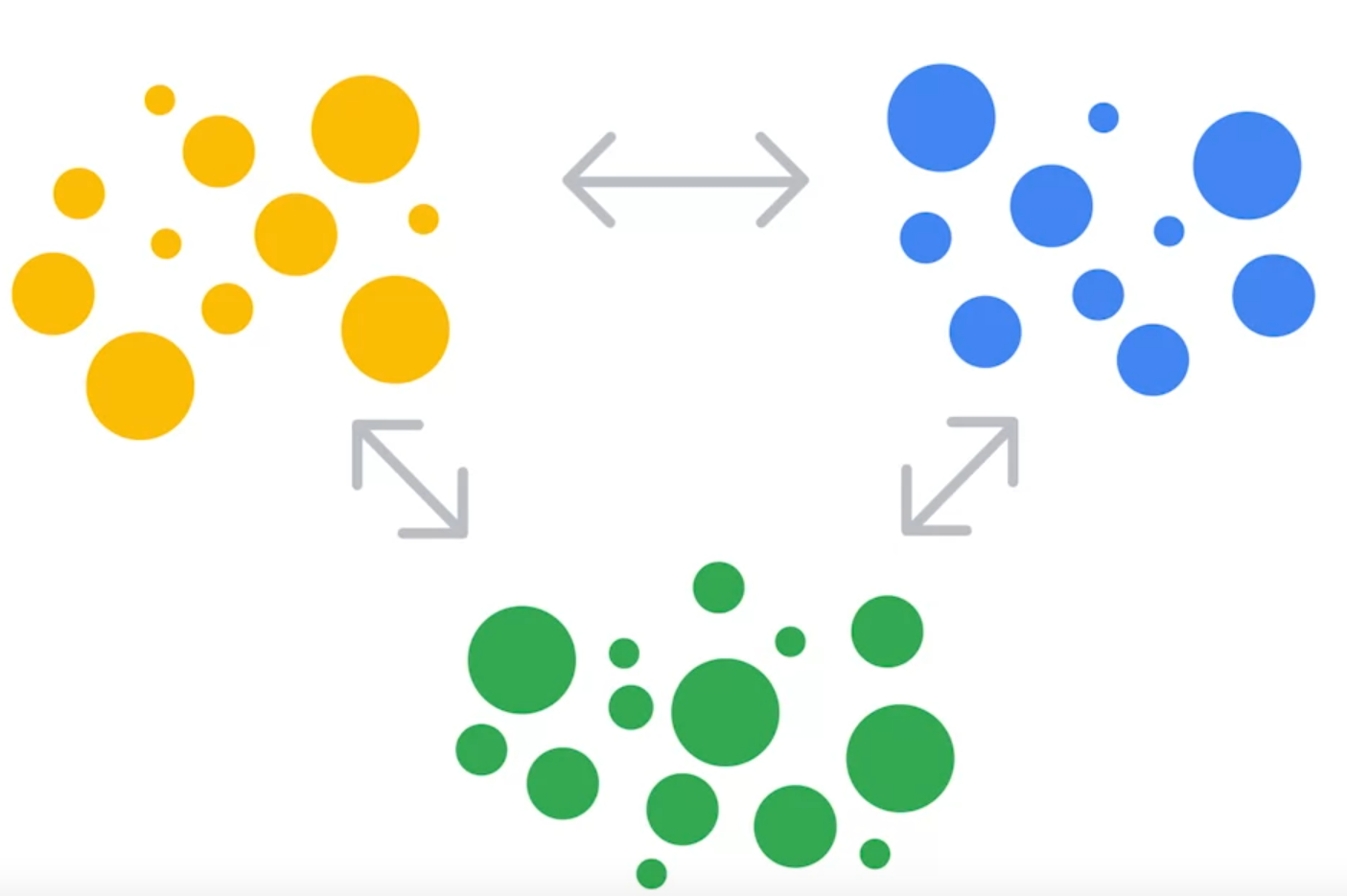
# Evaluate K means

We don’t have any metrics. The goal of K means is just grouping based on similarity. Domain knowledge came in handy at evaluating. A good clustering should be:

