Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

DATA-DRIVEN TRANSFORMATION

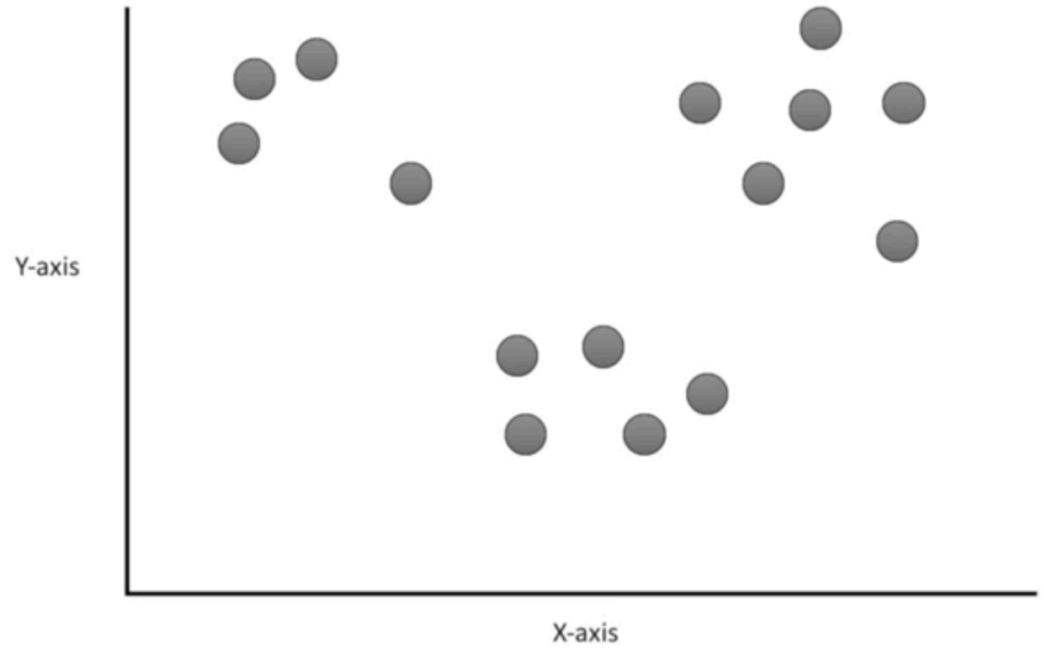
# INTRODUCTION TO CLUSTERING

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

# K-MEANS CLUSTERING

K-means clustering finds **K groups** of **data points close to each other** when visualizing the data.



# ALGORITHM

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

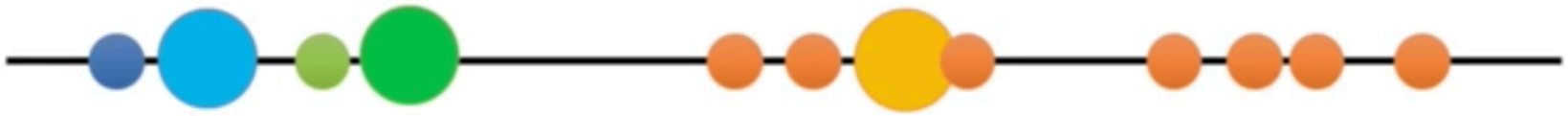
In this case, we’ll select K=3. That is to say, we want to identify 3 clusters.

https://www.youtube.com/watch?v=4b5d3muPQmA

# ALGORITHM

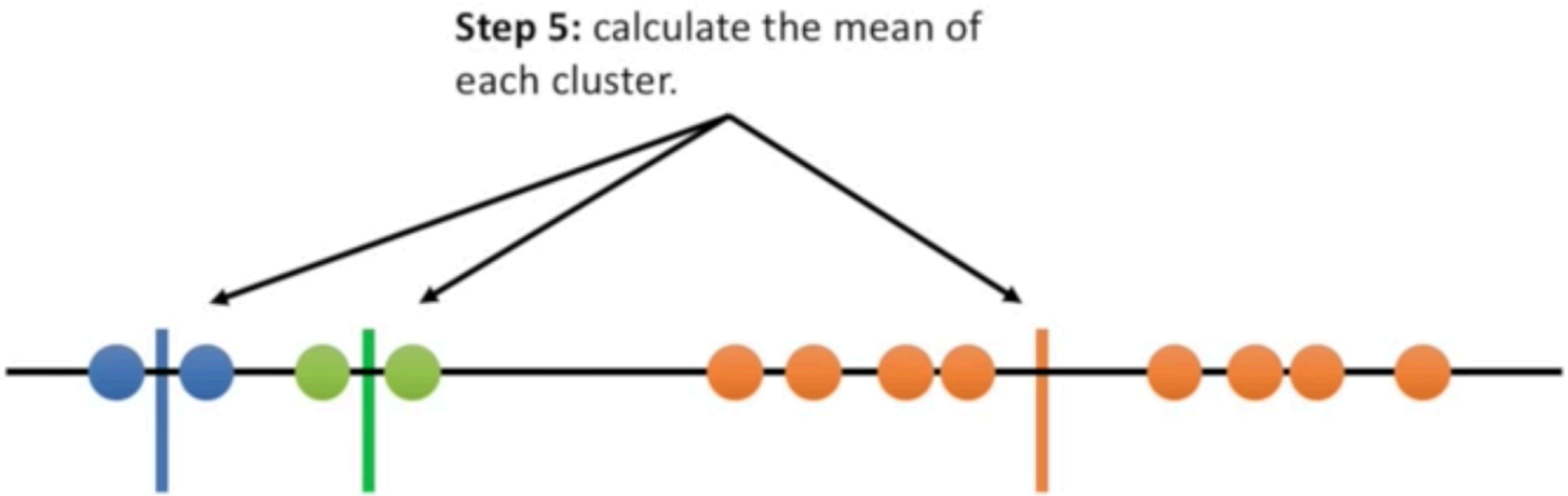
Step 2: Randomly select 3 distinct data points.

