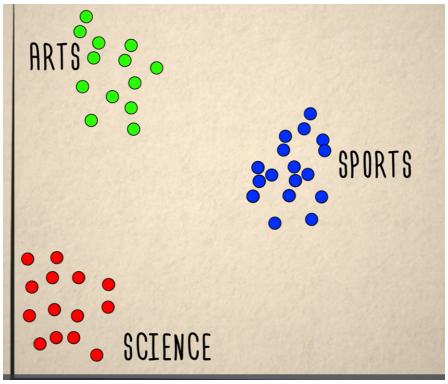
Assumption of K-Means Clustering - Part 1

We already examined some of the assumptions in K-Means clustering. Also we discussed what to do if we want to cluster the data but our application or our data don't quite fit those assumptions.

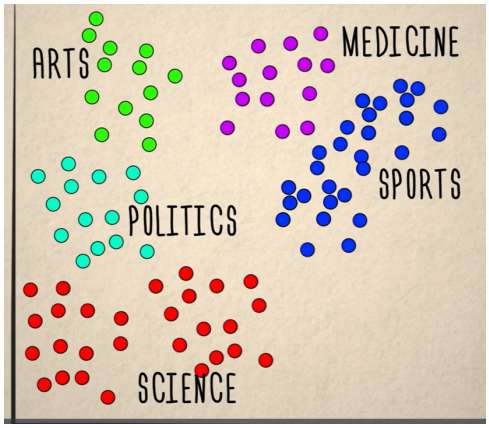
We know K-Means Clustering is just one particular type of fixed K clustering, that is the value of K is fixed . Here we also assume the number of clusters is finite.
Even if we don't know the number of clusters in advance, we often still assume that the number of clusters in our data is some fixed and finite number.

In this article, we will discuss why that assumption might not always be right.
Often bigdata doesn't just mean that we have a single very large data set. We might have **Streaming Data** (where new data is being added to our dataset all the time) .



For instance, English language wikipedia alone averages around 800 new articles per day right now.
If we build a new fitness app, maybe it will not only get new users all the time, but users will interact with each other everyday.

So, what changes when we have Streaming Data?
In the case of wikipedia, no matter how many articles we have read in the past, there are always new topics to read about.
We can think of this as the "Wikipedia Phenomenon".



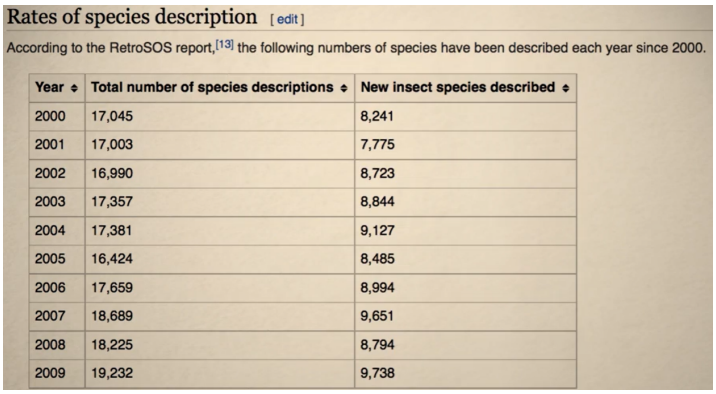
Suppose we have 100 articles from wikipedia. We could cluster those articles and find some topics to group articles together.
But if we have thousands of wikipedia articles, then we expect to find new topics(shown in figure) that we haven't seen in the first 100 articles.

This file is meant for personal use by tomperry7@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

As we read more and more wikipedia articles, getting to millions and billions of articles, we expect to keep finding new topics.
In this case it might be okay to run clustering with some fixed number of clusters for a hundred articles or any fixed number of articles.
But the number of clusters grows with the size of the data set. We can also say that we want the complexity of our model to be able to grow with the size of the data.

This happens in a lot of other types of data as well. For example,

1. Humans have been discovering species for a very long time and yet at any time if we do a search on google news, we will surely find that there are articles about all kinds of new species that were discovered in the past week. So, we are still discovering new species all the time.



Rates of species description [edit]

According to the RetroSOS report,[13] the following numbers of species have been described each year since 2000.

Year	Total number of species descriptions	New insect species described
2000	17,045	8,241
2001	17,003	7,775
2002	16,990	8,723
2003	17,357	8,844
2004	17,381	9,127
2005	16,424	8,485
2006	17,659	8,994
2007	18,689	9,651
2008	18,225	8,794
2009	19,232	9,738

2. Take the example of our fitness app that we have been referring to for some time. We can ask people about their exercise routines. And no matter how many people we ask, we might expect that we will keep discovering new and unusual exercises that we haven't seen among the previous users.

3. In genetics, we might expect to find some new and unknown ancestral populations as we study more individual's DNA. Or suppose we look at newborns in a hospital and as we study more and more about newborns, we might expect to find more new and unique health issues among those newborns, and we want to be prepared for the new health issues, so that we can treat the newborns better.

This file is meant for personal use by tomperry7@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

4. Consider a social network, as more and more people join the social network, we expect to see new friend groups and interests represented in the network.

In all these cases, we don't want a fixed number of cluster K, for clustering. Here we want K to be able to grow as the size of the data grows.

One solution to this problem is provided by **Non-Parametric Bayesian Methods**.
Here, the 'Non-Parametric' part means many parameters or infinitely many parameters.

These methods let the number of instantiated parameters grow with the size of the data. But these methods are more complex than K-Means Clustering. But some very recent work on **MAD-BAYES** shows how to turn Non-Parametric Bayesian methods into K-Means like problems and algorithms for clustering in particular and Unsupervised learning in general.

An additional read on Non-Parametric Bayesian Methods can be found at Here.

This file is meant for personal use by tomperry7@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.