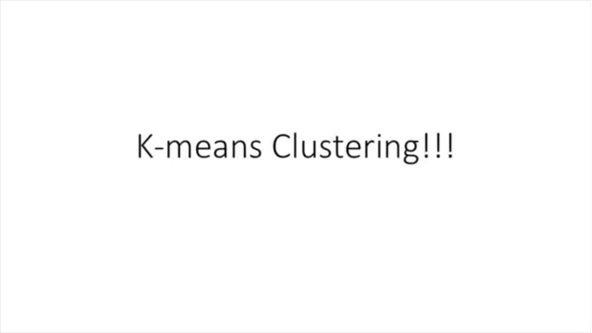
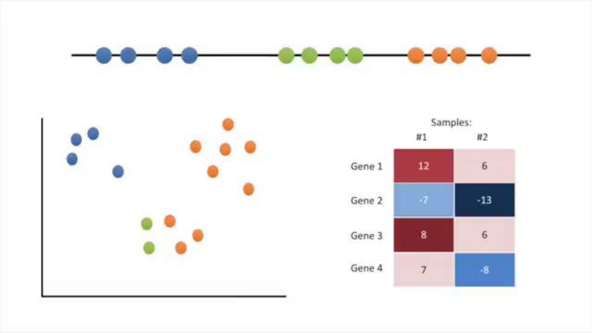


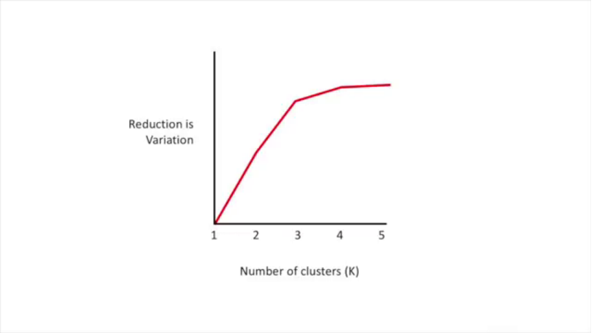
K-means Clustering!!!

1

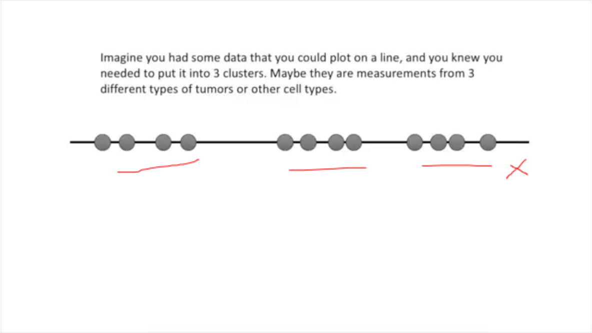


	Samples:	
	#1	#2
Gene 1	12	6
Gene 2	-7	-13
Gene 3	8	6
Gene 4	7	-8

2

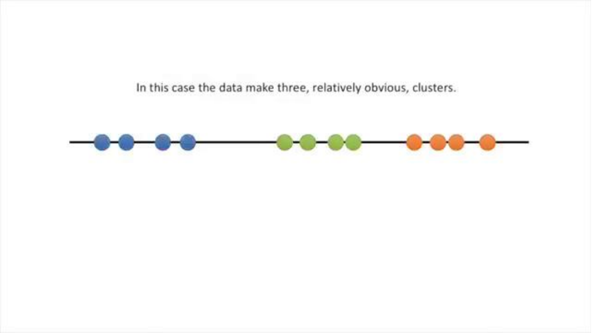


3



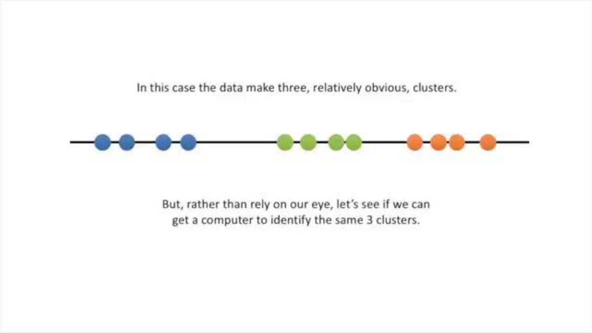
Imagine you had some data that you could plot on a line, and you knew you needed to put it into 3 clusters. Maybe they are measurements from 3 different types of tumors or other cell types.

4



In this case the data make three, relatively obvious, clusters.

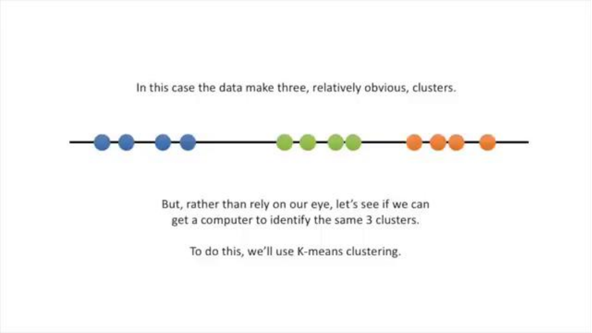
5



In this case the data make three, relatively obvious, clusters.

But, rather than rely on our eye, let's see if we can get a computer to identify the same 3 clusters.

6

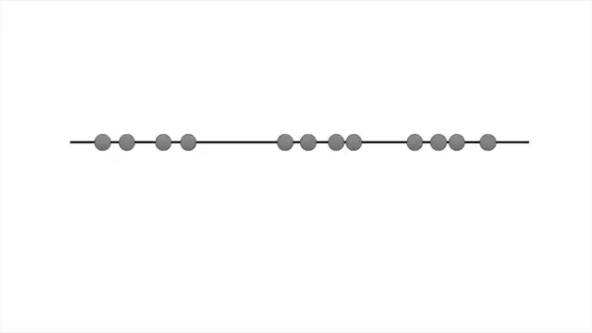


In this case the data make three, relatively obvious, clusters.

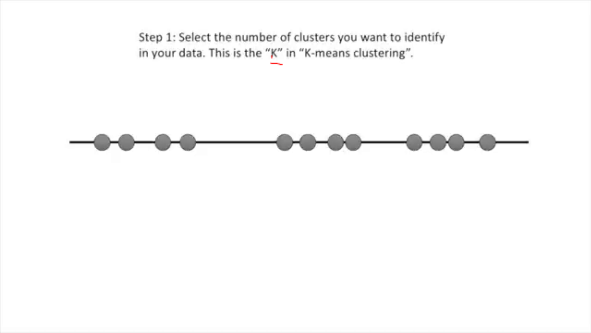
But, rather than rely on our eye, let's see if we can get a computer to identify the same 3 clusters.

To do this, we'll use K-means clustering.

7

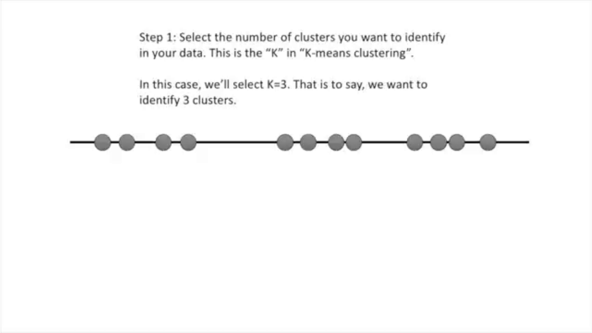


8



Step 1: Select the number of clusters you want to identify in your data. This is the "K" in "K-means clustering".

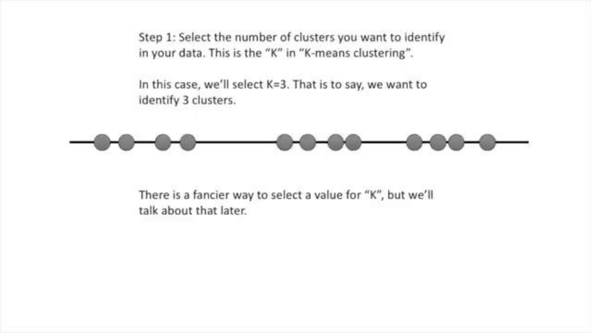
9



Step 1: Select the number of clusters you want to identify in your data. This is the "K" in "K-means clustering".

