Lesson 20

Clustering

Kush Kulshrestha

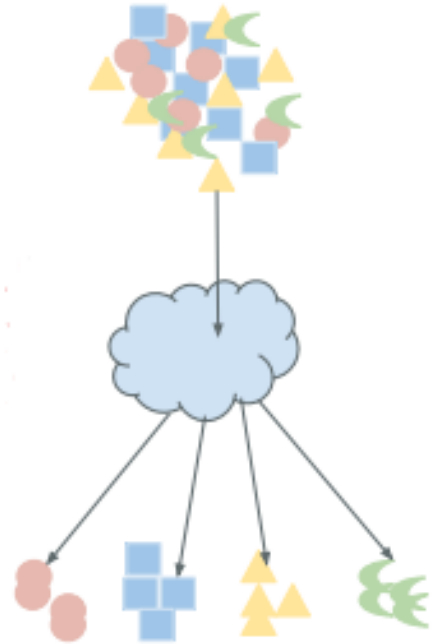
Supervised vs Unsupervised

Supervised Learning

- For every x there is a y
- Goal is to predict y using x
- Most of the industrial problems are solved using supervised learning.

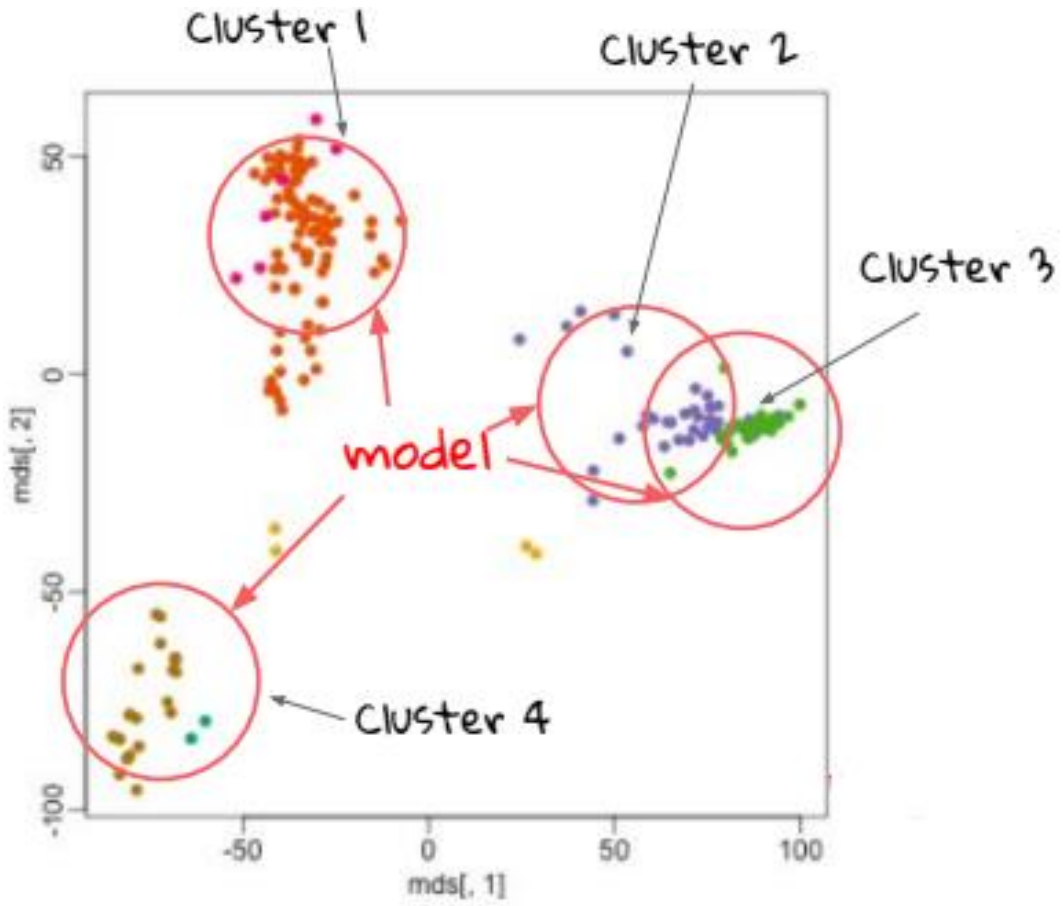
Unsupervised Learning

- For every x there is no y.
- Goal is not prediction, but to investigate the x
- Unsupervised learning has less use cases comparatively. It is used as one of the pre-processing step before supervised learning.