# ALGORITHM



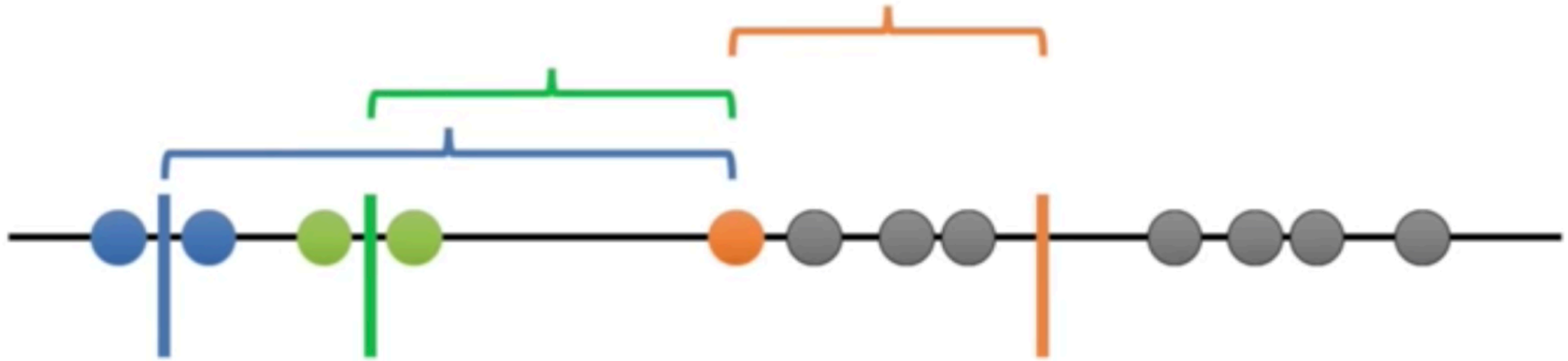
these points are
closest to the orange cluster,
so they'll go in that one