In this case, we'll select K=3. That is to say, we want to identify 3 clusters.

10

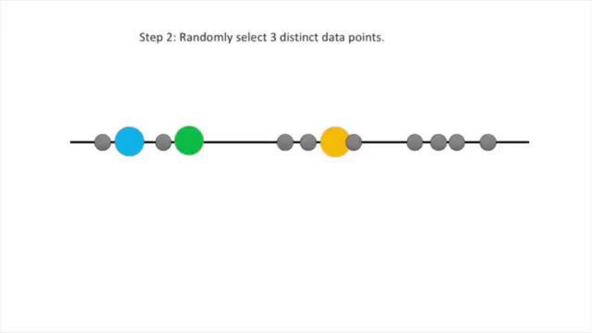


Step 1: Select the number of clusters you want to identify in your data. This is the "K" in "K-means clustering".

In this case, we'll select K=3. That is to say, we want to identify 3 clusters.

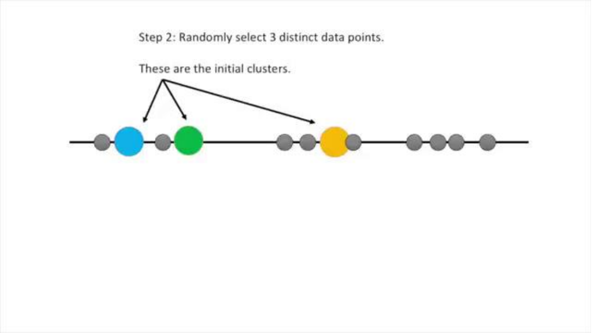
There is a fancier way to select a value for "K", but we'll talk about that later.

11



Step 2: Randomly select 3 distinct data points.

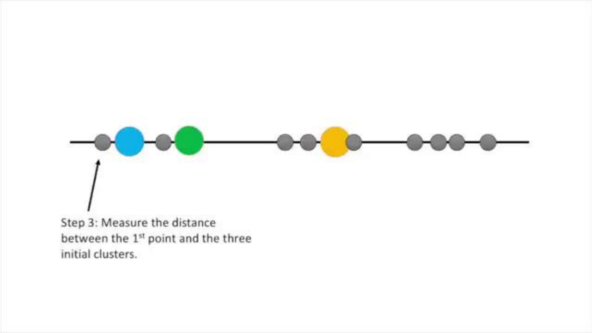
12



Step 2: Randomly select 3 distinct data points.

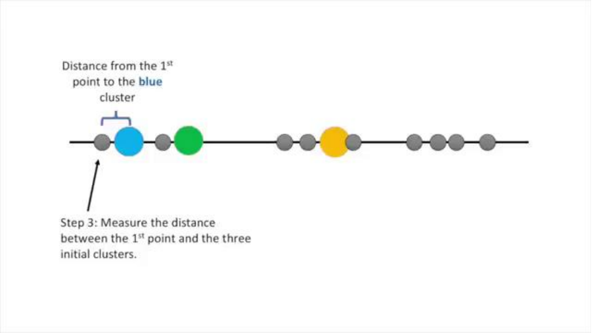
These are the initial clusters.

13



Step 3: Measure the distance between the 1st point and the three initial clusters.

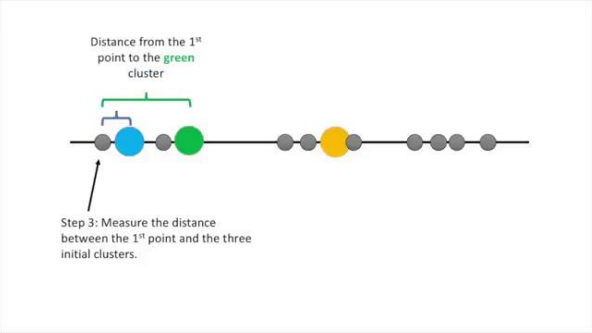
14



Distance from the 1st point to the **blue** cluster

Step 3: Measure the distance between the 1st point and the three initial clusters.

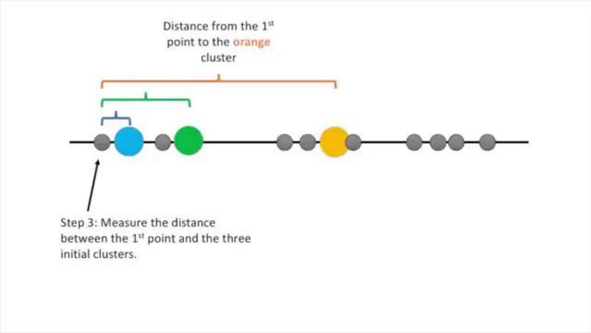
15



Distance from the 1st point to the **green** cluster

Step 3: Measure the distance between the 1st point and the three initial clusters.

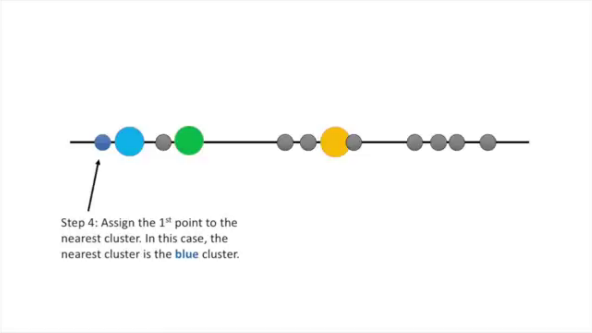
16



Distance from the 1st point to the orange cluster

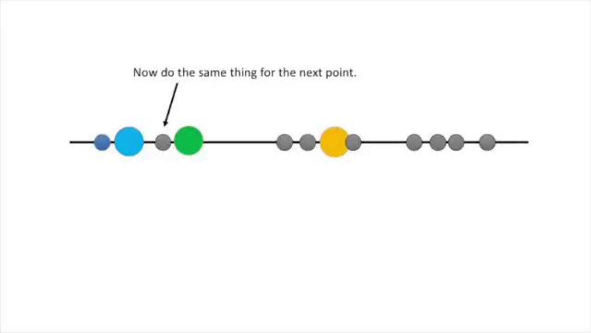
Step 3: Measure the distance between the 1st point and the three initial clusters.

17



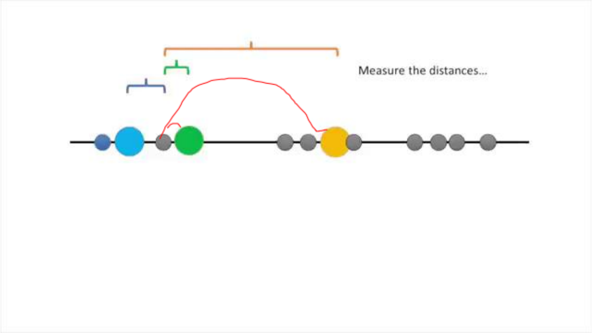
Step 4: Assign the 1st point to the nearest cluster. In this case, the nearest cluster is the **blue** cluster.

18



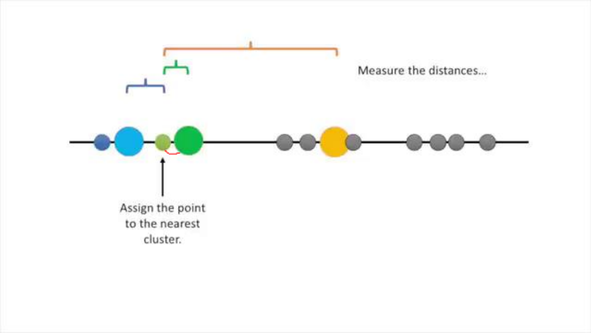
Now do the same thing for the next point.