Clustering

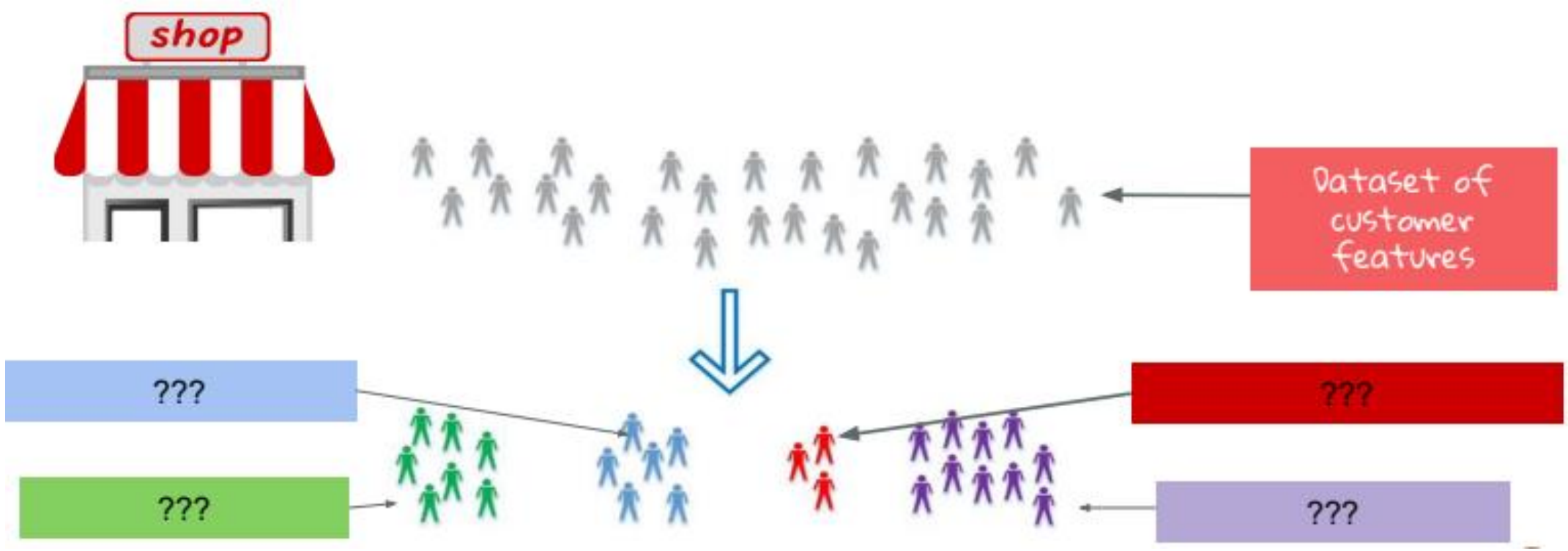
- Clustering splits data in order to find out how observations are similar on a number of different features.
- We are not predicting a true Y.
- The clusters are the model. We decide the number of clusters, represented as K.



Clustering

Imagine that you are the owner of an online clothing store. You want to segment your customers by their buying habits.