# ALGORITHM

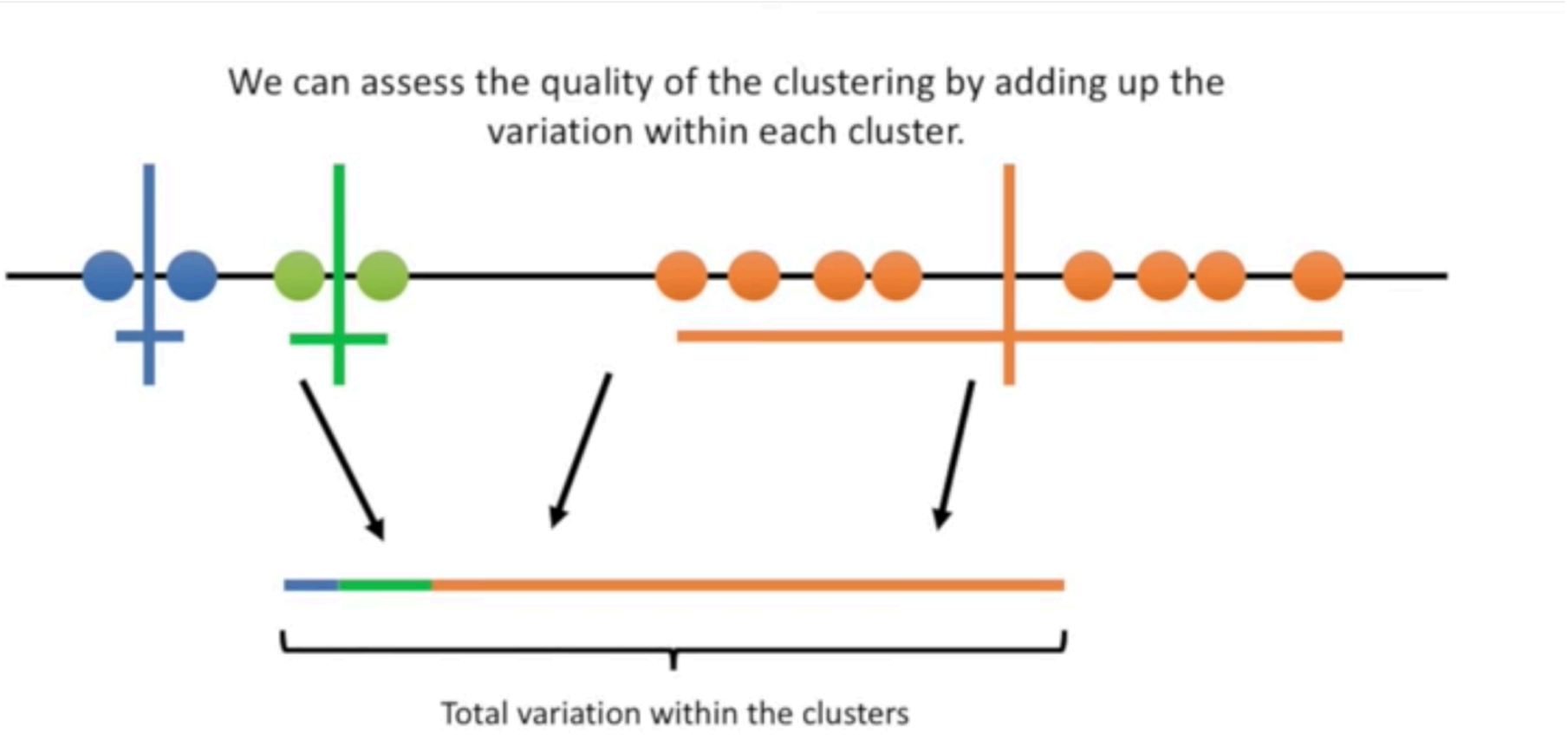


# ALGORITHM

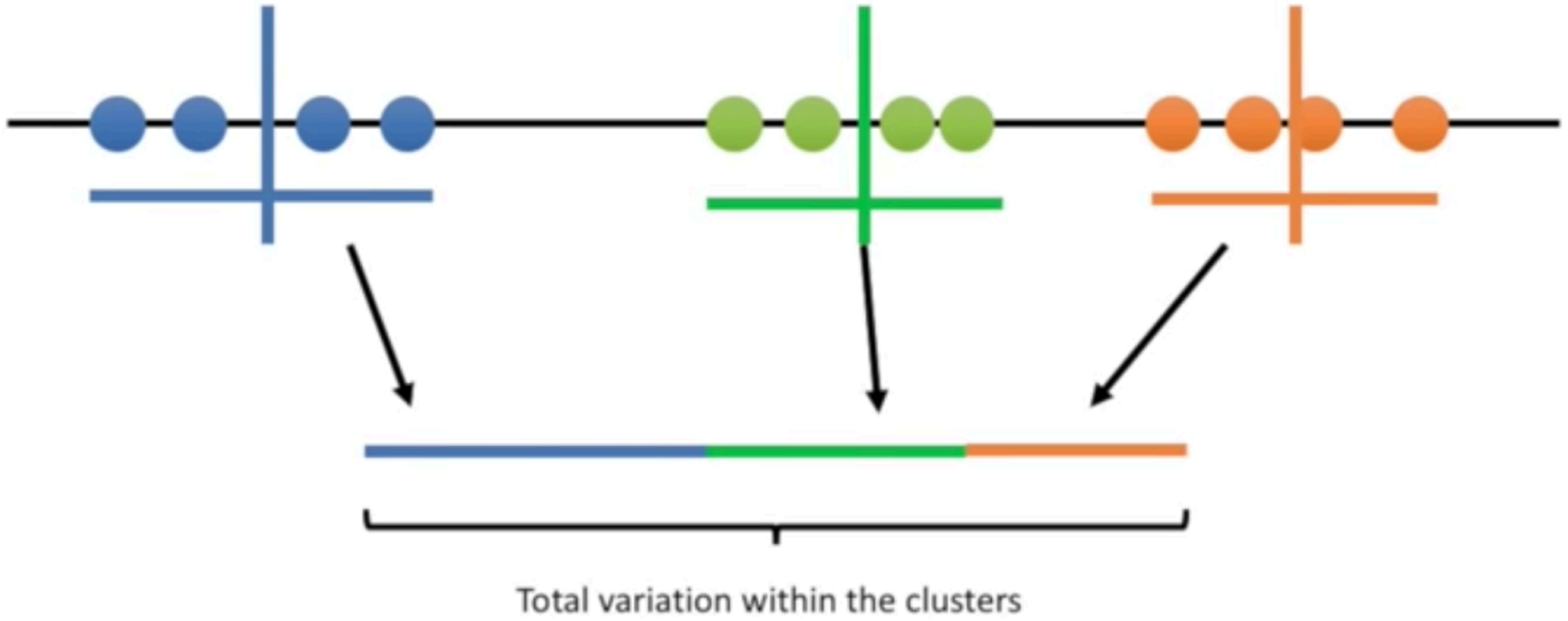
Then we repeat what we just did (measure and cluster) using the mean values.



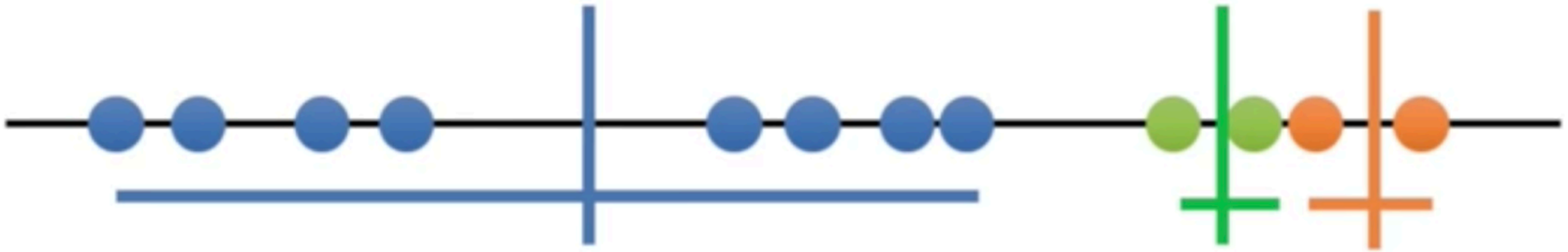
# ALGORITHM


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

# ALGORITHM



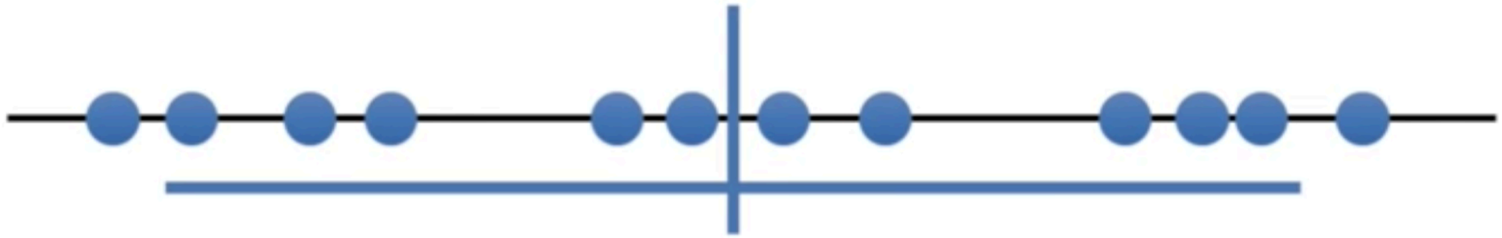
# ALGORITHM



3rd cluster attempt:

# HOW TO CHOOSE K ?

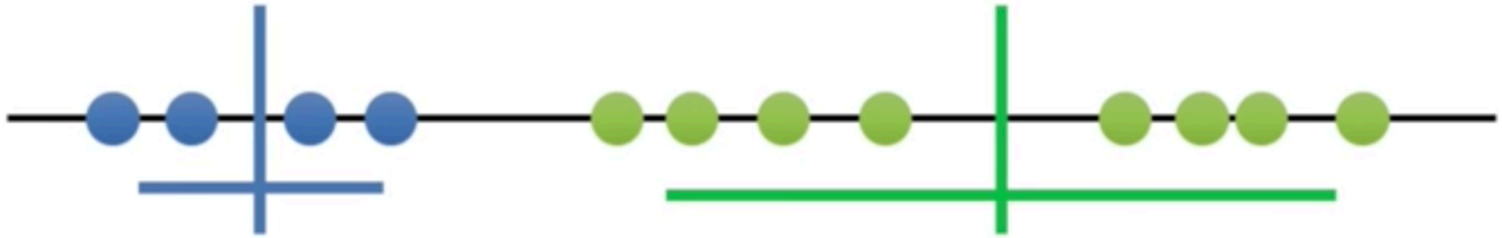
Start with K = 1



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

# HOW TO CHOOSE K ?

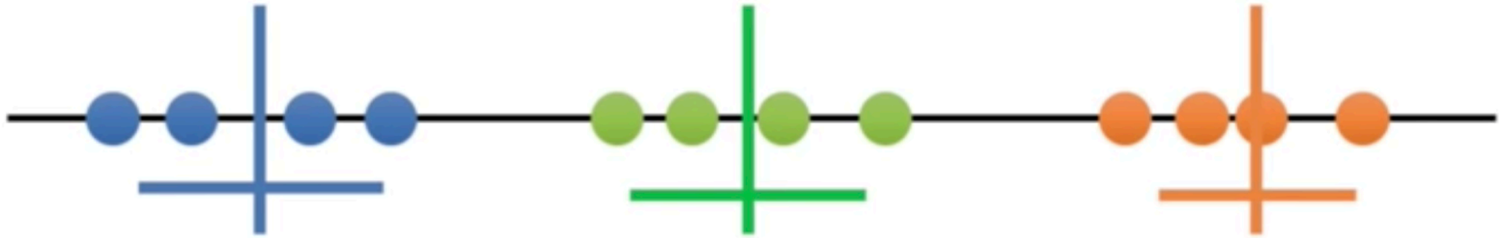
Now try K = 2



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

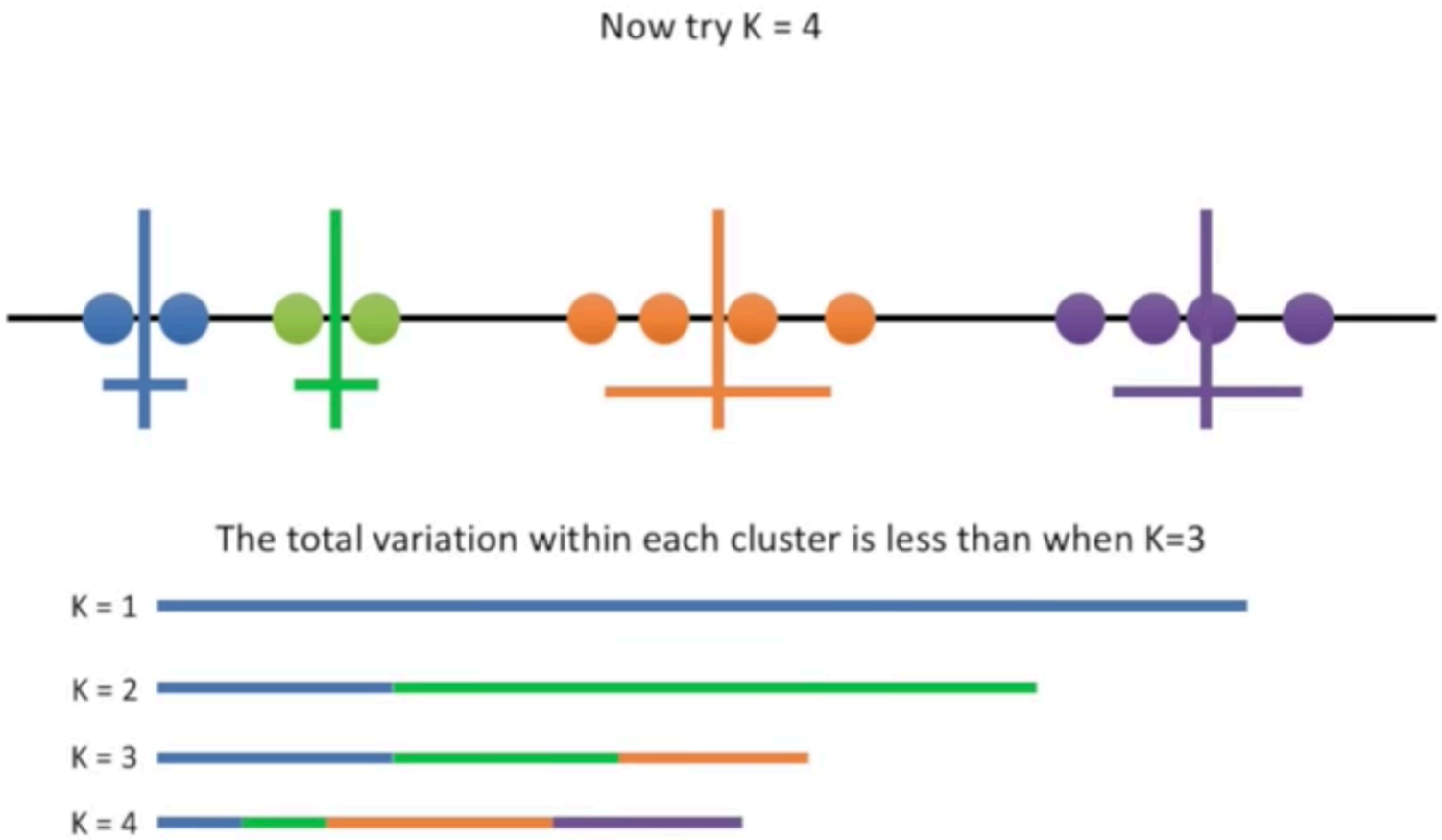
# HOW TO CHOOSE K ?