19



Measure the distances…

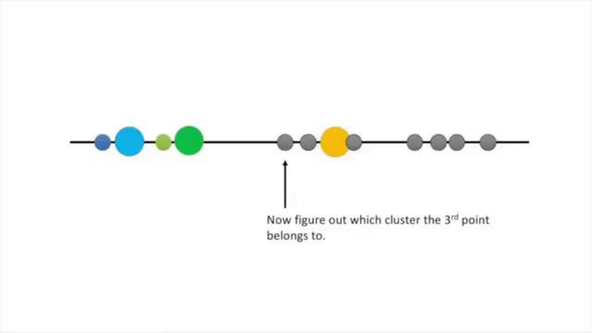
20



Measure the distances…

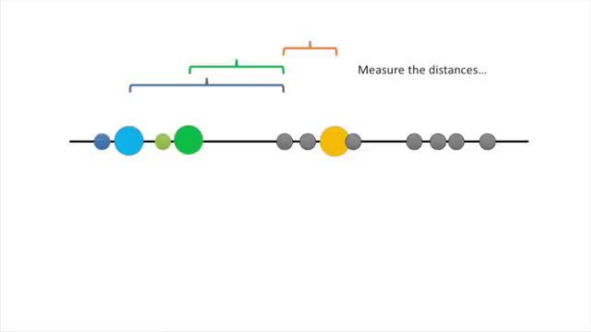
Assign the point to the nearest cluster.

21



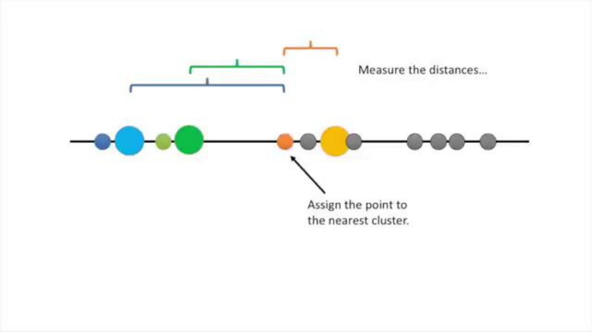
Now figure out which cluster the 3rd point belongs to.

22



Measure the distances…

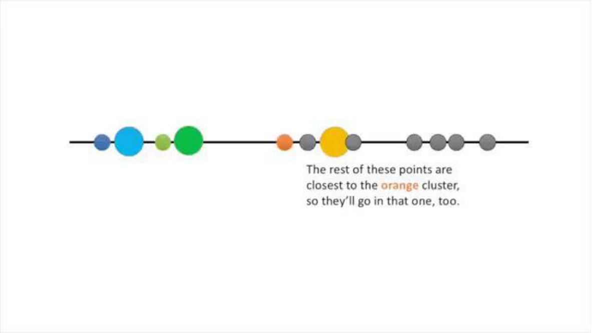
23



Measure the distances…

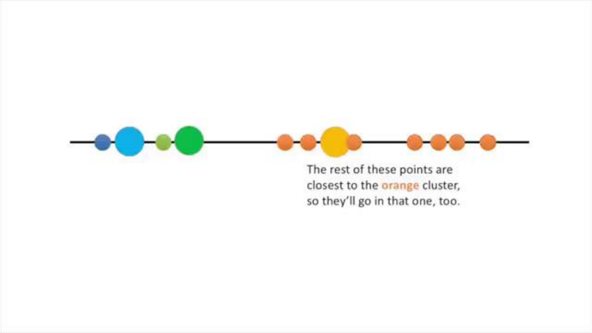
Assign the point to the nearest cluster.

24



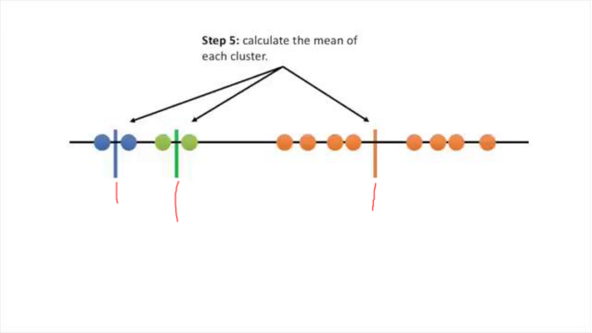
The rest of these points are closest to the orange cluster, so they'll go in that one, too.

25



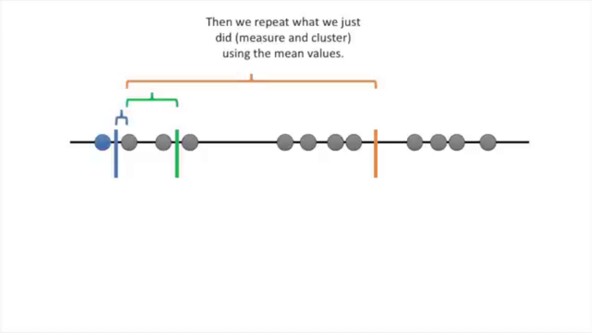
The rest of these points are closest to the orange cluster, so they'll go in that one, too.

26



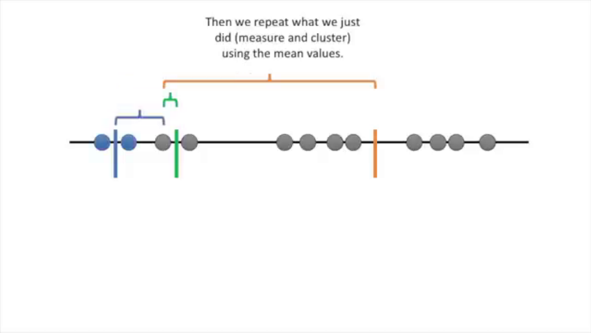
Step 5: calculate the mean of each cluster.

27



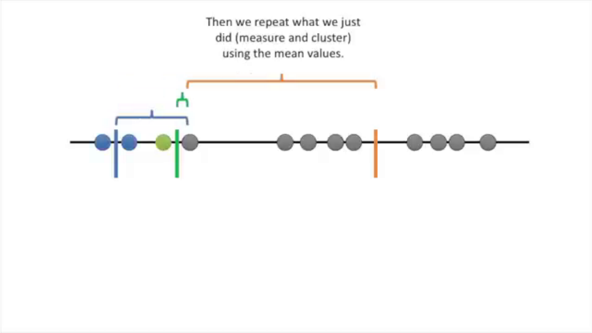
Then we repeat what we just did (measure and cluster) using the mean values.

28



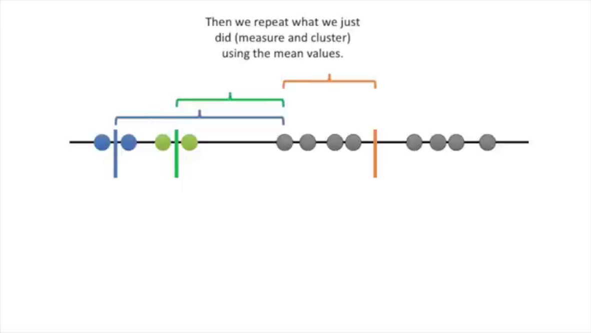
Then we repeat what we just did (measure and cluster) using the mean values.

29



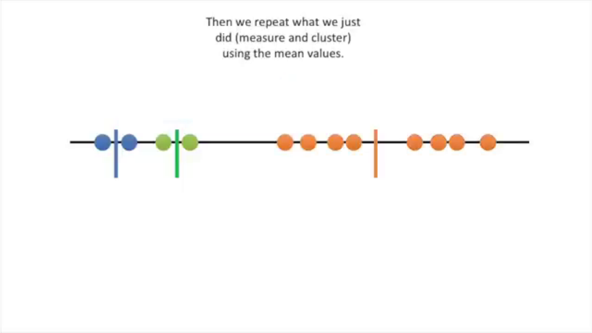
Then we repeat what we just did (measure and cluster) using the mean values.

30



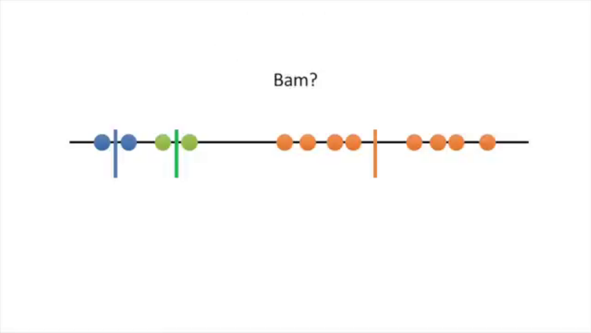
Then we repeat what we just did (measure and cluster) using the mean values.