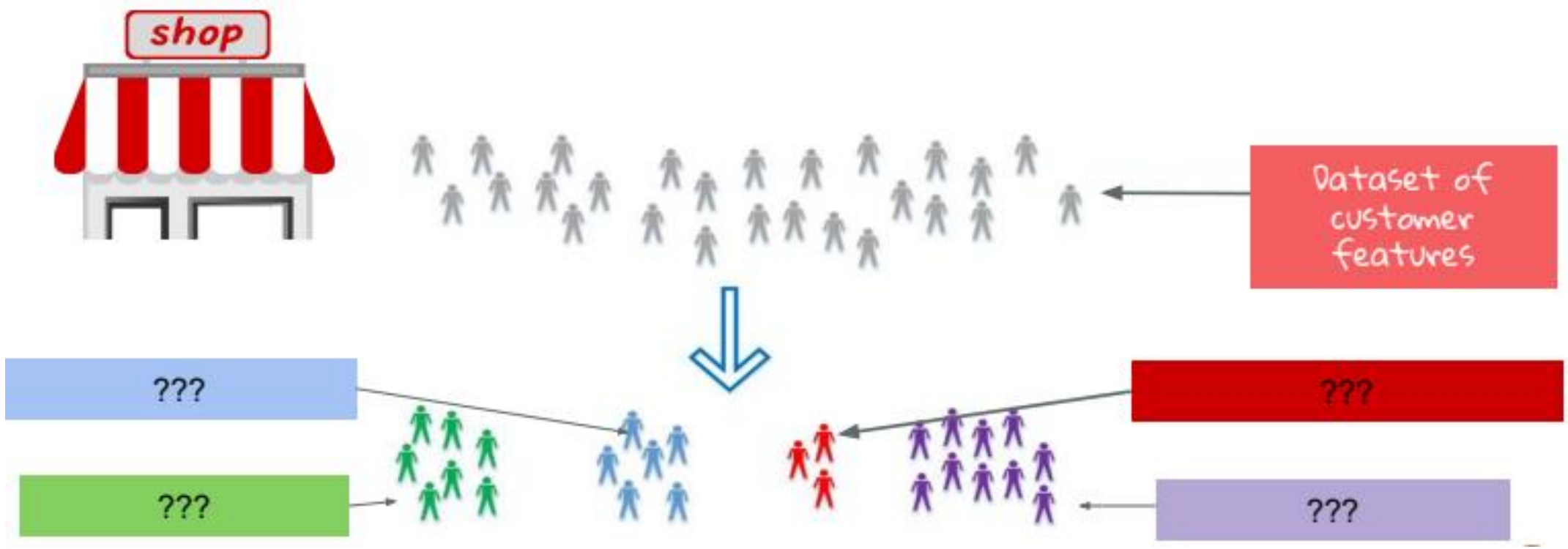
- As the owner of the store, you think that some customers are bargain-hunters while others are price-insensitive; some come in every week while others drop in during the holidays.
- But you don’t actually know definitively which ones are which.



Clustering

You could take a supervised approach to this problem, and try to predict how much a customer will spend, or how frequently a customer will drop in.

But here, we're just trying to see what the data will tell us. This is what makes clustering an unsupervised problem.



Clustering

Since we have an online website for our store, we have valuable data about how our consumers behave online.