Now try K = 3

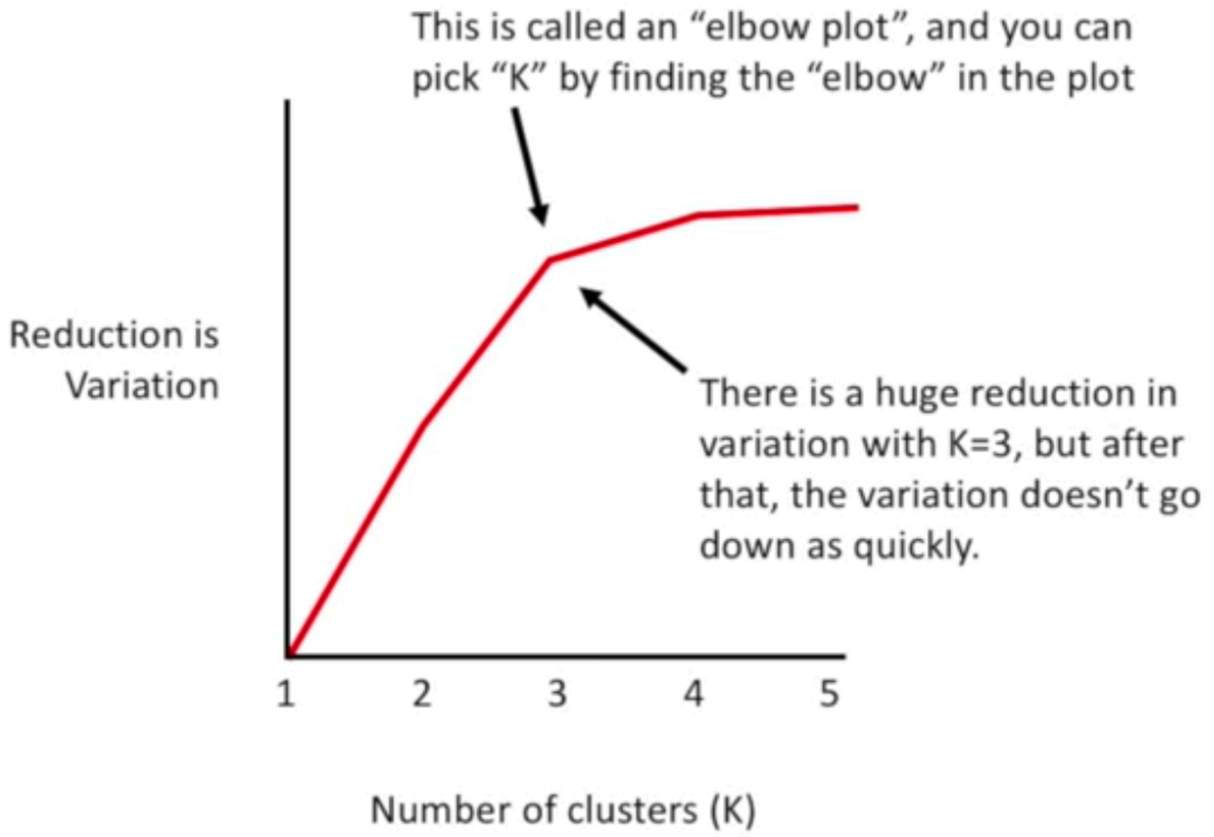


K = 3 is even better! We can quantify how much better by comparing the total variation within the 3 clusters to K = 2

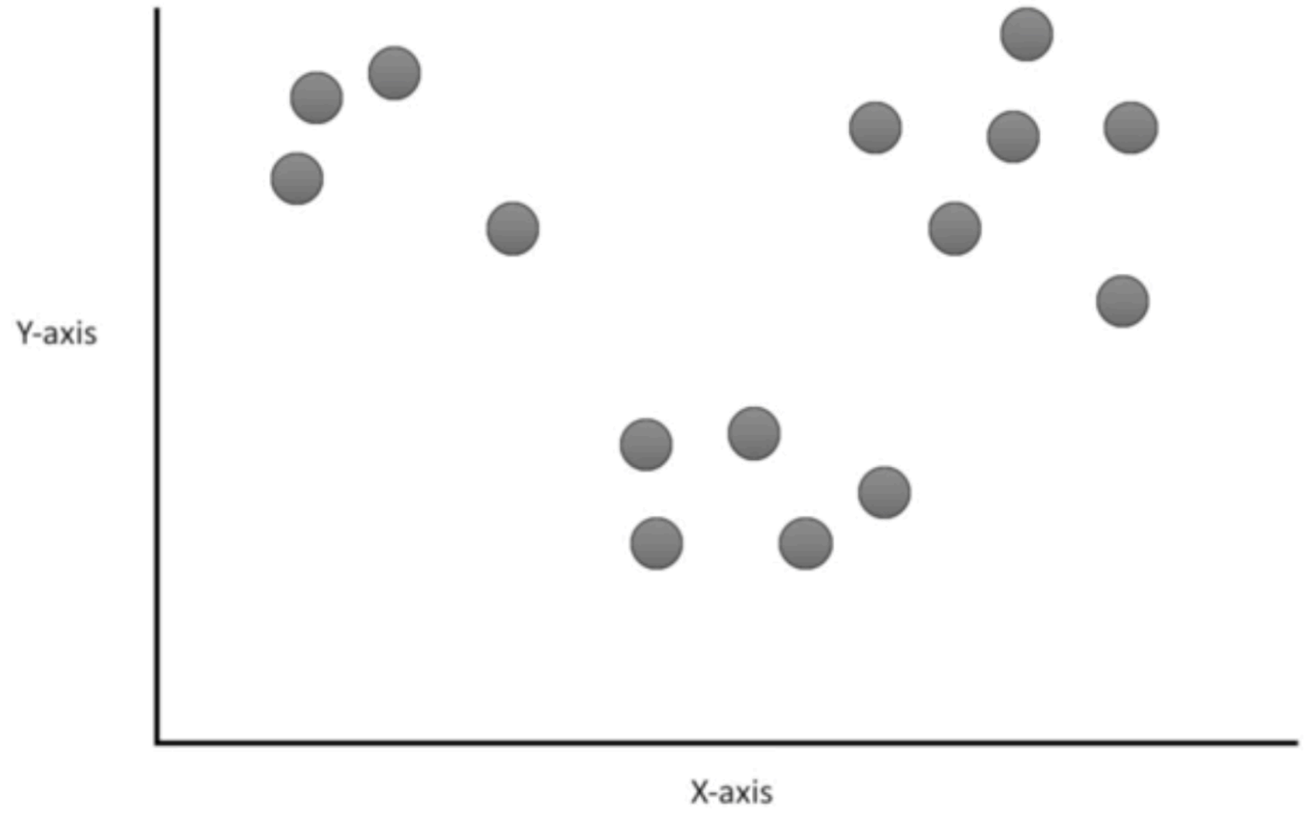
# HOW TO CHOOSE K ?