1. Clearly identifiable clusters



1. There should be space between clusters



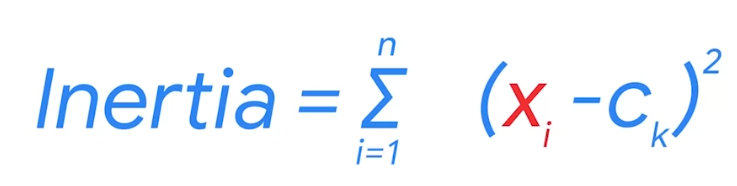
For this they use a metric called Inertia (Sum of squared distances between each observation and its nearest centroid)

**Silhouette Score** are used to find the optimal K value and avoid local minima.

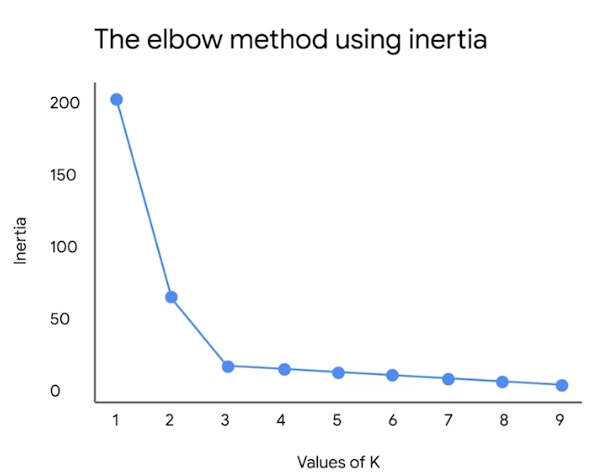
## Inertia

Inertia measures how tightly the data points are clustered around the centroids. It is the sum of the squared distances between each data point and the centroid of the cluster it belongs to.

Purpose: The goal of K-Means is to minimize the inertia. Lower inertia means that data points are closer to their respective cluster centroids, implying that the clusters are more compact.



where Xi is a data point, C is the centroid of the cluster C(k) that the data point belongs to, and n is the total number of data points.

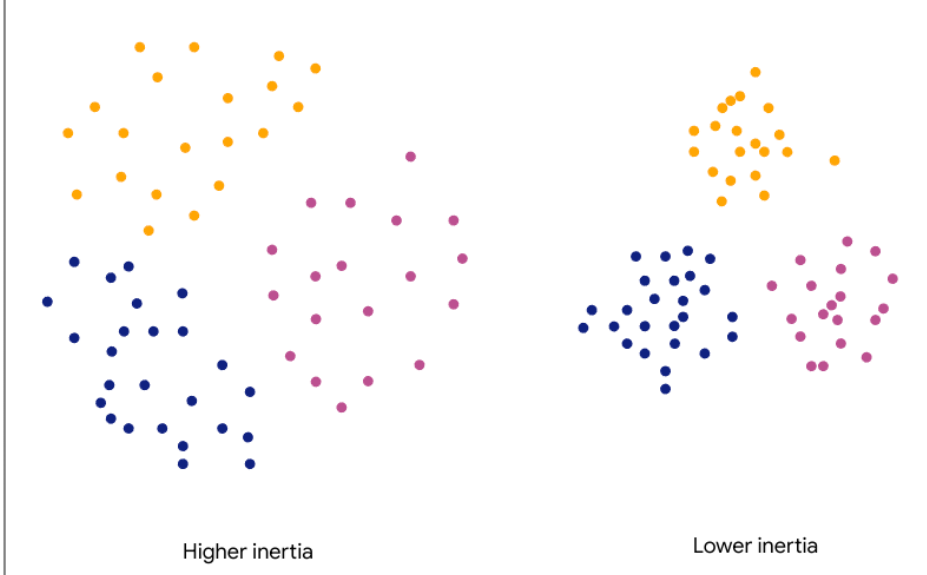


This metric is good for optimize K values

A good way to choose K is when K starts a turning point or the elbow.

