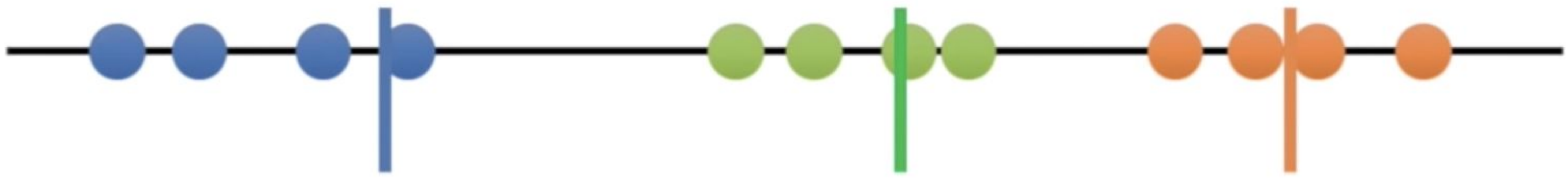
7. Add the variation within cluster

The goal now is to minimize the total variation. How?? -> **Iterate**



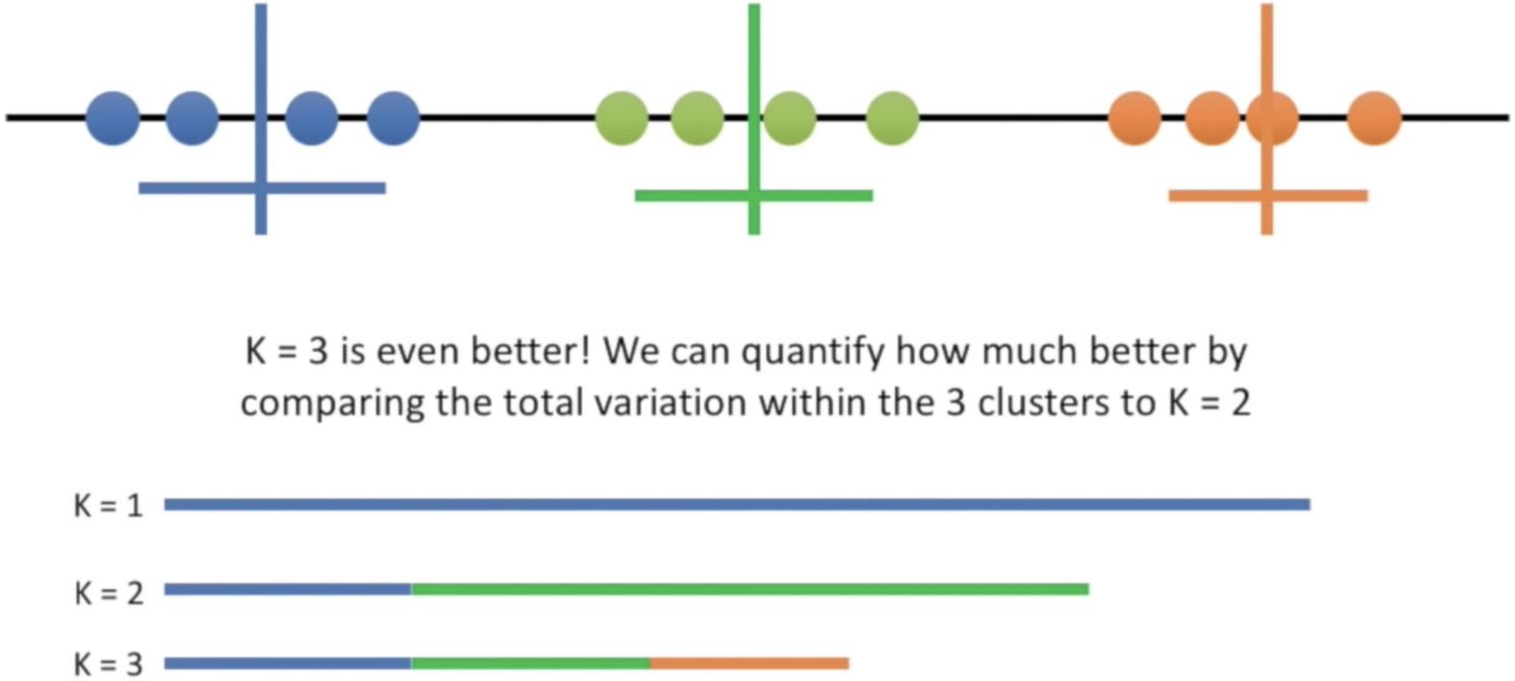
k-means theory

We have finally reach the cluster that minimize the total variation