31



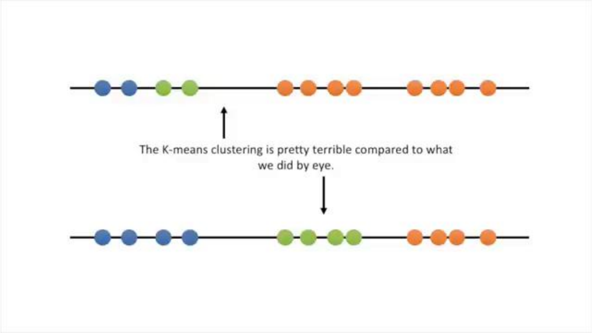
Then we repeat what we just did (measure and cluster) using the mean values.

32



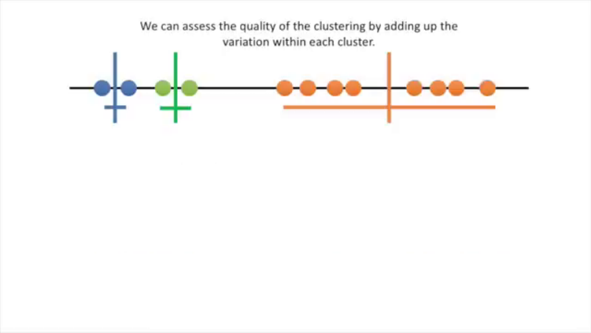
Bam?

33



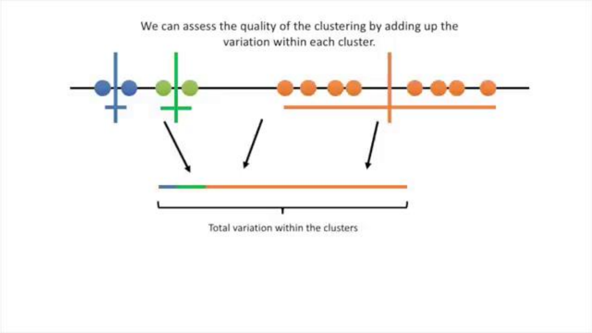
The K-means clustering is pretty terrible compared to what we did by eye.

34



We can assess the quality of the clustering by adding up the variation within each cluster.

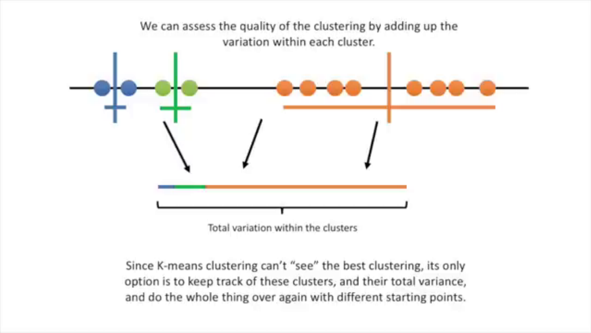
35



We can assess the quality of the clustering by adding up the variation within each cluster.

Total variation within the clusters

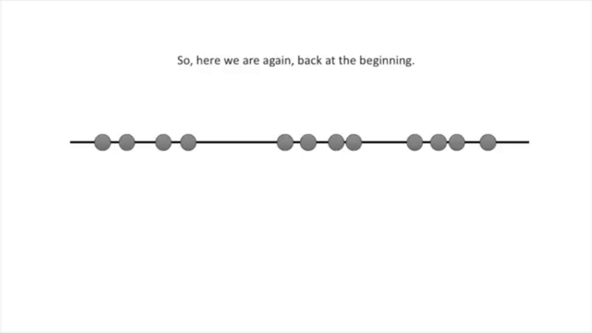
36



We can assess the quality of the clustering by adding up the variation within each cluster.

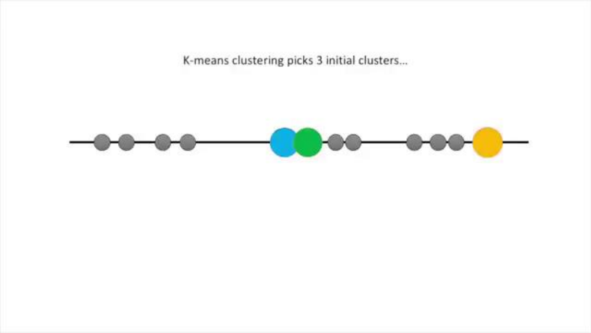
Since K-means clustering can't "see" the best clustering, its only option is to keep track of these clusters, and their total variance, and do the whole thing over again with different starting points.

37



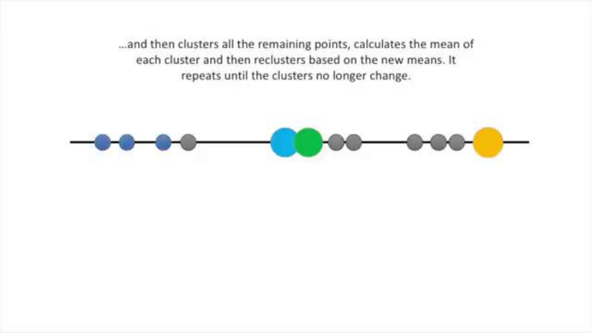
So, here we are again, back at the beginning.

38



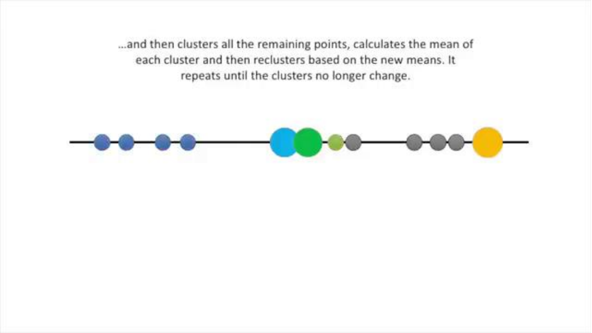
K-means clustering picks 3 initial clusters...

39



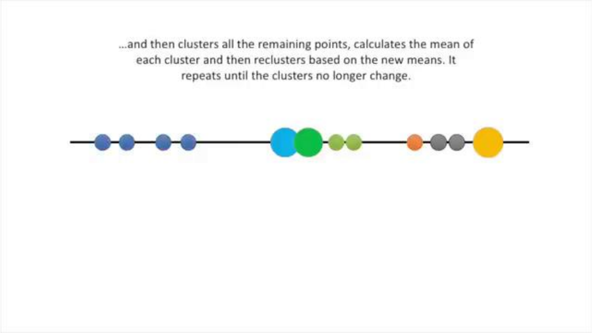
...and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

40



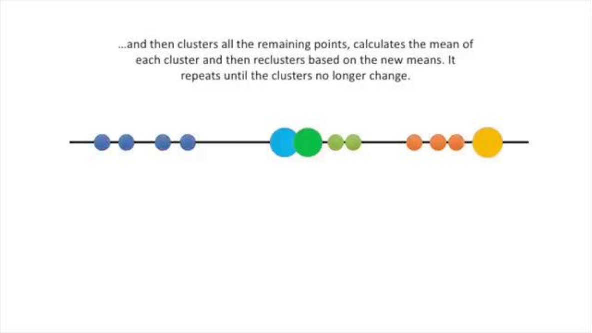
...and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

41



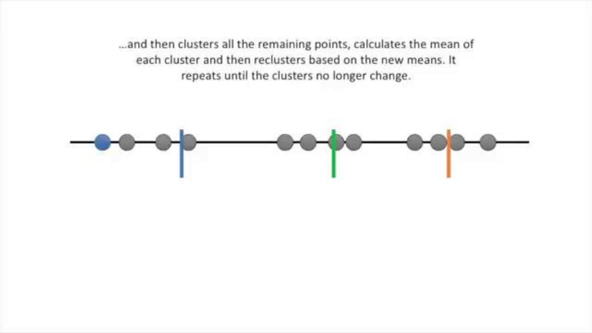
...and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