Example:

The three clusters on the left have higher inertia than the three clusters on the right, because they are less compactly positioned around their respective centroids.



Note, however, that inertia only measures intracluster distance. Therefore, both of the clusterings in the figure below have the same inertia.

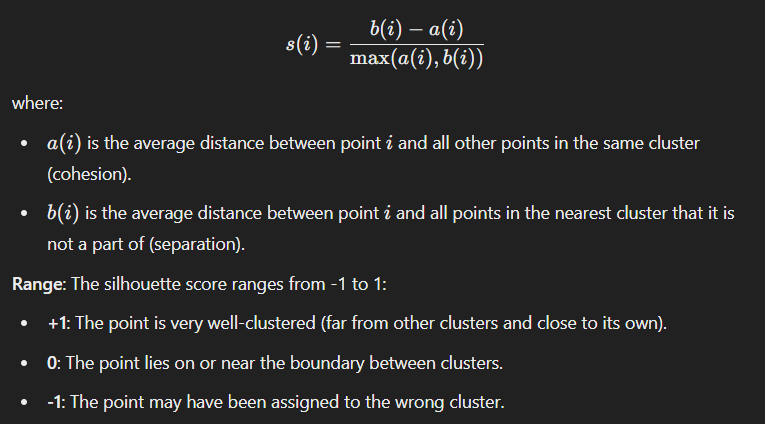


Remember, you want inertia to be low, but if you add more and more clusters with only minimal improvement to inertia, you’re only adding complexity without capturing real structure in the data.