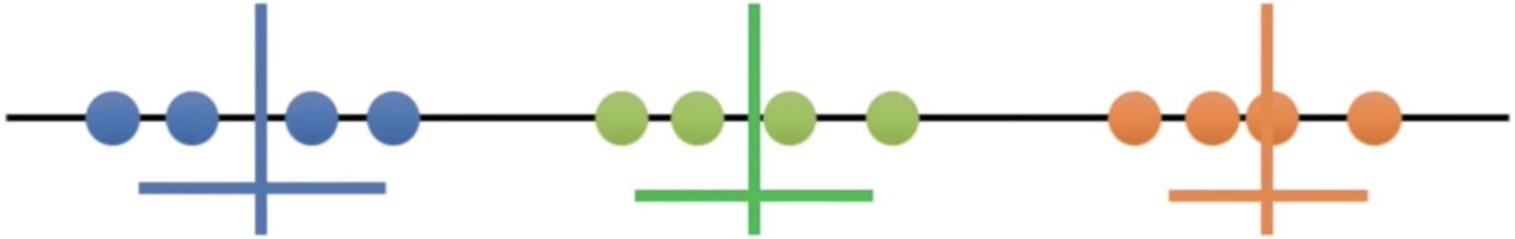


k-means theory

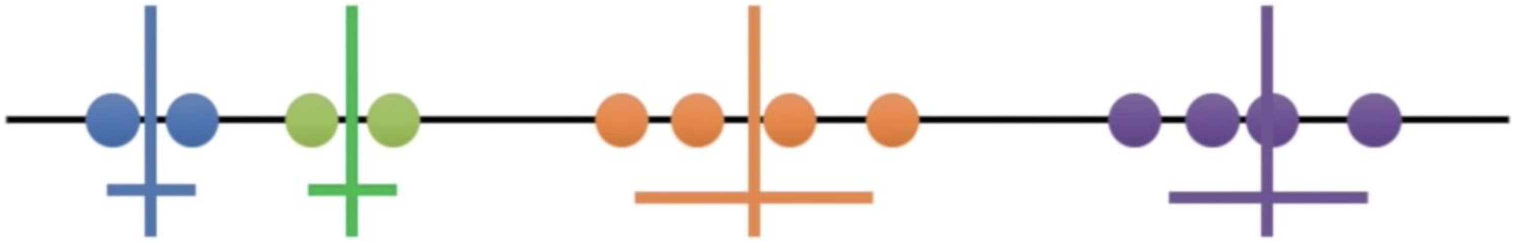
What is the best K? The one that minimize the variation

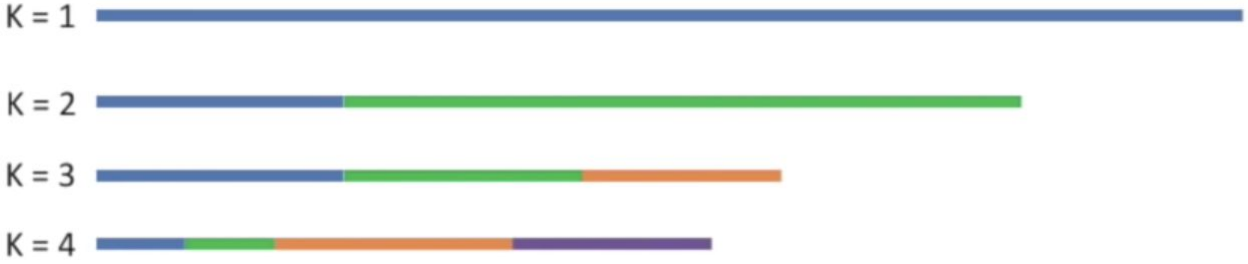


What is the best K? The one that minimize the variation



What if we try k = 4?