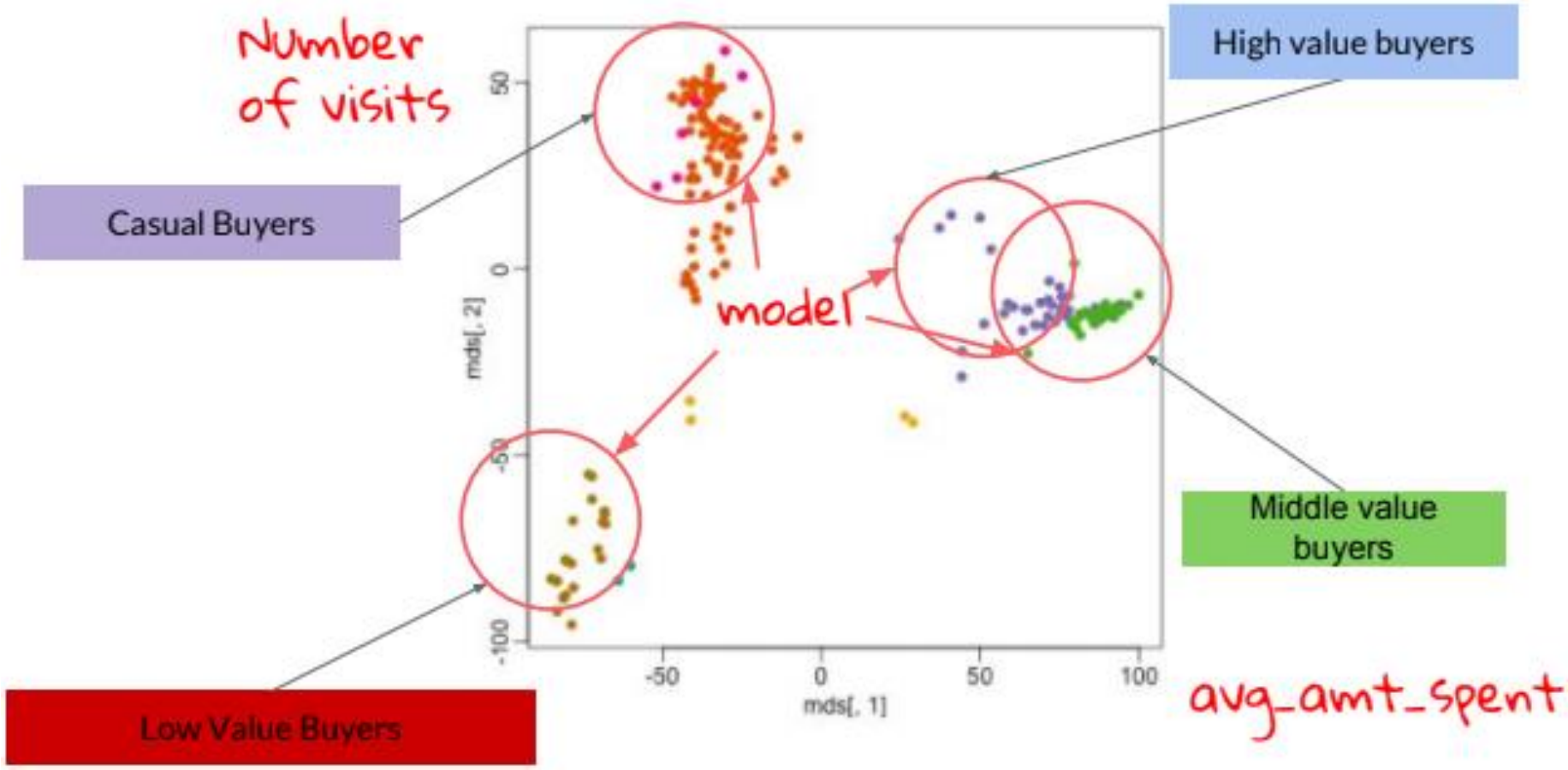
We select features that will determine how clusters are formed in our algorithm.

Since we want to segment customers by their buying habits, we probably want to form clusters using the features "number of visits" and "average amount spent." Let's start with these.

customer_id	Number of visits	Avg_amt_spent	traffic type	%_of_visits_during_sales
	X1	X2	X3	X4
1237482	5	$92	organic	20%
1213345	50	$35	Email_sale	100%
2323764	20	$200	Email_new_collection	10%
2326734	1	$40	organic	100%