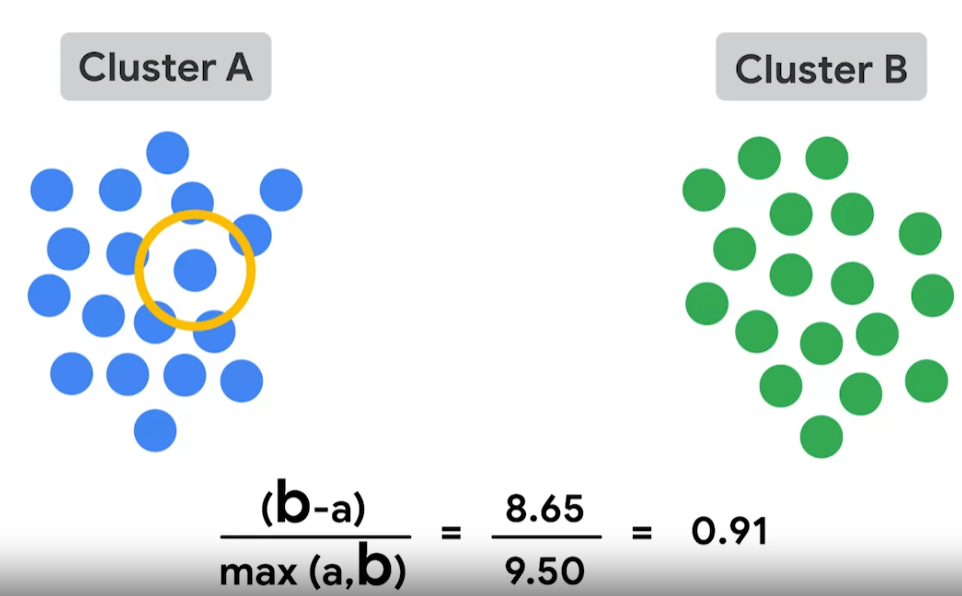
## Silhouette score

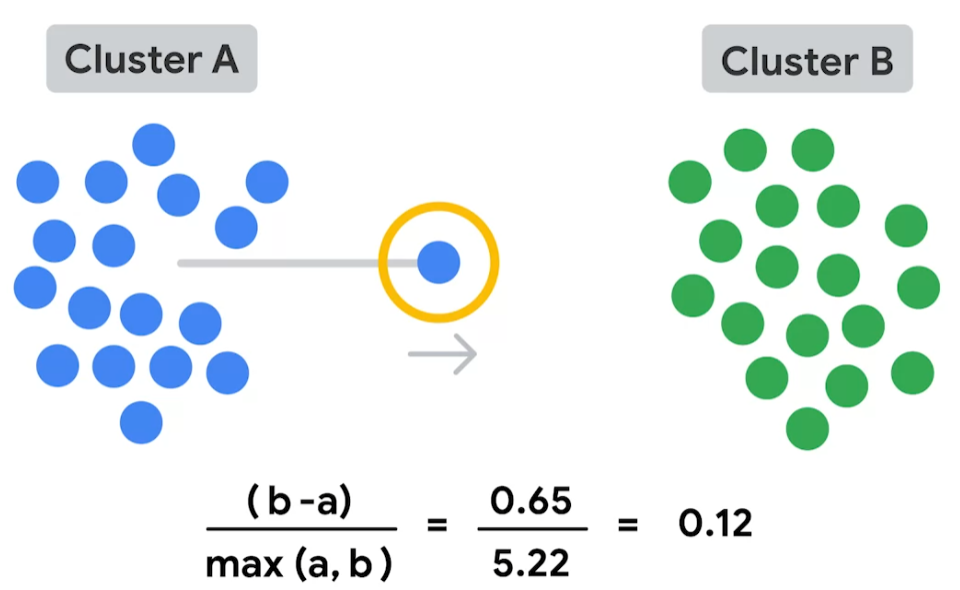
The silhouette score measures how similar an object is to its own cluster compared to other clusters. It quantifies the quality of clustering by measuring how well-separated clusters are and how cohesive the points within a cluster are. It takes the distance between clusters.

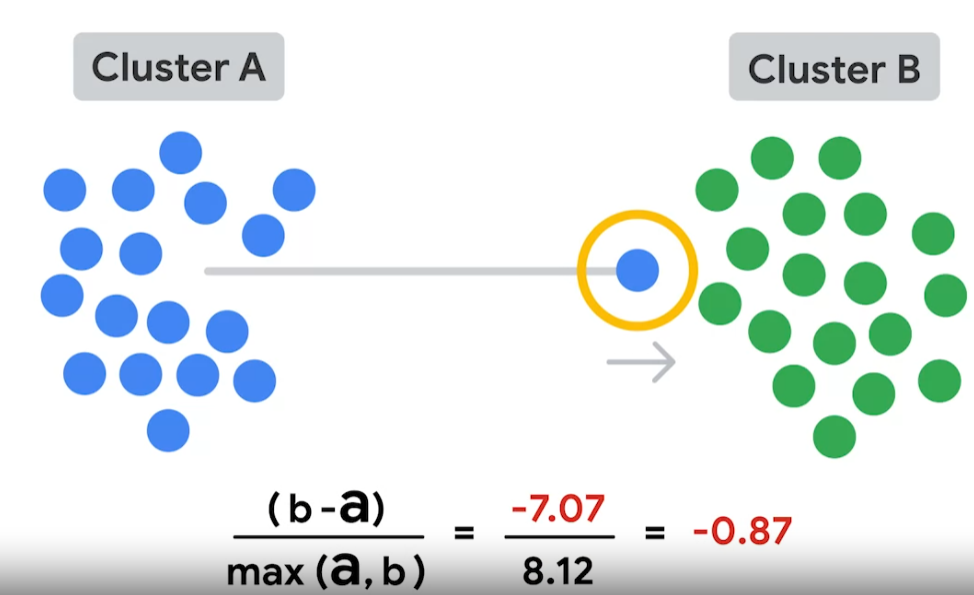
The **silhouette score** is the mean silhouette coefficient over all the observations in a model. The greater the silhouette score, the better defined the model clusters, because the points in a given cluster are closer to each other, and the clusters themselves are more separated from each other.



For example:







The points in the three clusters on the left have lower silhouette coefficients than the points in the three on the right.