42



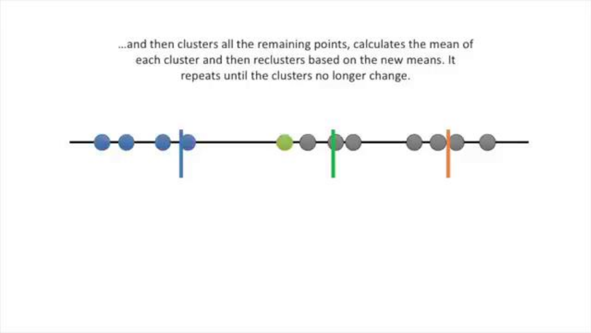
...and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

43



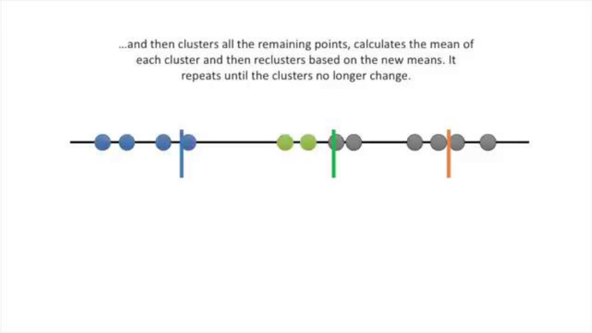
...and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

44



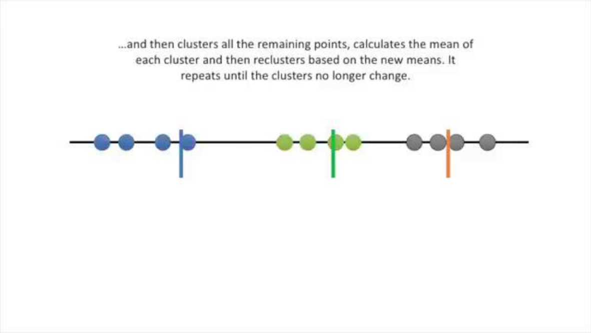
…and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

45



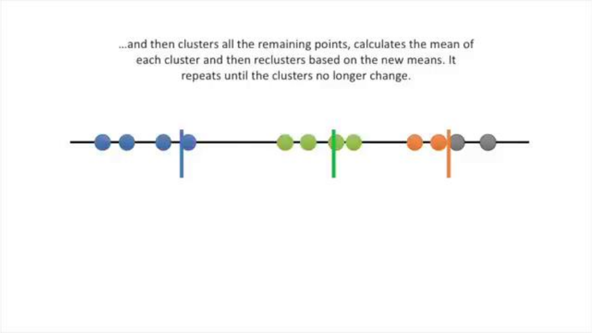
…and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

46



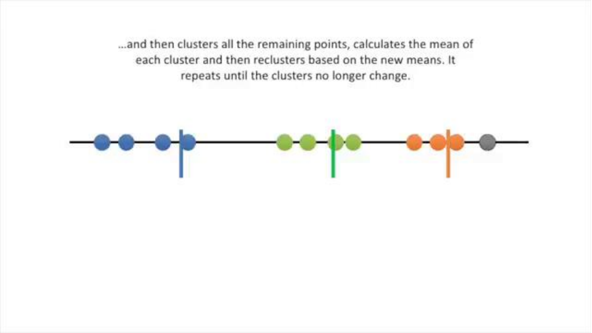
…and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

47



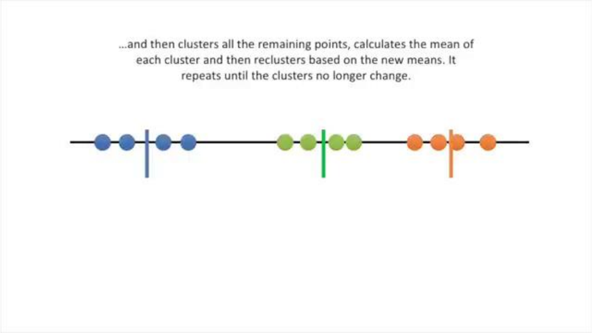
…and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

48



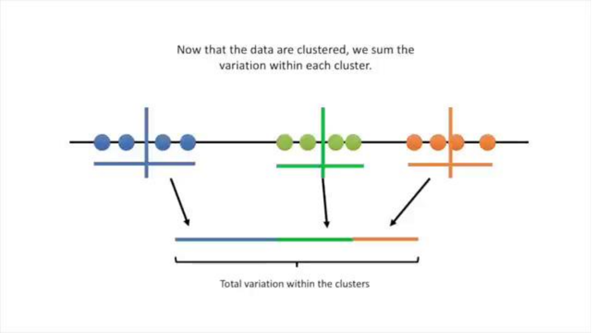
...and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

49



...and then clusters all the remaining points, calculates the mean of each cluster and then reclusters based on the new means. It repeats until the clusters no longer change.

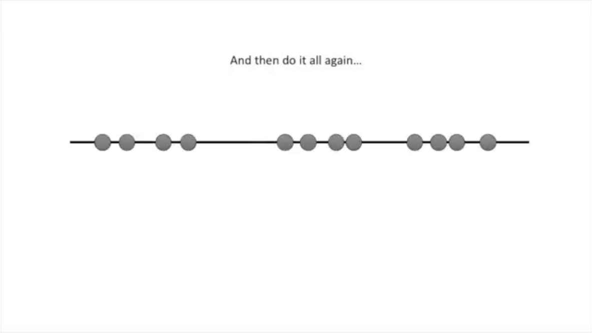
50



Now that the data are clustered, we sum the variation within each cluster.

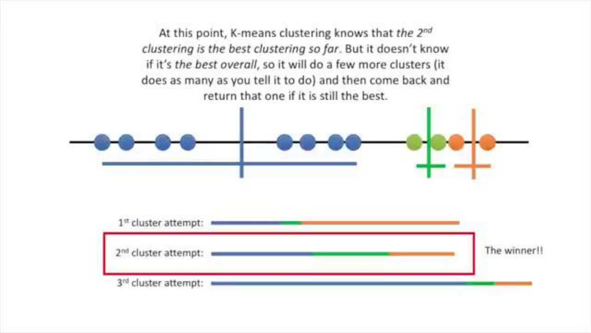
Total variation within the clusters

51



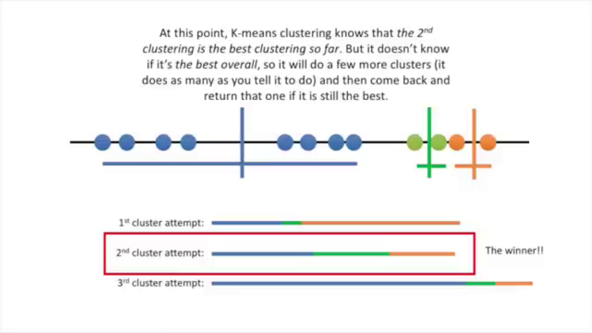
And then do it all again...

52



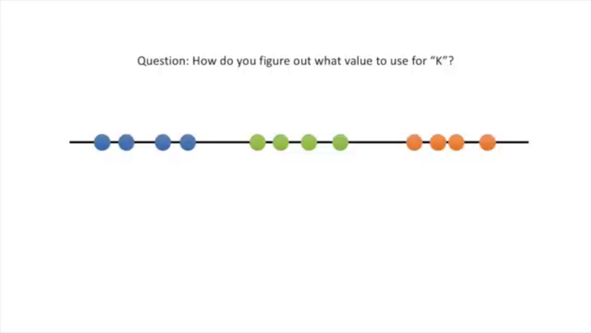
At this point, K-means clustering knows that *the 2nd clustering is the best clustering so far*. But it doesn't know if it's *the best overall*, so it will do a few more clusters (it does as many as you tell it to do) and then come back and return that one if it is still the best.