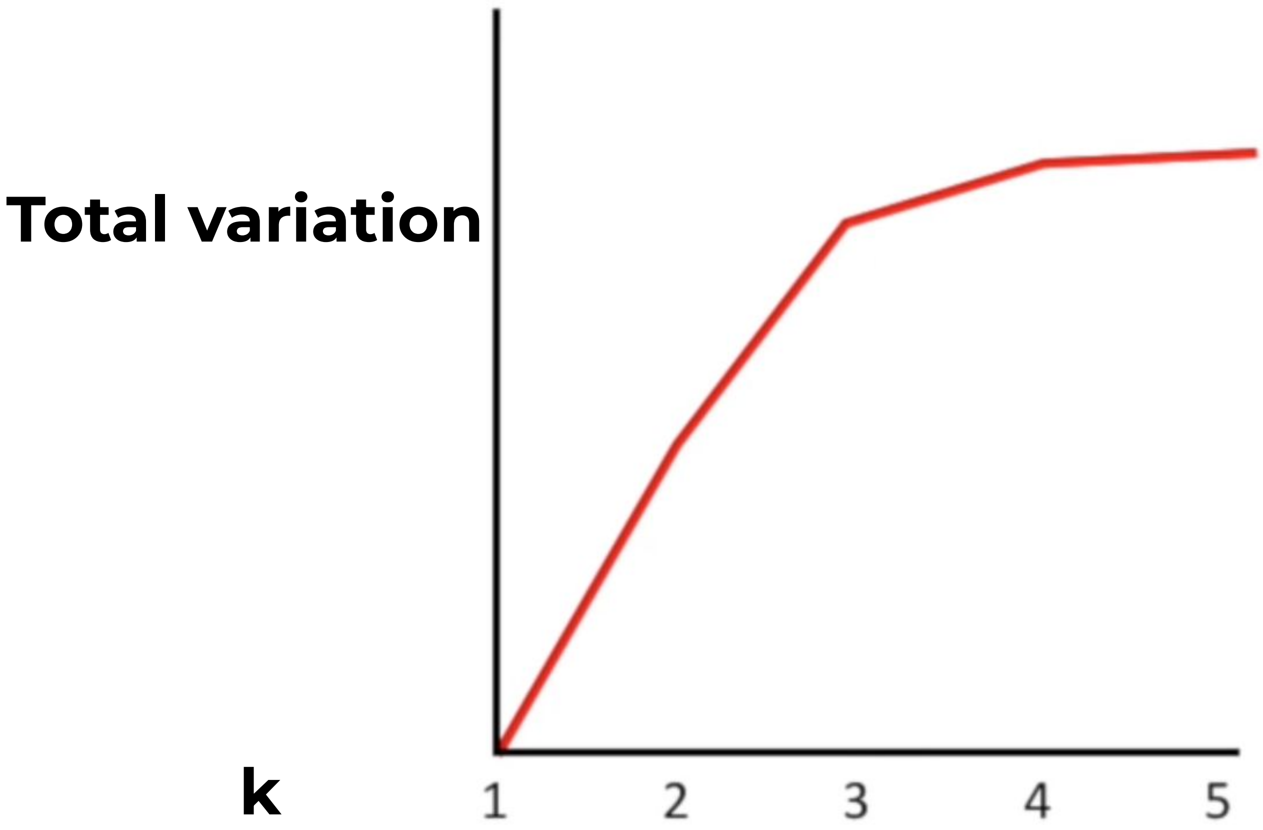
k-means theory

Each time we add a cluster (increases k by 1) the total variation is smaller.

So the best solution is when there is only one cluster per data point, innit??

