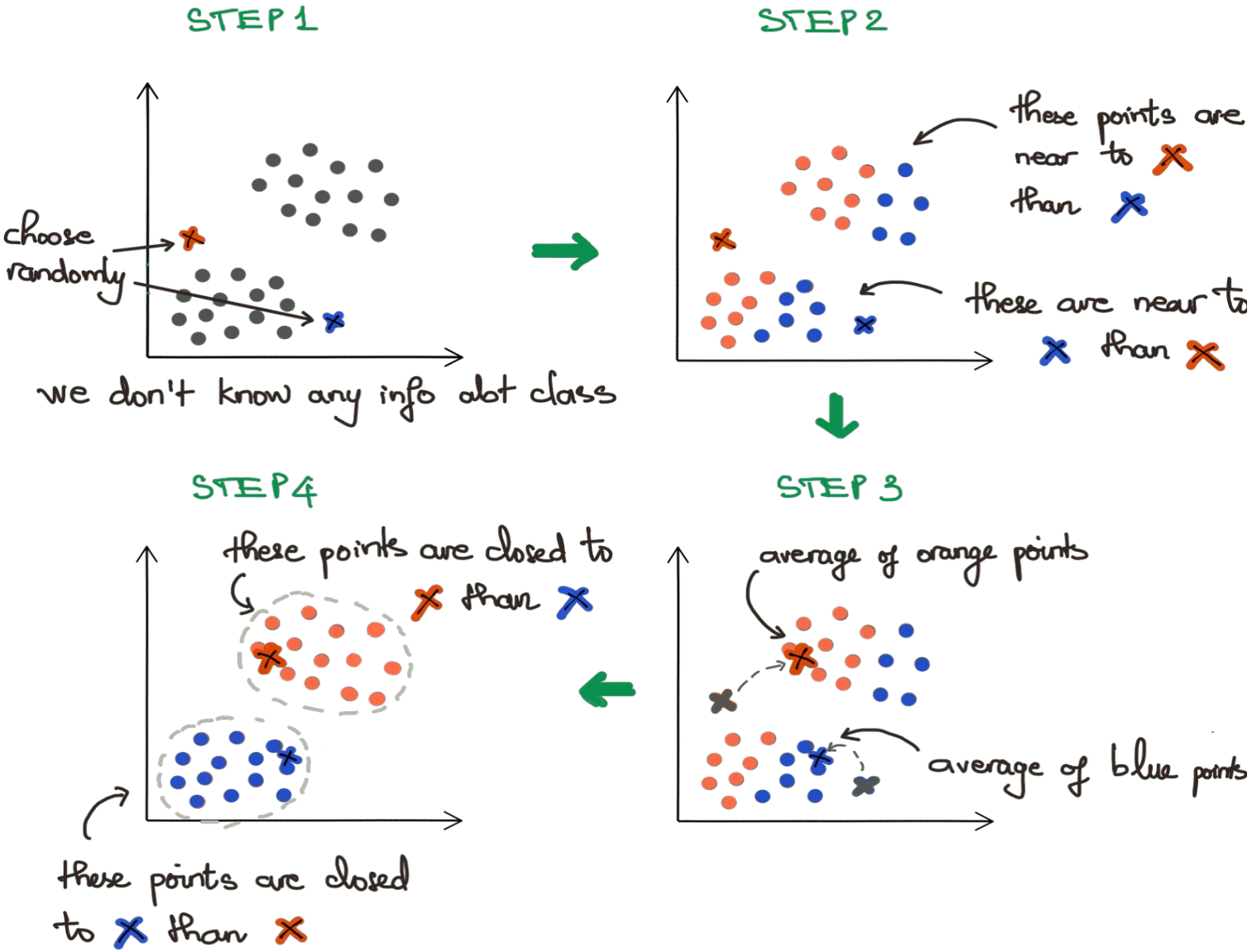
Recap- K-means clustering



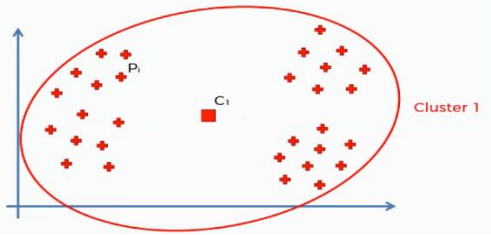
Deciding cluster numbers

WCSS (within-cluster sum of squares) helps us to determine the optimal number of clusters

$$\text{WCSS} = \sum_{P_i \text{ in Cluster 1}} \text{distance}(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} \text{distance}(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} \text{distance}(P_i, C_3)^2$$

