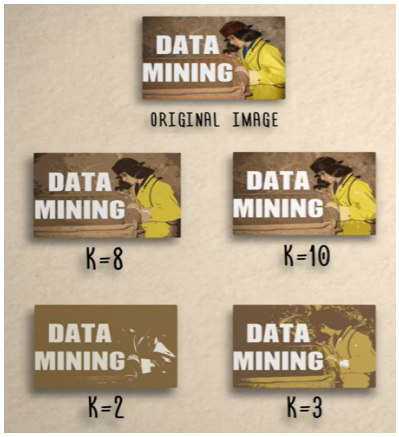
Examples of K-Means Clustering Problems

As we know, Clustering is basically grouping data according to similarity. In this article, we will see some different notions of grouping.
One of the most immediate and obvious issues with K-Means clustering in a variety of applications is that it requires the user to specify the number of clusters K.

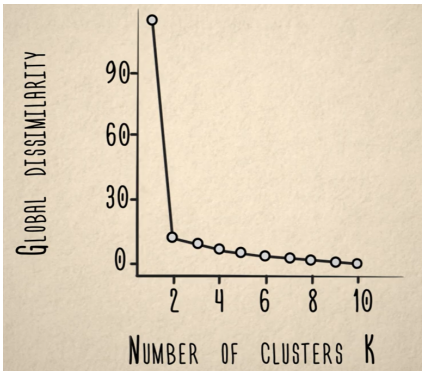
Sometimes this is not a major problem. For example, K-Means can be used for image segmentation. In this application each data point might be Red, Green and Blue, which are the intensity values for each pixel.

The K-Means algorithm finds K colors that can be used to approximate the true and typically much wider range of colors in the image.

Here, K might be determined by how much we want to compress the image.
We can see in the image that, increasing the value of K increases the quality of the compressed image in this case.



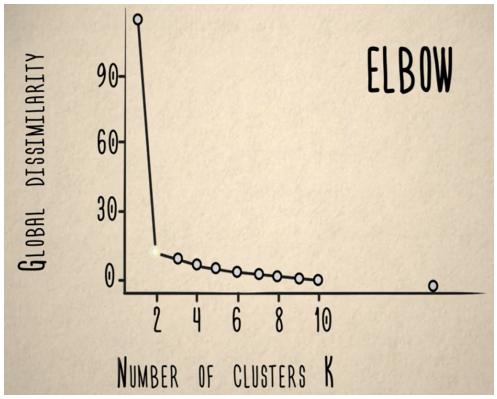
In many other cases though, K is unknown in advance. Suppose we get some new data and we need to determine both K as well as the latent clusters. One option is to use popular heuristics.



One heuristics arises from plotting the global dissimilarity across a wide range of values of K.
We can't choose K immediately from such a figure because the K-Means Objective function will always decrease as we increase K.

If we just choose K as the value that minimizes the K-Means objective function, we would have to choose K equal to the number of data points i.e N. But in that case it would become useless clustering.

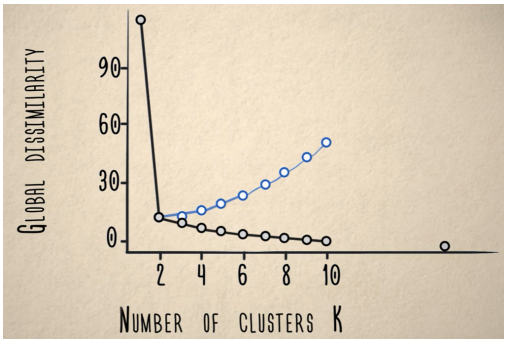
An alternative heuristics is to look for an ELBOW in a plot of Global Dissimilarity vs the number of clusters. Elbow point is a point in the figure where the gain from the new clusters suddenly slows down.



We can set K to be the number of clusters at this Elbow. In the figure which is K=2.

Other more theoretically motivated options for choosing the value of K exist as well.