If K=N then the variation is 0

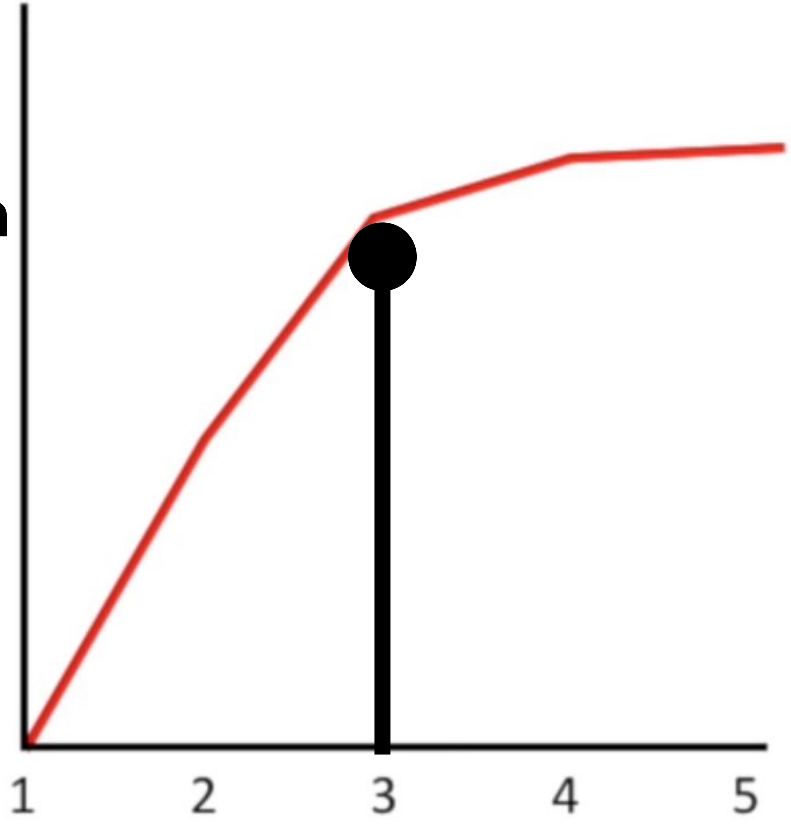
Solution : The elbow method



k-means theory

Solution : The elbow method

Total variation



k

k-means theory

- The objective of K-means is simple: group similar data points together and discover underlying patterns. To achieve this objective, K-means looks for a fixed number (k) of clusters in a dataset.

k-means theory

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.

2. Repeat until convergence: {

   For every i, set
   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

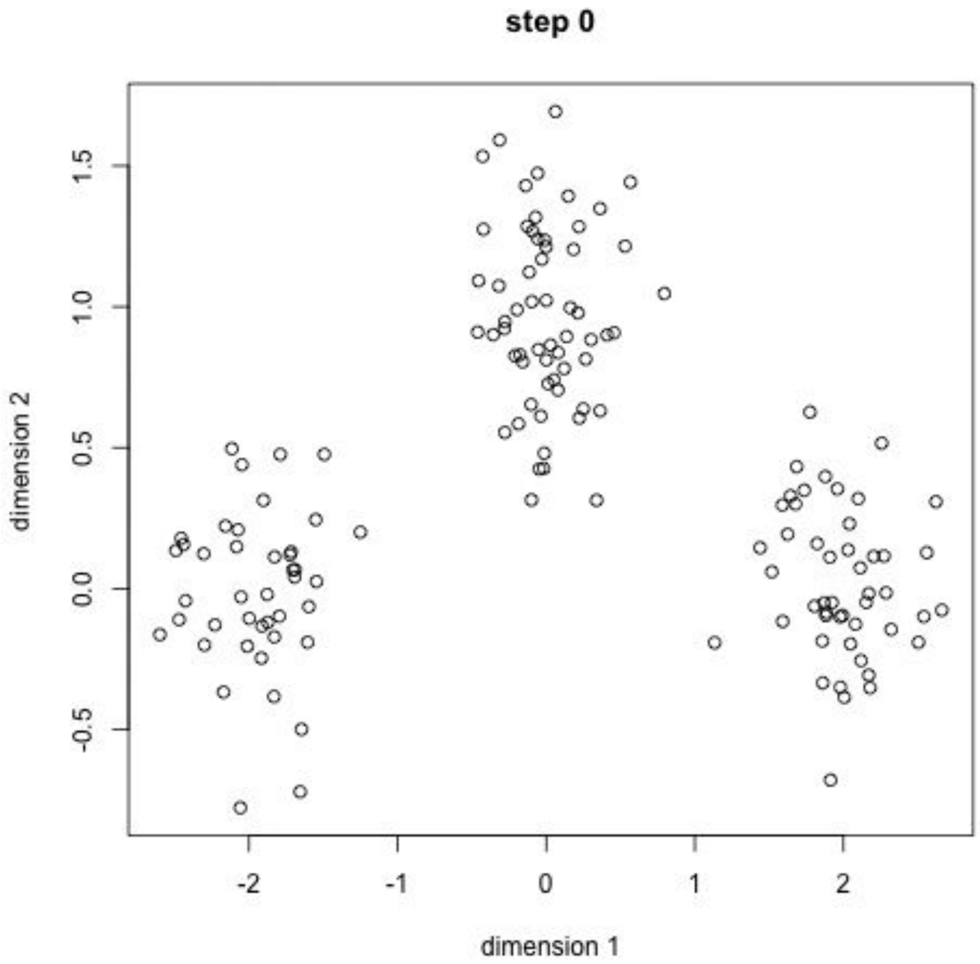
   For each j, set
   $$\mu_j := \frac{\sum_{i=1}^m 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m 1\{c^{(i)} = j\}}.$$