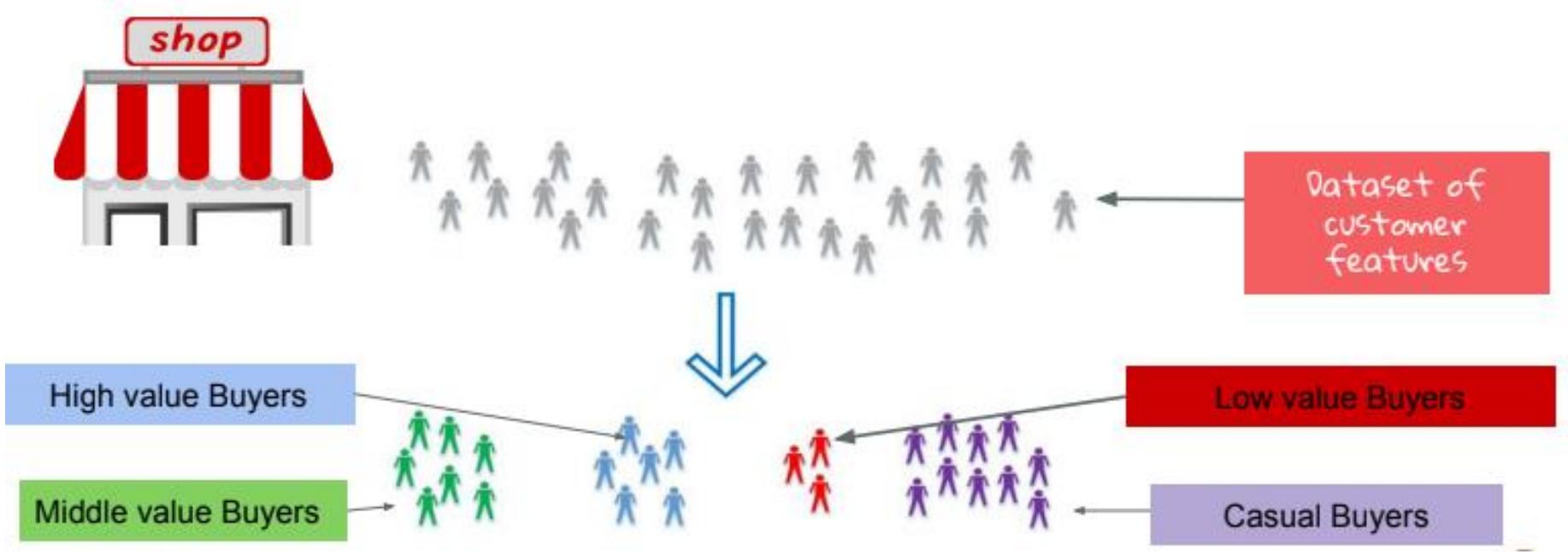
Clustering

By applying clustering on “number of visits” and “average amount spent”, we can demonstrate clustering using two features in two dimensional space!



Clustering

And by tagging each of the cluster with appropriate name/context in business sense, we can focus on each group differently.

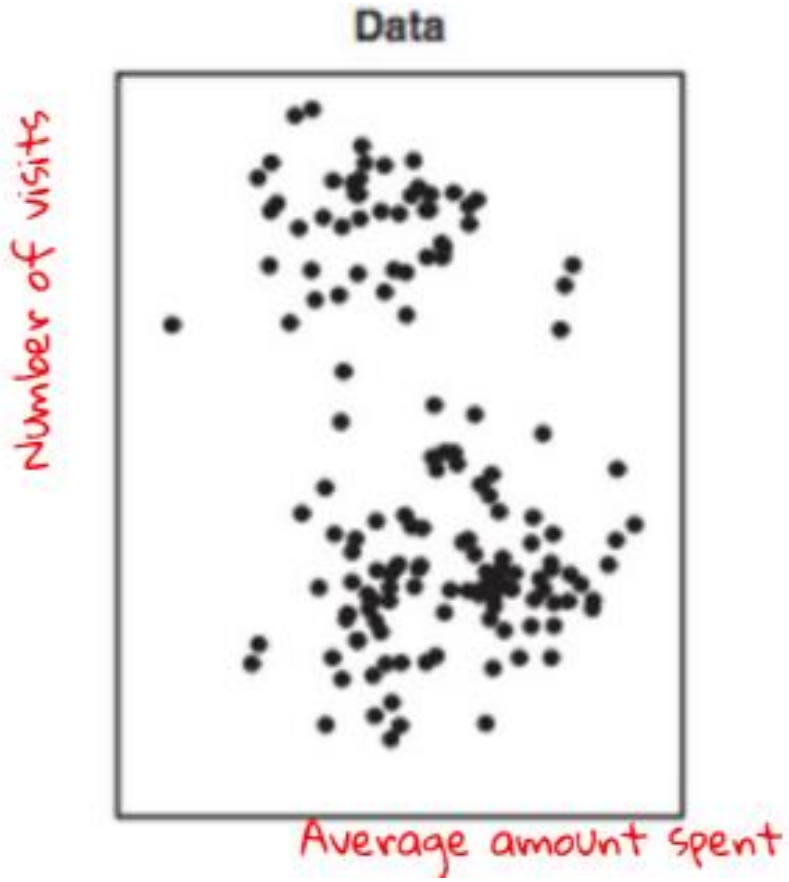


K-Means Clustering - Algorithm

Step – I

We start with a simple scatter plot of number of visits against average amount spent.

Decide the value of K – total number of clusters in your data.