53



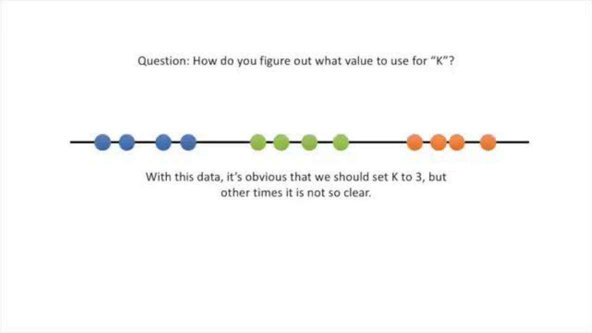
At this point, K-means clustering knows that *the 2nd clustering is the best clustering so far*. But it doesn't know if it's *the best overall*, so it will do a few more clusters (it does as many as you tell it to do) and then come back and return that one if it is still the best.

54



Question: How do you figure out what value to use for "K"?

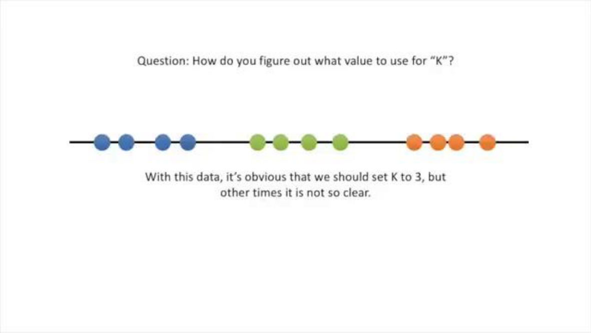
55



Question: How do you figure out what value to use for "K"?

With this data, it's obvious that we should set K to 3, but other times it is not so clear.

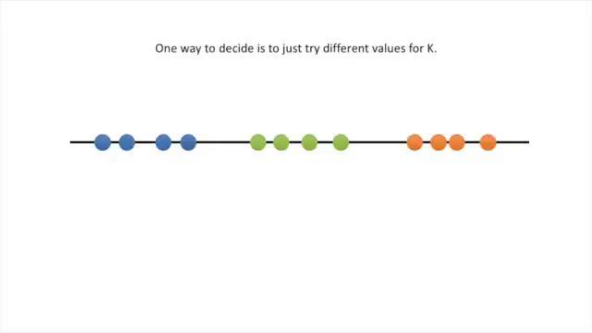
56



Question: How do you figure out what value to use for "K"?

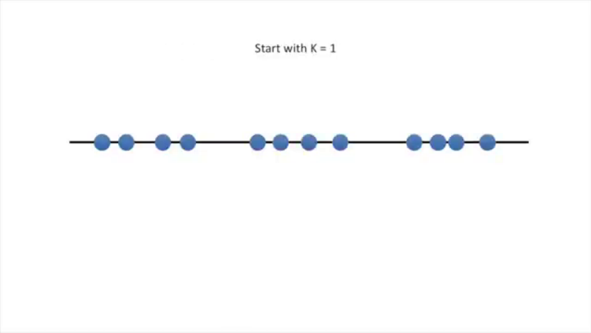
With this data, it's obvious that we should set K to 3, but other times it is not so clear.

57



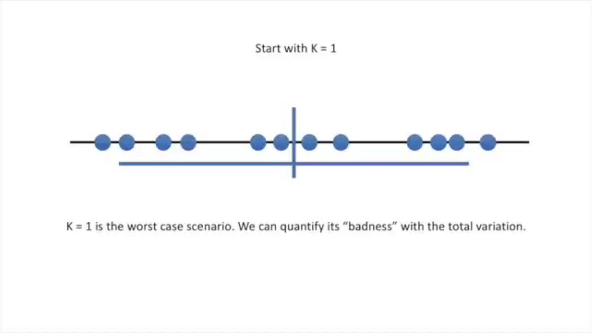
One way to decide is to just try different values for K.

58



Start with K = 1

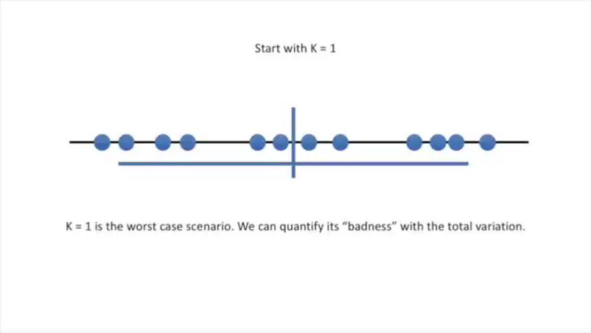
59



Start with K = 1

K = 1 is the worst case scenario. We can quantify its "badness" with the total variation.

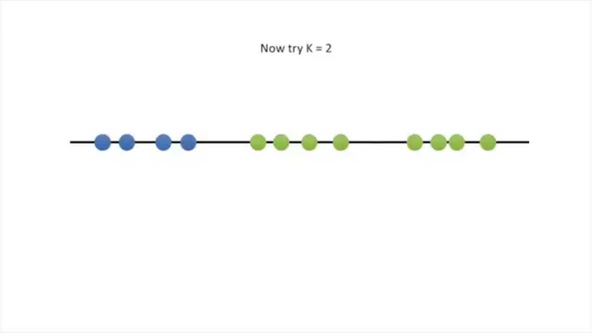
60



Start with K = 1

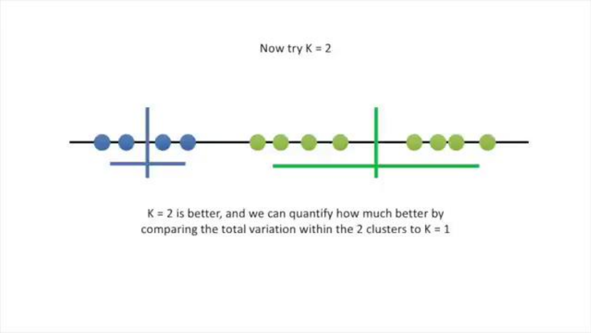
K = 1 is the worst case scenario. We can quantify its “badness” with the total variation.

61



Now try K = 2

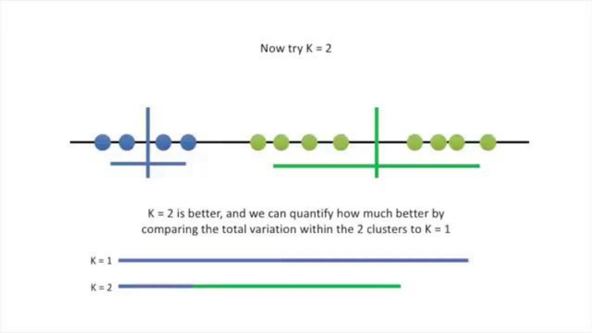
62



Now try K = 2

K = 2 is better, and we can quantify how much better by comparing the total variation within the 2 clusters to K = 1

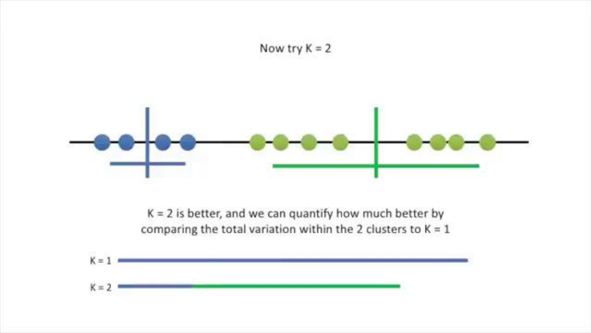
63



Now try K = 2

K = 2 is better, and we can quantify how much better by comparing the total variation within the 2 clusters to K = 1

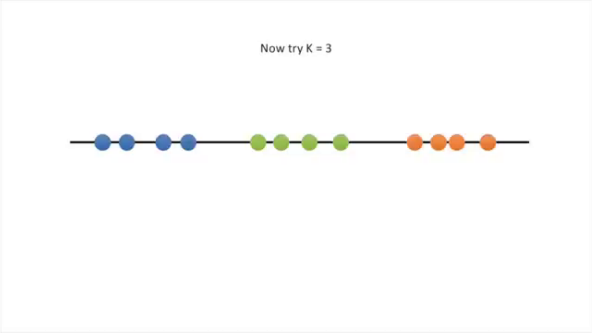
64



Now try K = 2

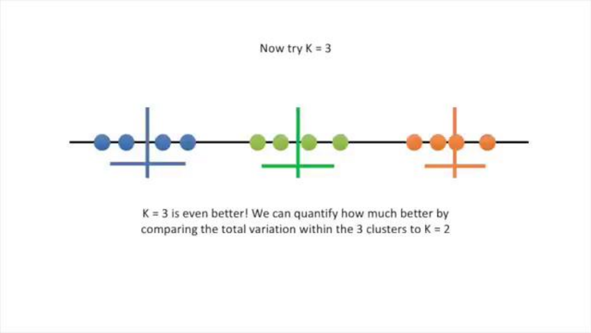
K = 2 is better, and we can quantify how much better by comparing the total variation within the 2 clusters to K = 1