   }

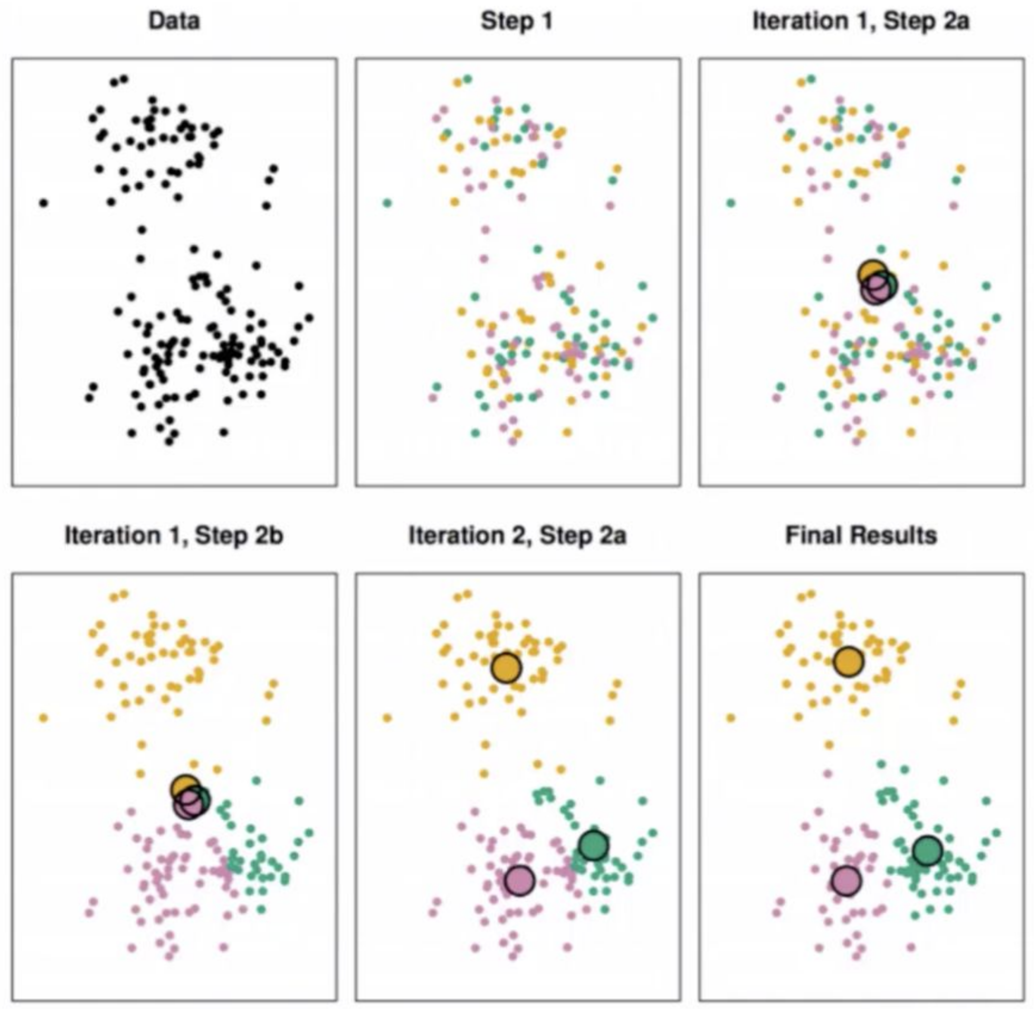
k-means theory

1. **you** choose a number of clusters **K**
2. kmeans randomly create K centroids (which are the imaginary data point that represent the cluster) and assign each data point to the cluster which centroid is closest
3. Until convergence repeat:
    a. for each cluster, compute the cluster centroid by taking the **mean** vector of data points in the cluster
    b. assign each data point to the cluster for which the centroid is the closest.