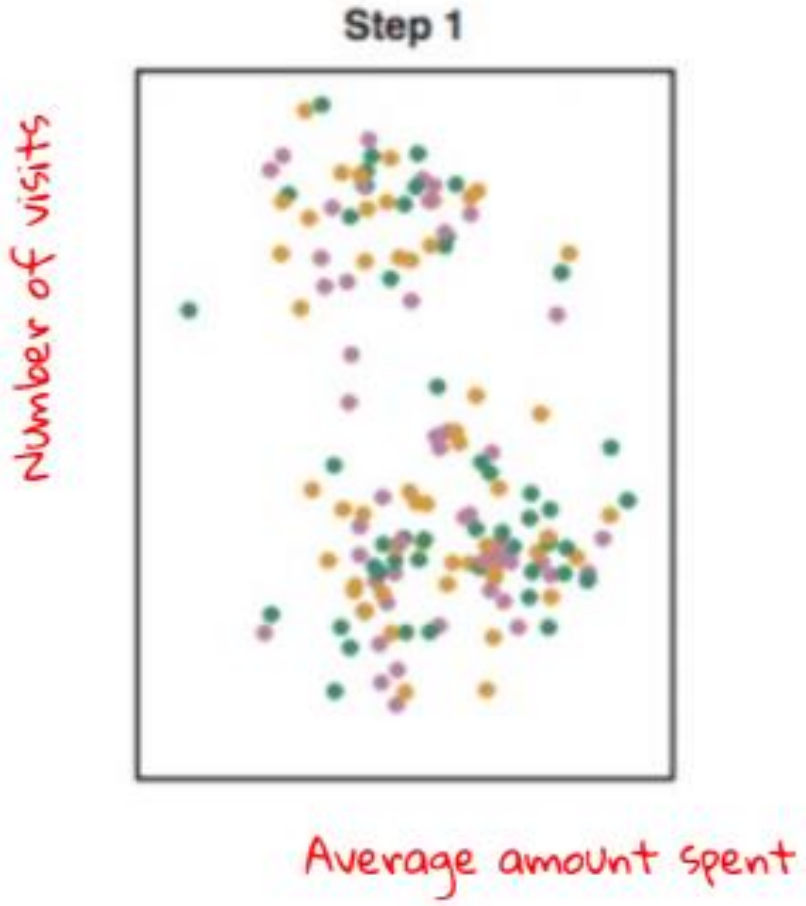


K-Means Clustering

Step – II

Randomly assign each customer to a cluster.

Data points have randomly been assigned into pink, yellow, and green groupings.