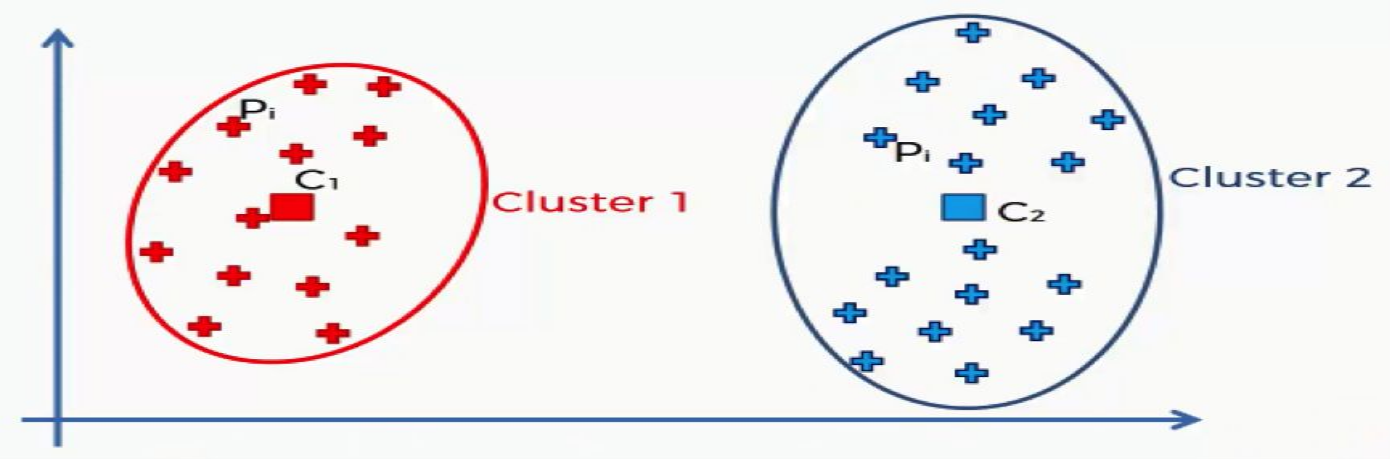
What this equation signifies is this: For Cluster 1, we'll take every point (Pi) that falls within the cluster, and calculate the distance between that point and the centroid (C1) for Cluster 1.

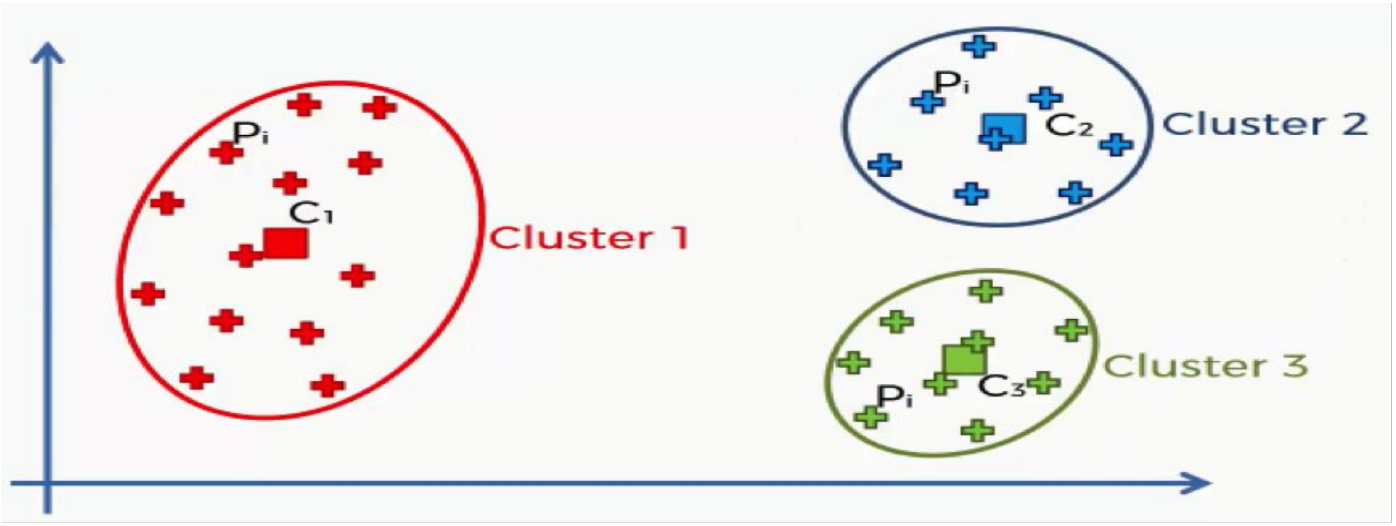
We then square these distances and, finally, calculate the sum of all the squared distances for that cluster. The same is done for all the other clusters. How does this help us in knowing what number of clusters we should use?