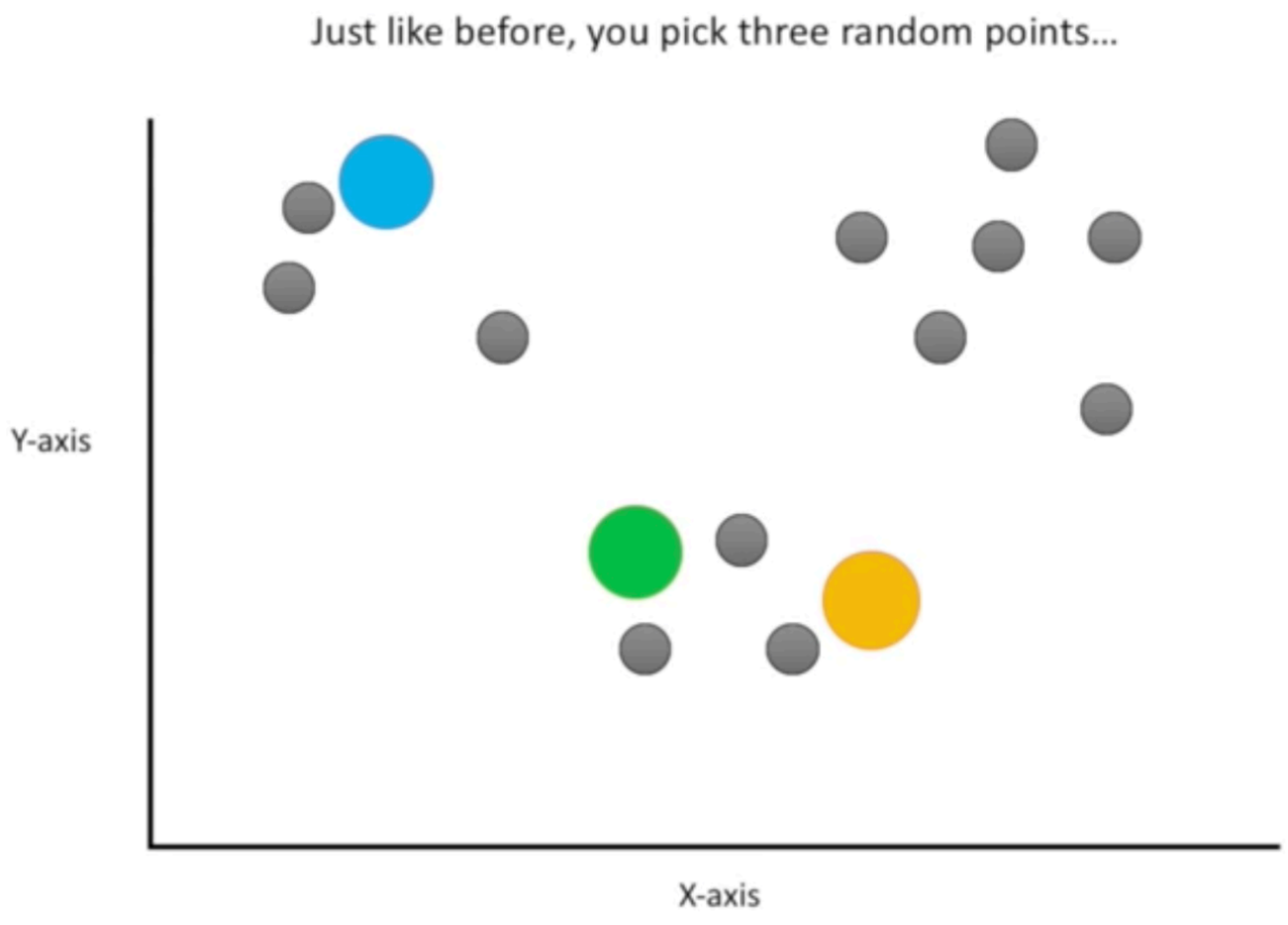
# HOW TO CHOOSE K ?