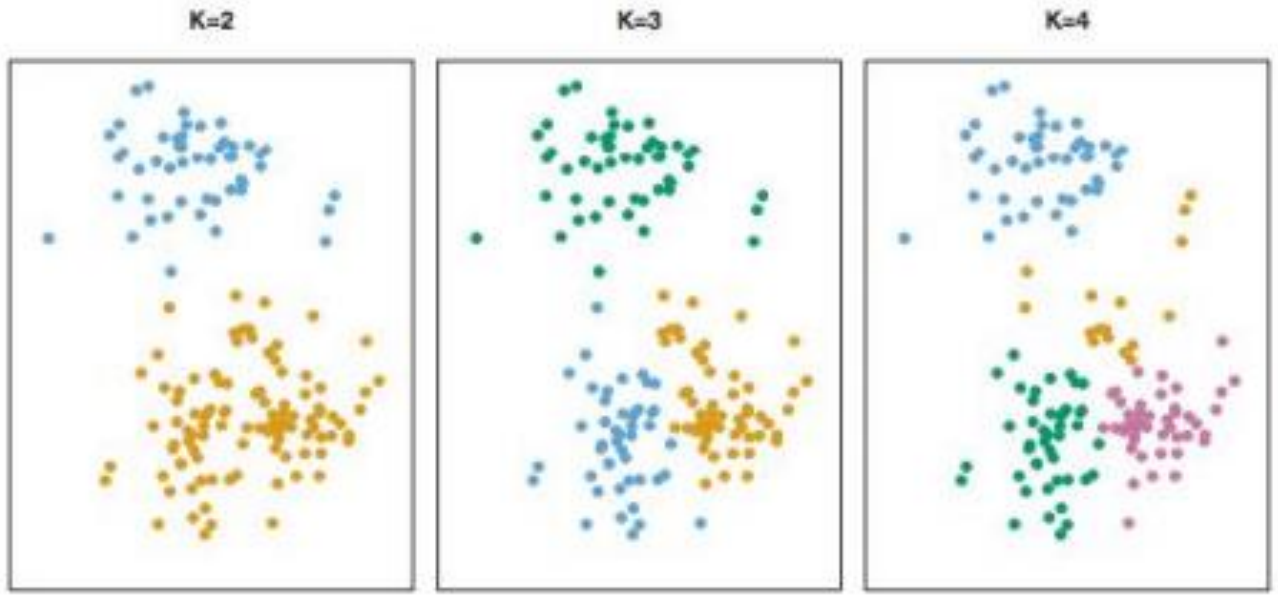


K-Means Clustering

In our example we specified 3 clusters, but we can tell the model whatever “K” we want.

A very large number of clusters will cause overfitting (for example, if you set n equal to the number of observations you will have a cluster for every observation!) This is not very useful for making sense of subsets of our data.

K is an example of hyper parameter, the value which is input to the algorithm and is decided by you.



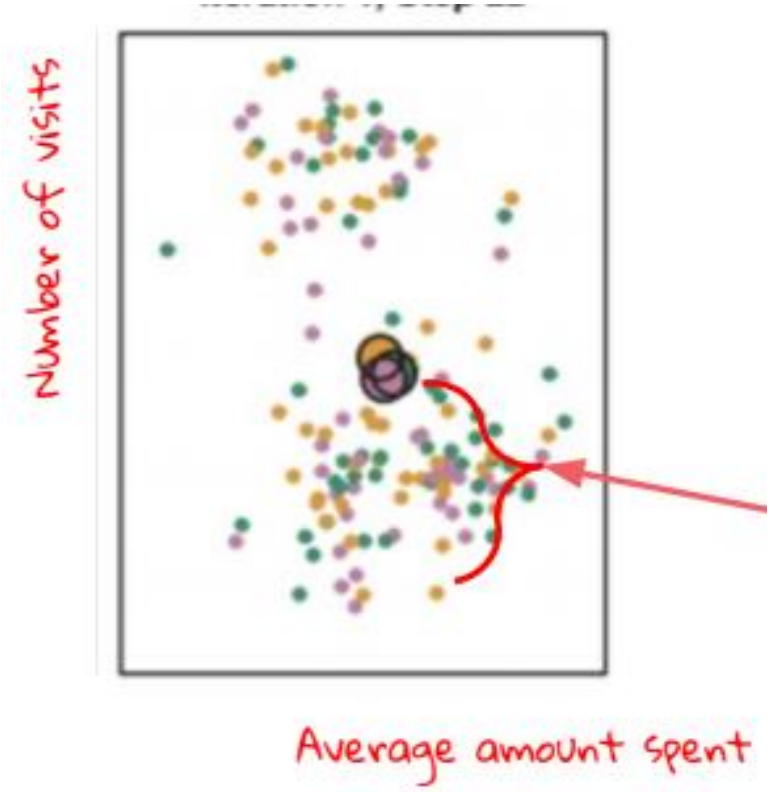
K-Means Clustering

Step – III

We calculate a centroid for each cluster. Our goal is to minimize the distance from centroid to any observation.

The centroids are calculated as the center of their randomly assigned group.

Here, centroids are really close together, and distance to the outer points is very large -- we can definitely do better!