65



Now try K = 3

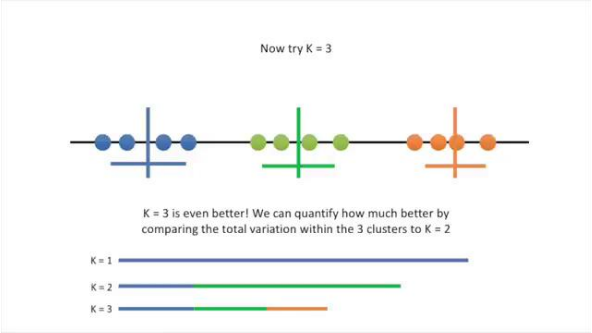
66



Now try K = 3

K = 3 is even better! We can quantify how much better by comparing the total variation within the 3 clusters to K = 2

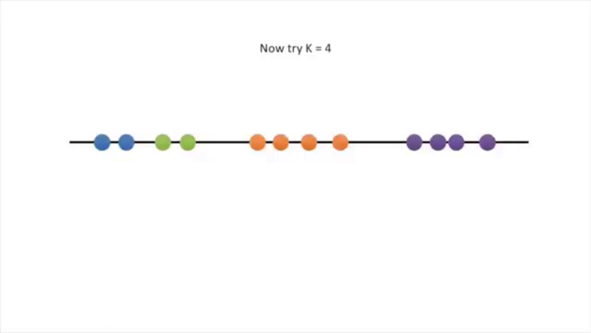
67



Now try K = 3

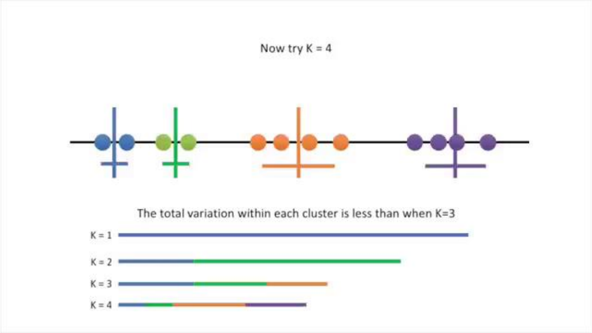
K = 3 is even better! We can quantify how much better by comparing the total variation within the 3 clusters to K = 2

68



Now try K = 4

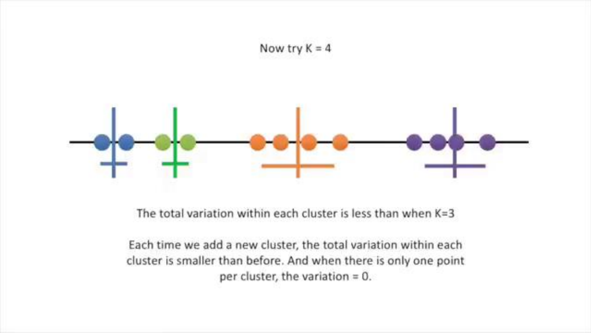
69



Now try K = 4

The total variation within each cluster is less than when K=3

70

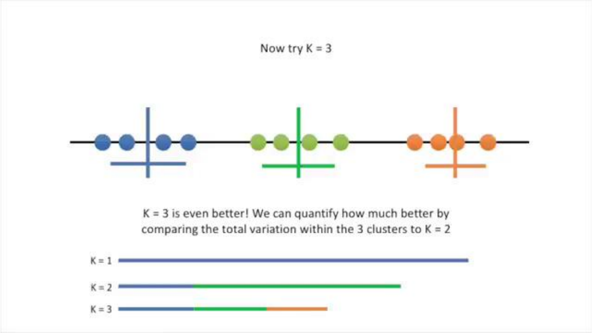


Now try K = 4

The total variation within each cluster is less than when K=3

Each time we add a new cluster, the total variation within each cluster is smaller than before. And when there is only one point per cluster, the variation = 0.

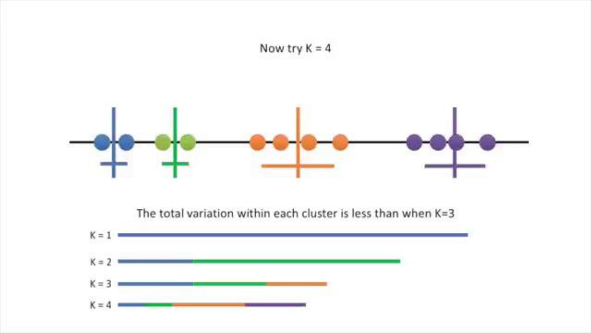
71



Now try K = 3

K = 3 is even better! We can quantify how much better by comparing the total variation within the 3 clusters to K = 2

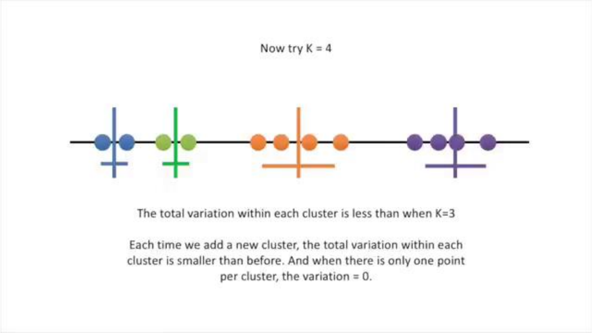
72



Now try K = 4

The total variation within each cluster is less than when K=3

73

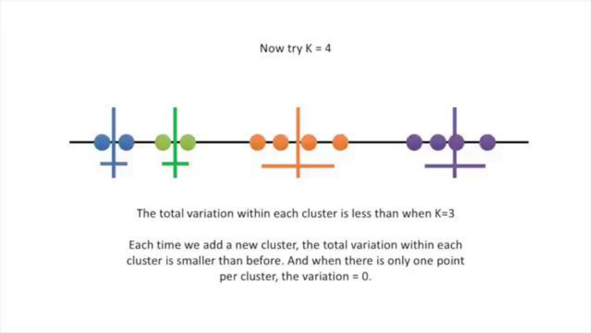


Now try K = 4

The total variation within each cluster is less than when K=3

Each time we add a new cluster, the total variation within each cluster is smaller than before. And when there is only one point per cluster, the variation = 0.

74

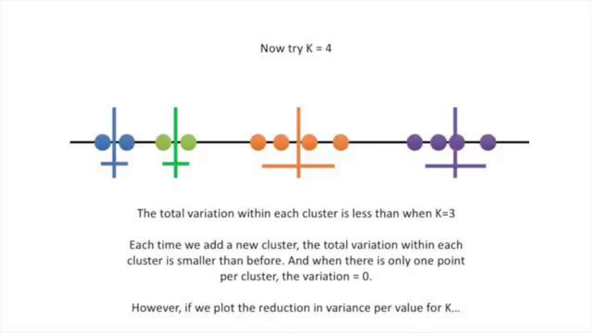


Now try K = 4

The total variation within each cluster is less than when K=3

Each time we add a new cluster, the total variation within each cluster is smaller than before. And when there is only one point per cluster, the variation = 0.