# MULTIDIMENSIONAL DATA

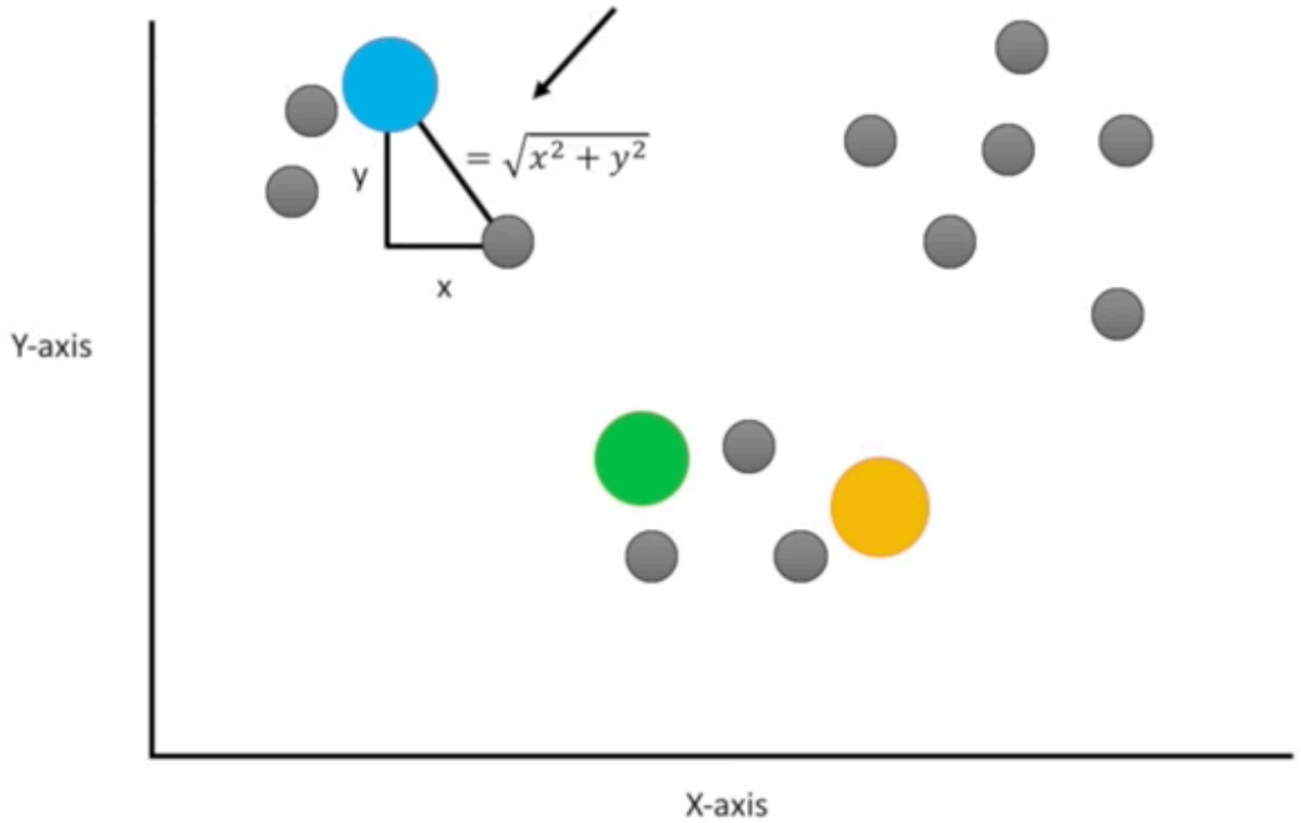


# MULTIDIMENSIONAL DATA



# MULTIDIMENSIONAL DATA

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




When we have 2 axes, the Euclidean distance is:

$$\sqrt{x^2 + y^2}$$

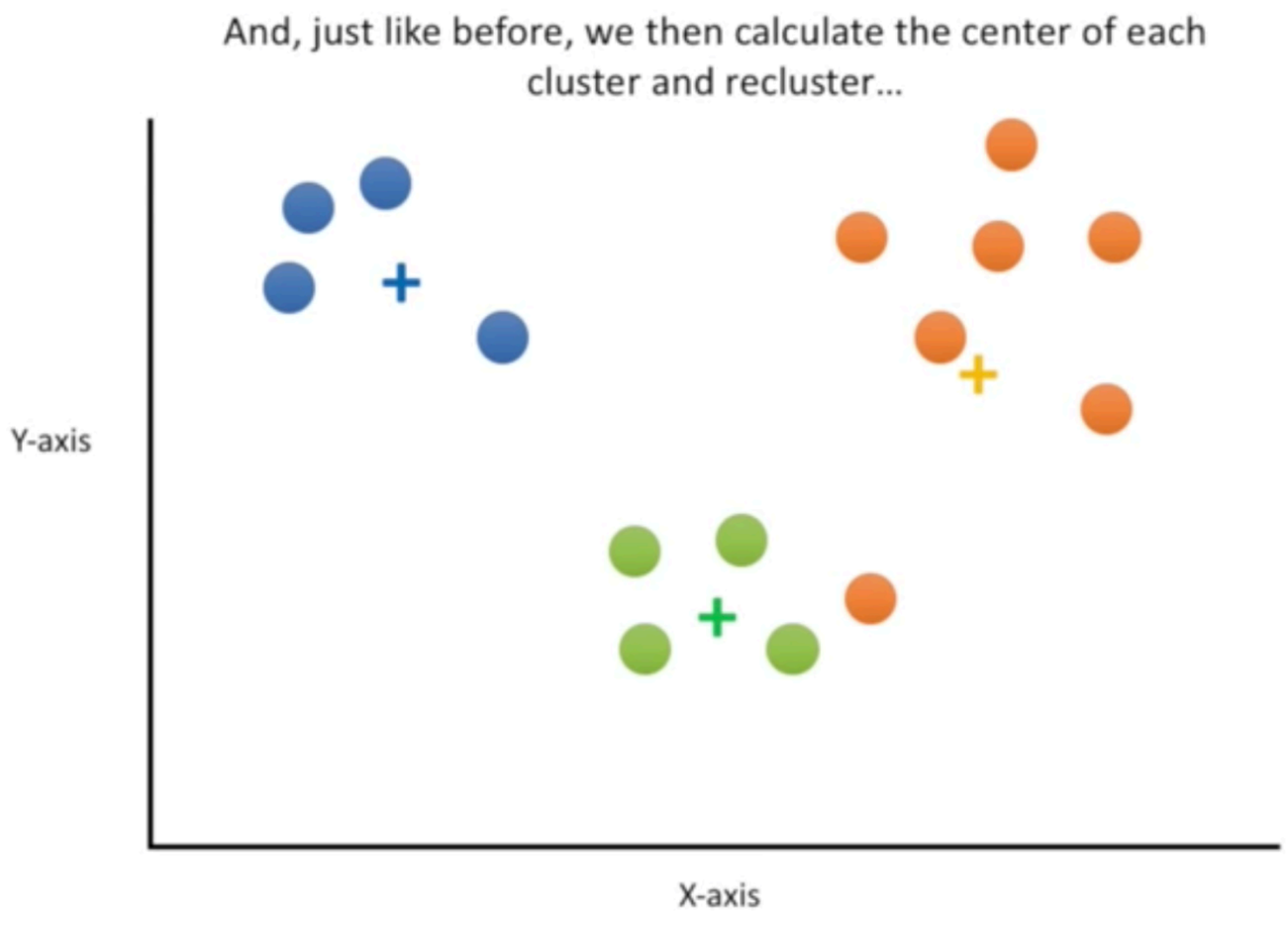
When we have 3 axes, the Euclidean distance is:

$$\sqrt{x^2 + y^2 + z^2}$$

When we have 4 axes, the Euclidean distance is:

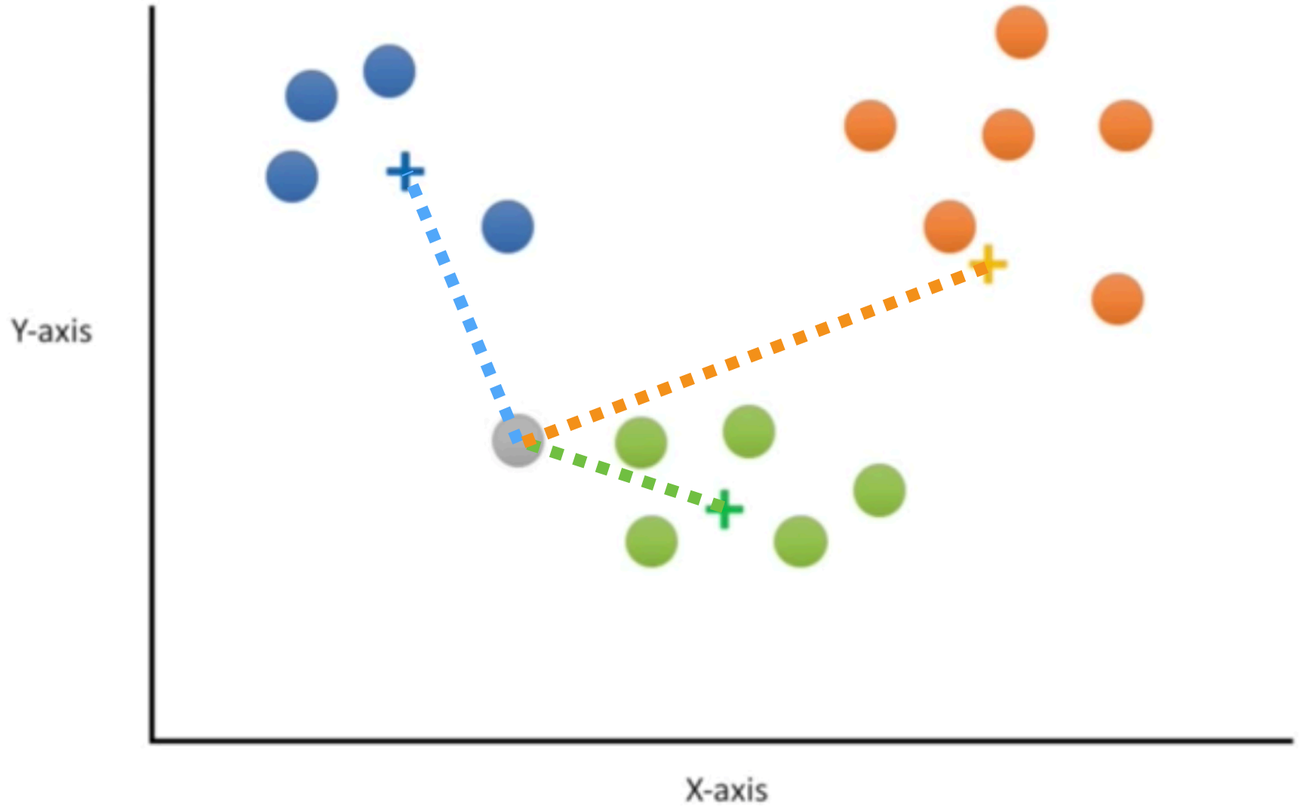
$$\sqrt{x^2 + y^2 + z^2 + a^2}$$

# MULTIDIMENSIONAL DATA

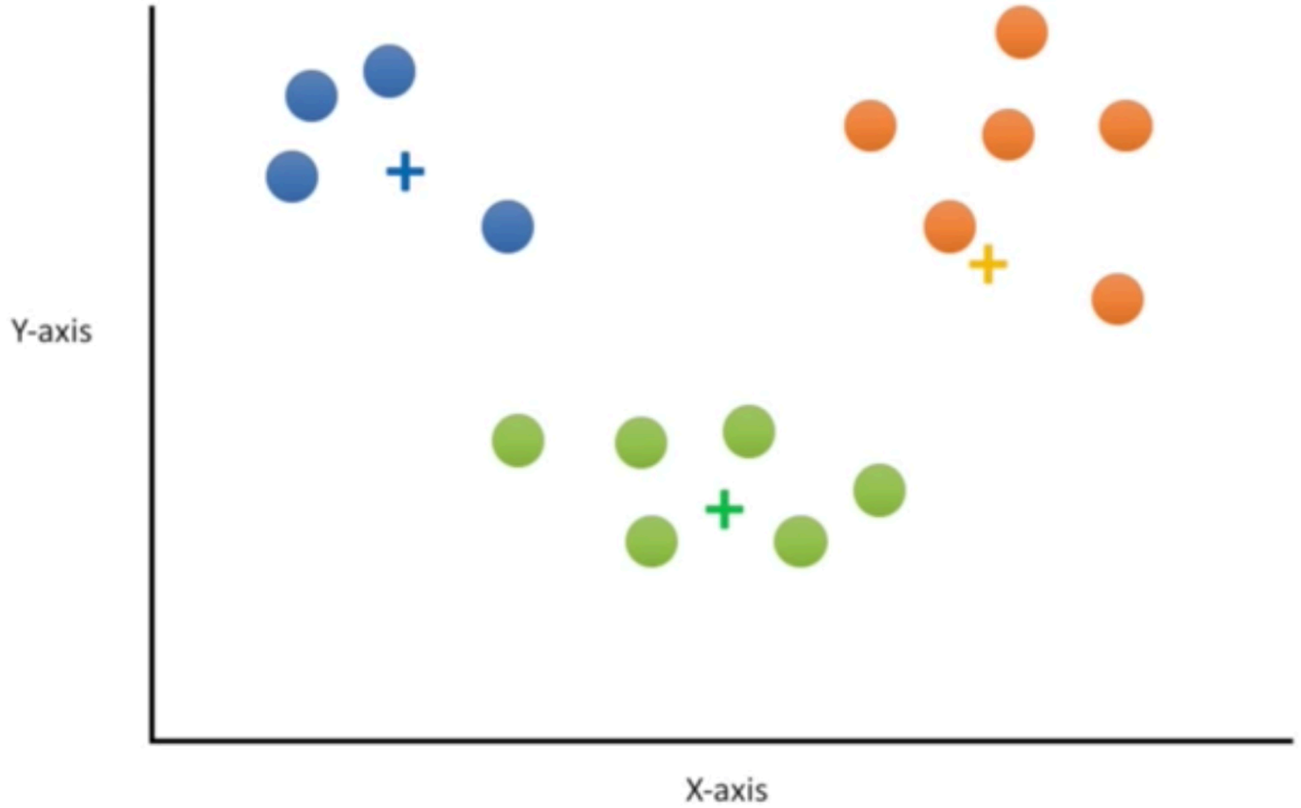


# MULTIDIMENSIONAL DATA

To classify a **new data point**, we chose the **closest cluster mean**.



# MULTIDIMENSIONAL DATA



Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

DATA-DRIVEN TRANSFORMATION

# QUESTION & DISCUSSION