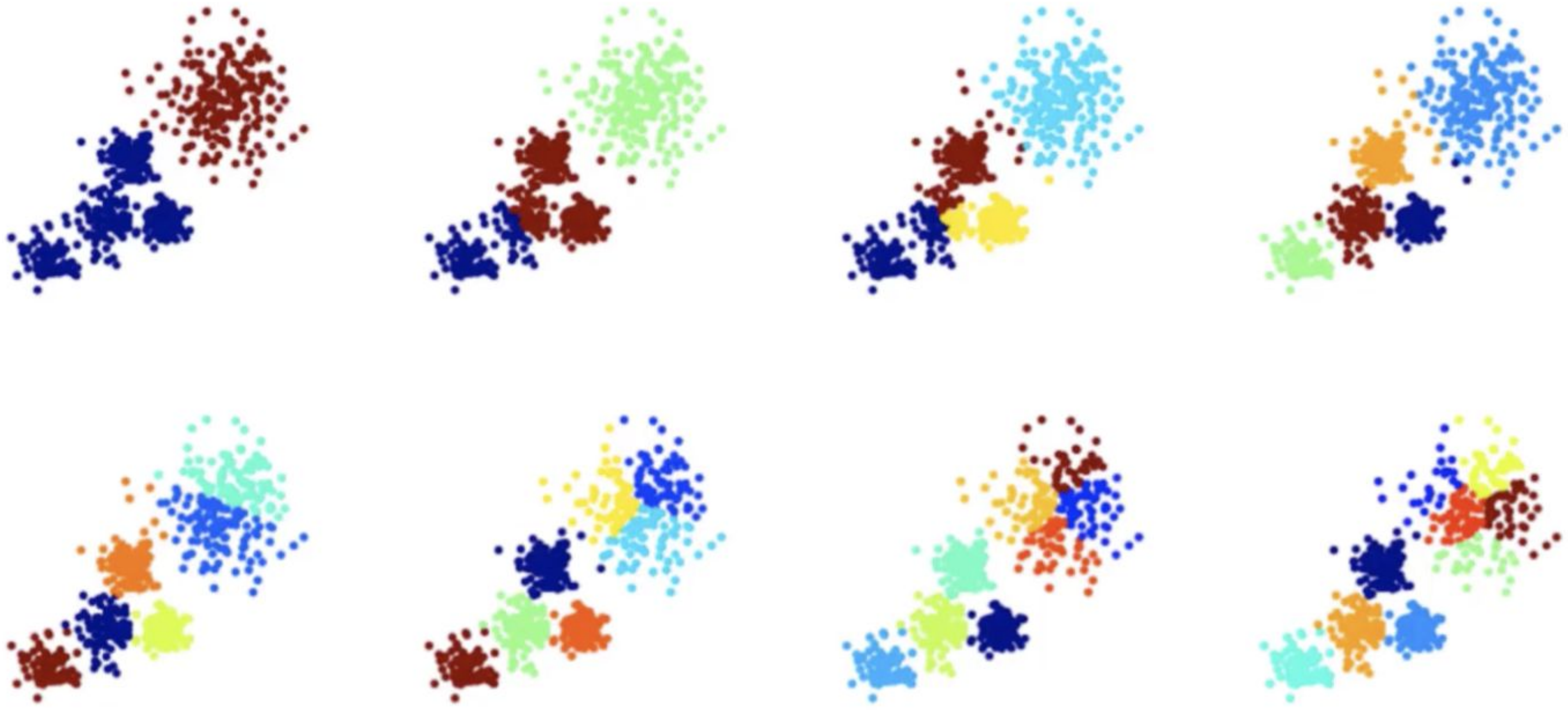
k-means theory



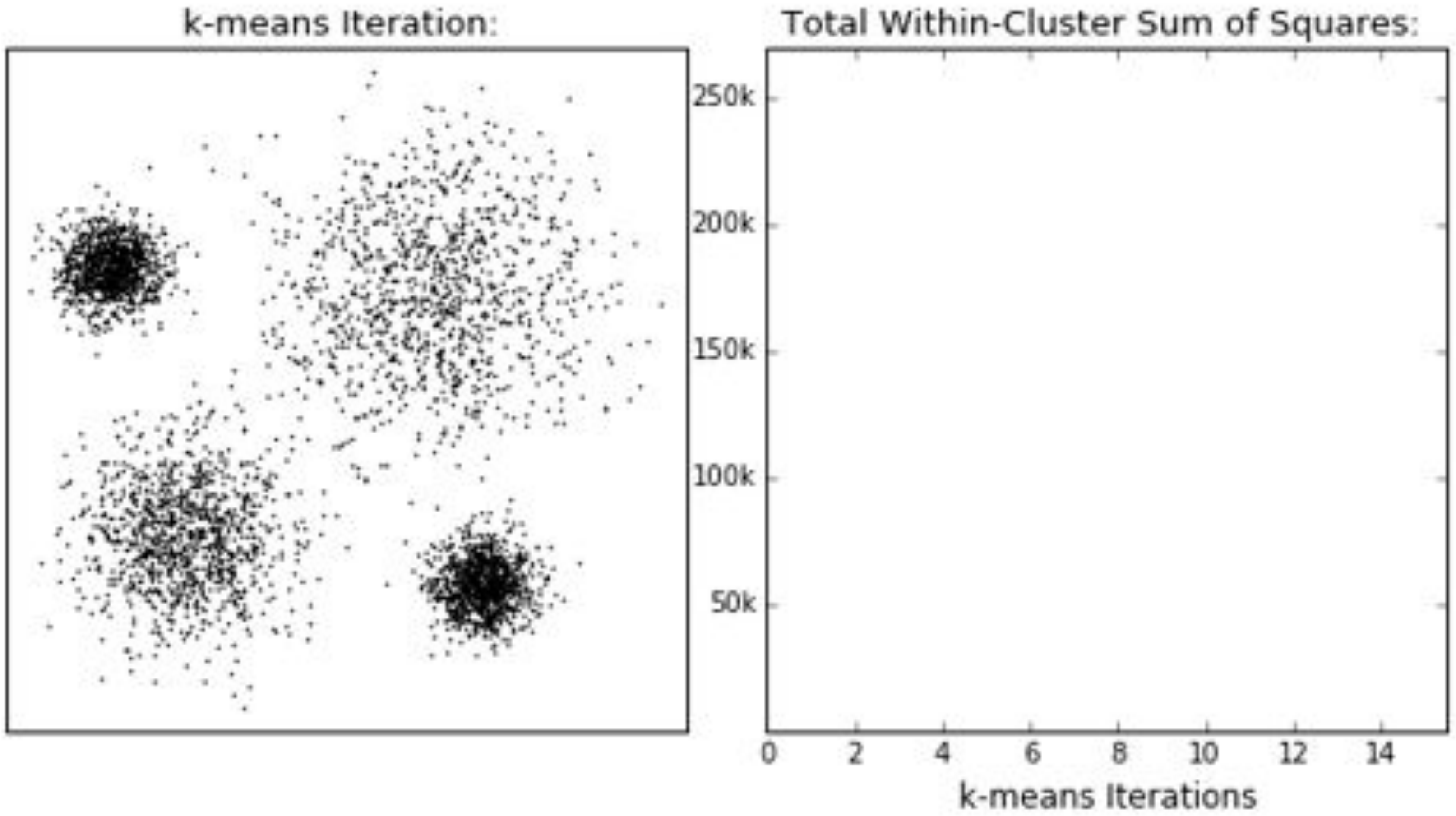