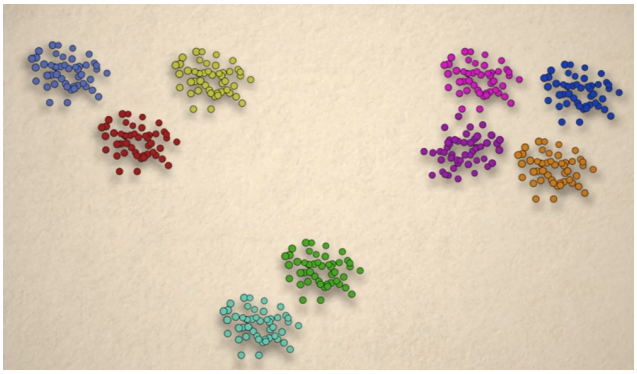
These methods include Gap Statistic as well as methods that change the objective of K-Means. The idea of this method is to explicitly account for the fact that we pay a price for more complex models.



In particular, we can start with the original K-Means Objective function and add a penalty term that grows with the number of parameters.

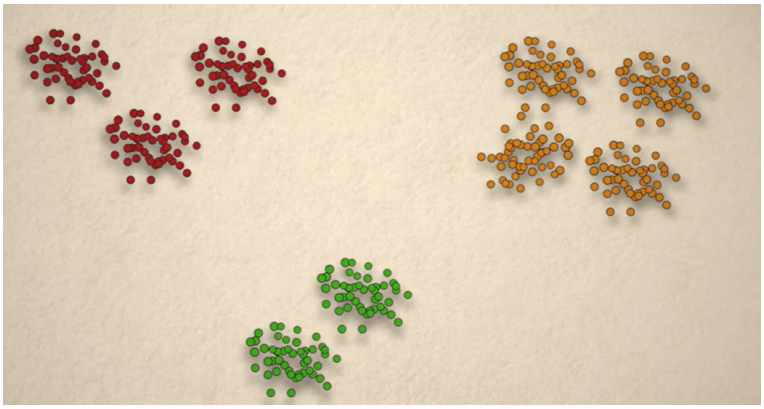
Some appropriate penalties are given by AIC, BIC or other so-called information criteria.

Once these penalties are added, we can plot the new objective function as a function of K, and now there will typically be a non-trivial minimum. We can choose K to be the number of clusters of this minimum.
There are also some more broader issues in choosing the value of K. Sometimes there is not a clear right value for K.



Suppose, we have a bunch of data on various organisms that we have studied in some environment. We might find that if we favor smaller clusters (see in fig), then we cluster the organisms into species and it's a great clustering.As we have picked up an interesting and meaningful latent grouping of the organisms.

But if we favor larger clusters, then we will end up clustering the organisms at the genus, family or order level which are also all useful and meaningful classifications used in biology.



So, clustering at any of these levels would give us an insight into the problem, but each level would result in a different number of clusters.

In this case, it might be useful to try a hierarchical clustering approach, such as **Agglomerative Clustering.**
These approaches explicitly model that some clusters might be composed of further sub-clusters.
So, all the clustering that we have talked about so far puts each data point into only one cluster. This is called **Hard Clustering.**

We also sometimes have clusters that are not perfectly separated. For instance, some data points might be on the border between two or more clusters. And we might not be sure about which cluster these data points should belong to.

An alternative to Hard Clustering that allows us to solve this pattern is **Soft clustering**. In soft clustering, we might allow each data point to have a different degree of membership in each cluster. For instance, a data point might have a probability distribution over its belonging in different clusters. For many data points, this probability distribution might express that we are very sure that the data point belongs in one particular cluster. But in other cases near the border between clusters, the probability distribution for a data point might favor one cluster, but also make it clear that the data point could reasonably belong in another cluster.

In the soft clustering case, we might use alternatives like Fuzzy Clustering and Gaussian Mixture Models instead of K-Means clustering.