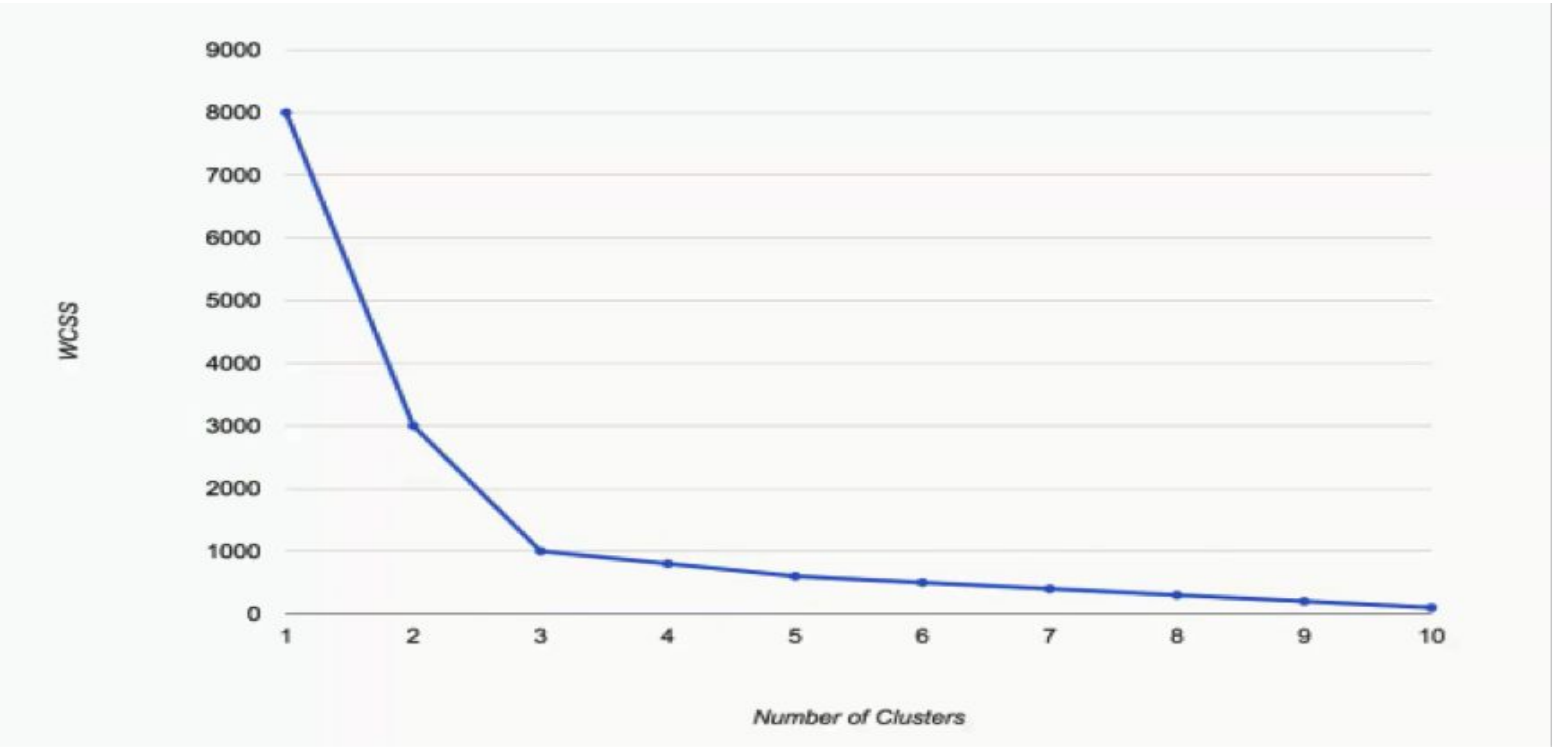
INTERNSHIPSTUDIO

Calculating WCSS

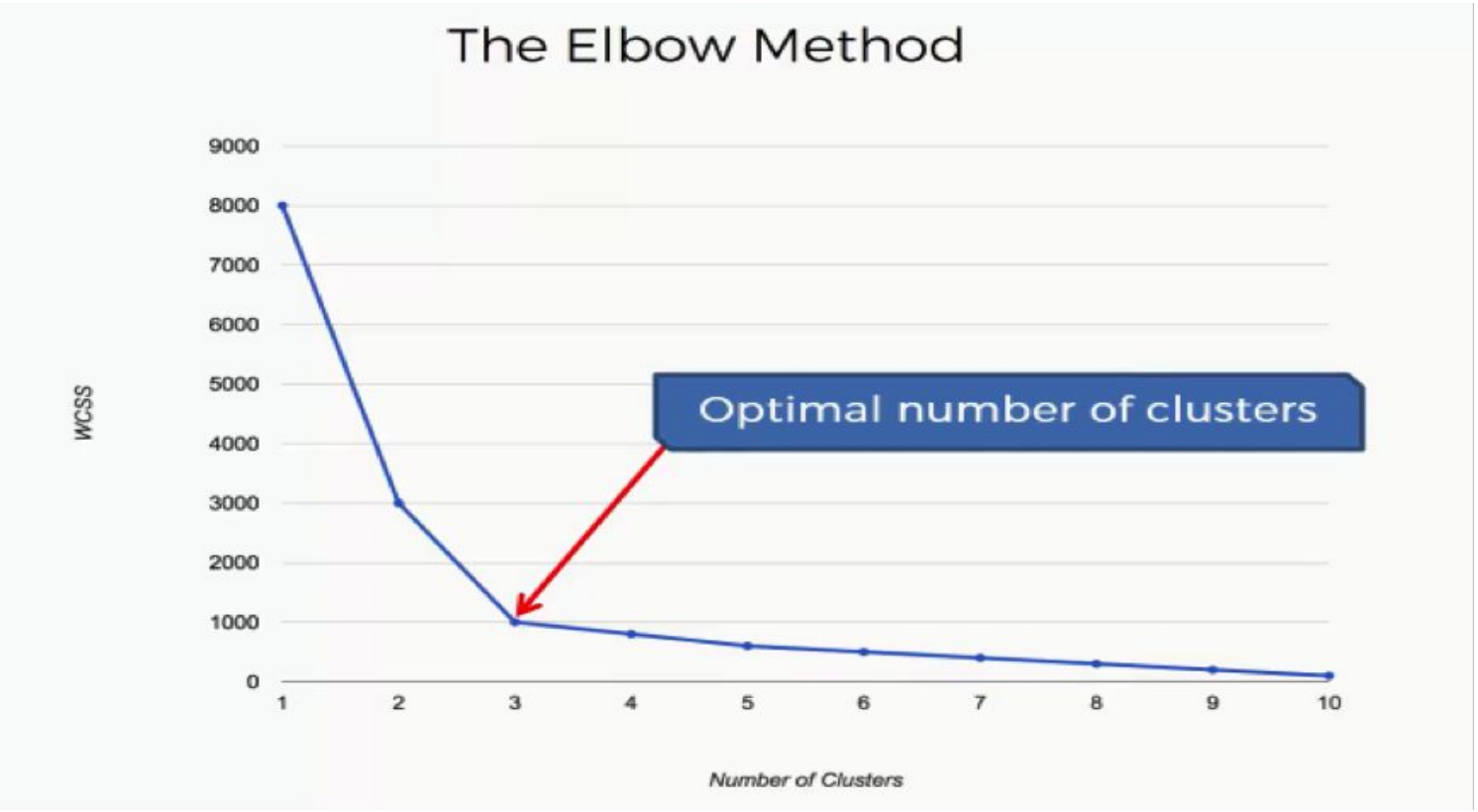




WCSS



- You can see that as we move from one cluster to two, the WCSS takes a massive fall from 8,000 to 3,000.
- As we move from two to three, the WCSS still decreases substantially, from 3,000 to 1,000. From that point on, however, the changes become very minimal, with each cluster only shaving off 200 WCSS points or less.
- That's our hint when it comes to choosing our optimal number of clusters. The keyword is: Elbow.



The "elbow" in your graph will not always be as obvious as in this example. You're likely to have situations where each person would choose a different point, each thinking that theirs is the optimal point.

That's where you have to make your judgment call as a data scientist. The Elbow Method can only give you a hint at where to look.