k-means theory



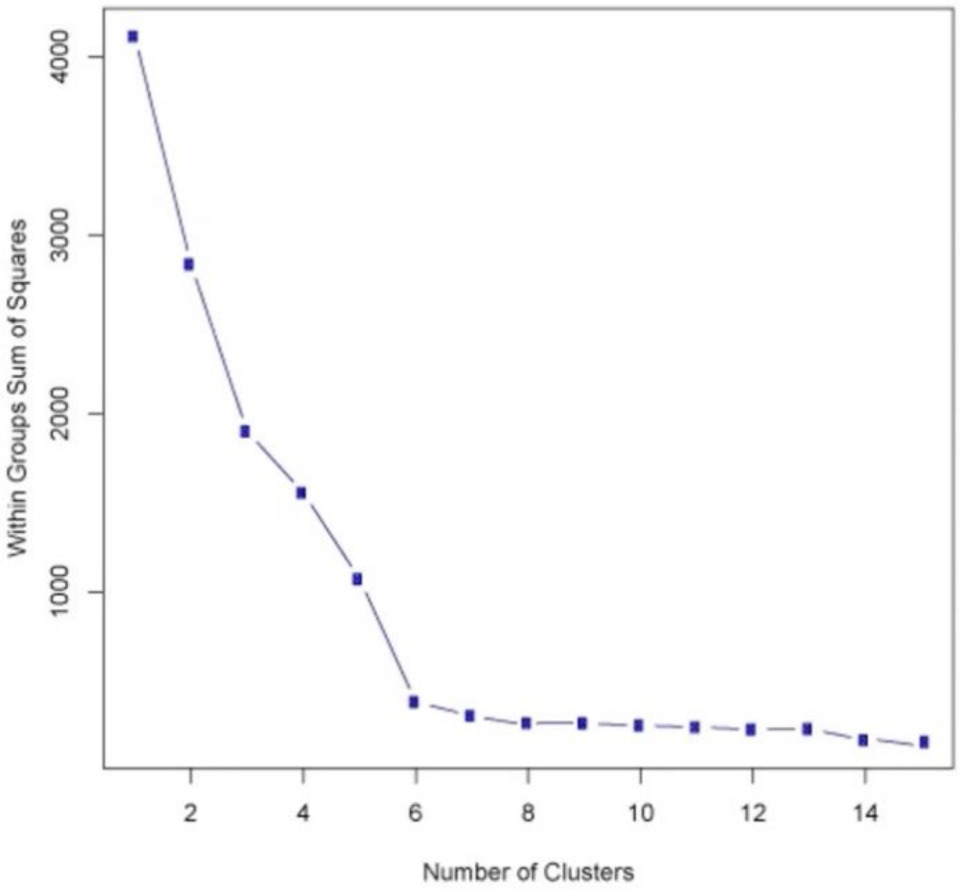
k-means theory