75



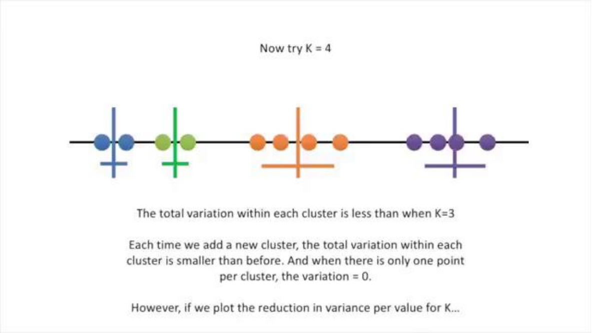
Now try K = 4

The total variation within each cluster is less than when K=3

Each time we add a new cluster, the total variation within each cluster is smaller than before. And when there is only one point per cluster, the variation = 0.

However, if we plot the reduction in variance per value for K…

76



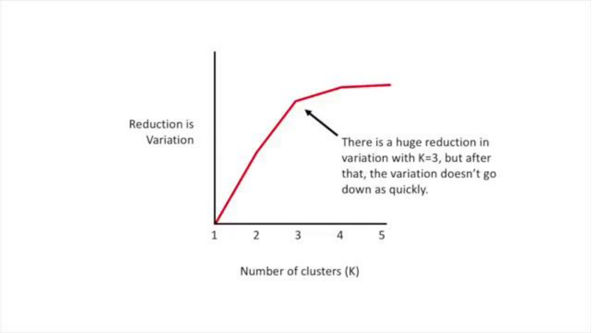
Now try K = 4

The total variation within each cluster is less than when K=3

Each time we add a new cluster, the total variation within each cluster is smaller than before. And when there is only one point per cluster, the variation = 0.

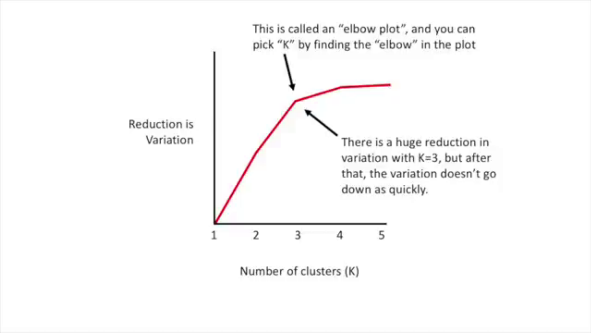
However, if we plot the reduction in variance per value for K…

77



There is a huge reduction in variation with K=3, but after that, the variation doesn't go down as quickly.

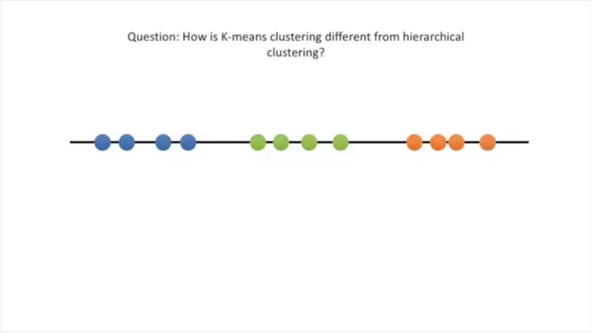
78



This is called an "elbow plot", and you can pick "K" by finding the "elbow" in the plot

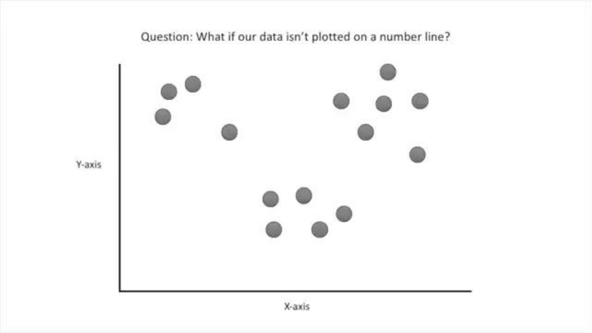
There is a huge reduction in variation with K=3, but after that, the variation doesn't go down as quickly.

79

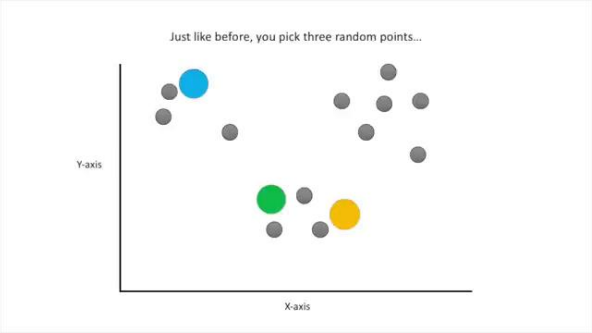


Question: How is K-means clustering different from hierarchical clustering?

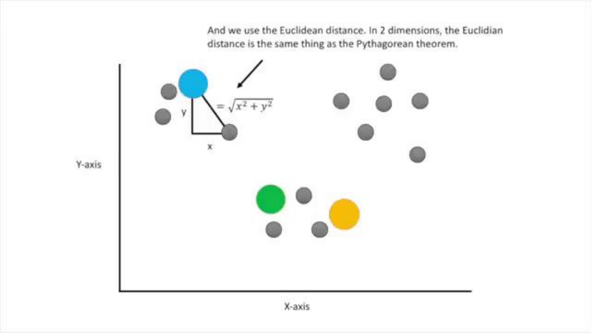
80


Question: What if our data isn't plotted on a number line?

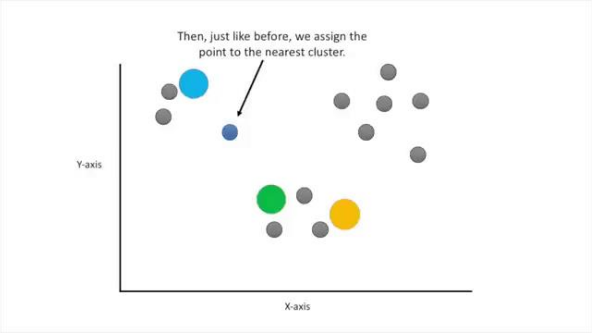
81


Just like before, you pick three random points…

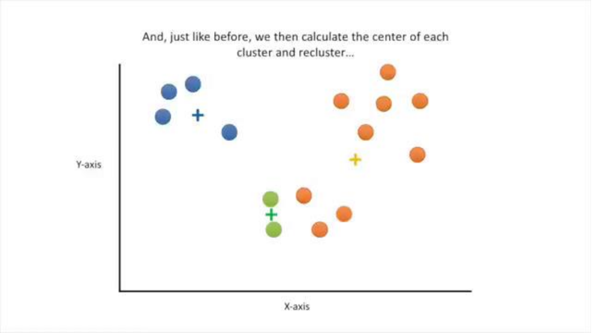
82


And we use the Euclidean distance. In 2 dimensions, the Euclidian distance is the same thing as the Pythagorean theorem.

83

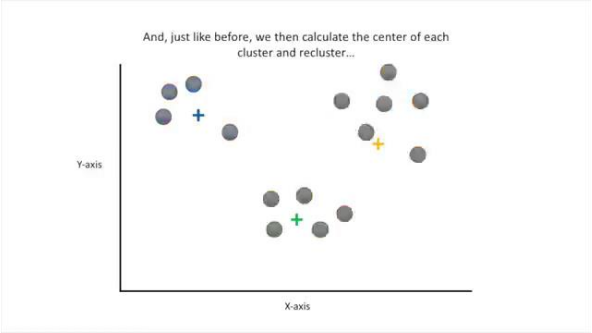

Then, just like before, we assign the point to the nearest cluster.

84



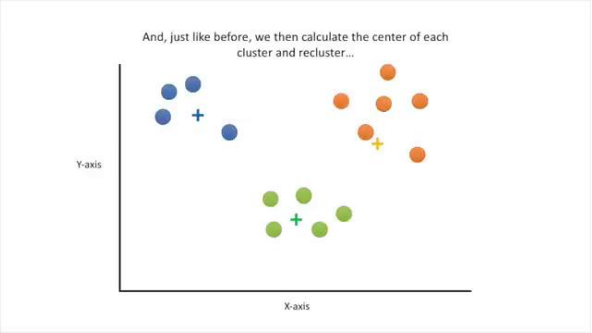
And, just like before, we then calculate the center of each cluster and recluster…

85



And, just like before, we then calculate the center of each cluster and recluster…

86



And, just like before, we then calculate the center of each cluster and recluster…

87