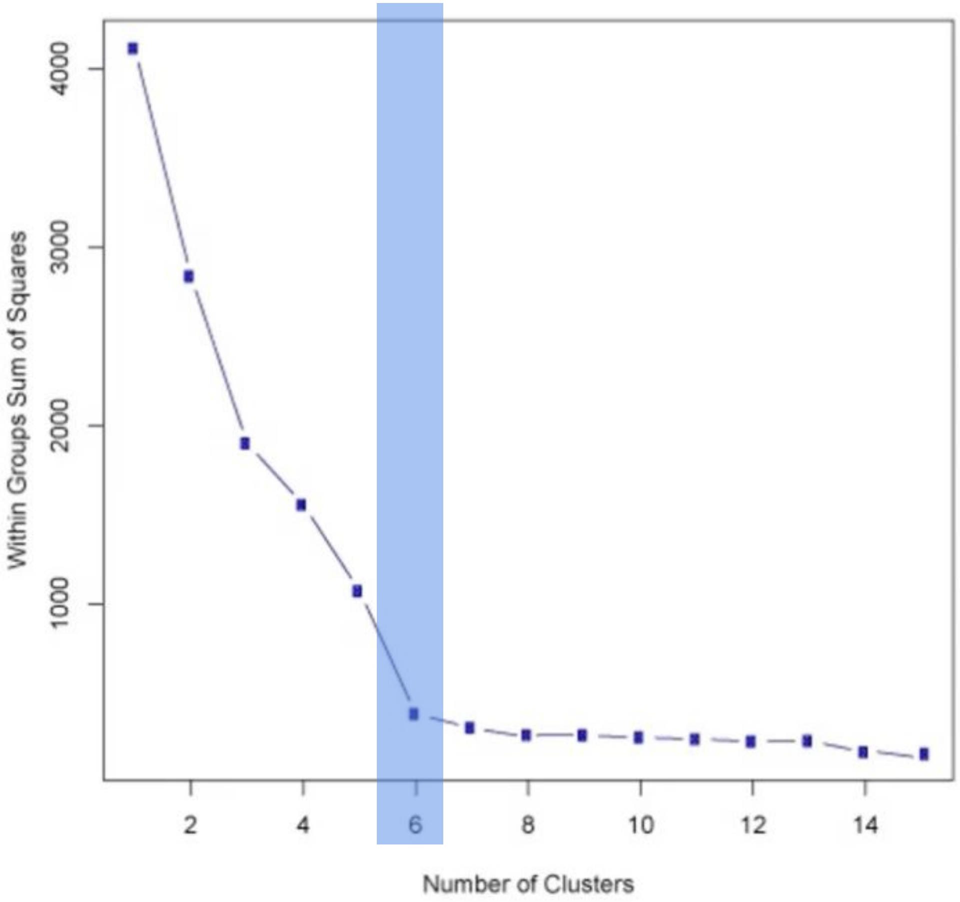
k-means theory



k-means theory



k-means theory



k-means theory