K-Means Clustering

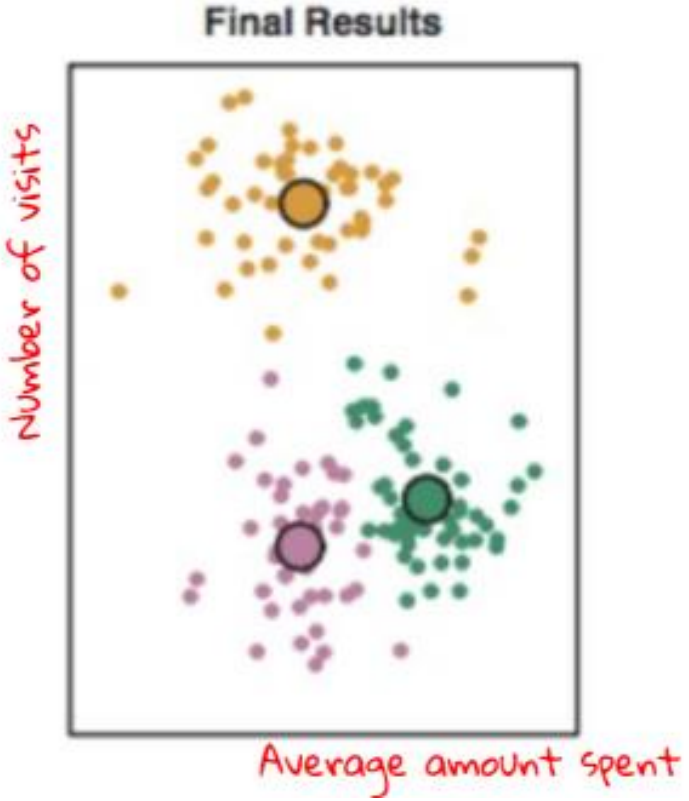
Step – IV

Now we re-assign each observation to the cluster group of the nearest centroid.

Recalculate centroid, with time centroids will go further apart.

With time, fewer observations will change clusters.

We keep repeating the centroid calculation / cluster re-assignments until nothing moves anymore - the model is complete!



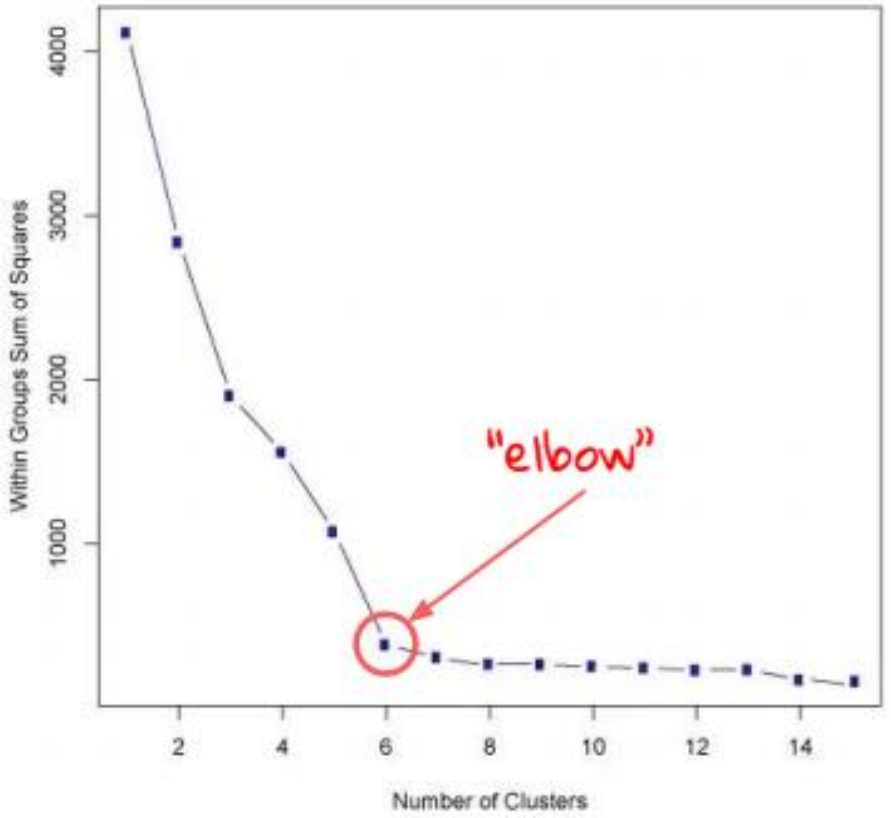
K-Means Clustering

Selecting Optimal value of K

An Elbow Plot visualizes how the error decreases as K increases.

If K == n (number of observations), then distance = 0 (each observation its own cluster)! This, as we know, is a case of overfitting.

In the graph to the right, we should choose K=6. This appears to be the number of clusters where the biggest gains in reducing error have already been made.



Pros and Cons

Advantages:

- Does not assume any underlying distribution (e.g. no normal distribution assumption like in linear regression)
- Easy to represent physically.
- Produces intuitive groupings.
- Can work in multiple dimensions.

Disadvantages:

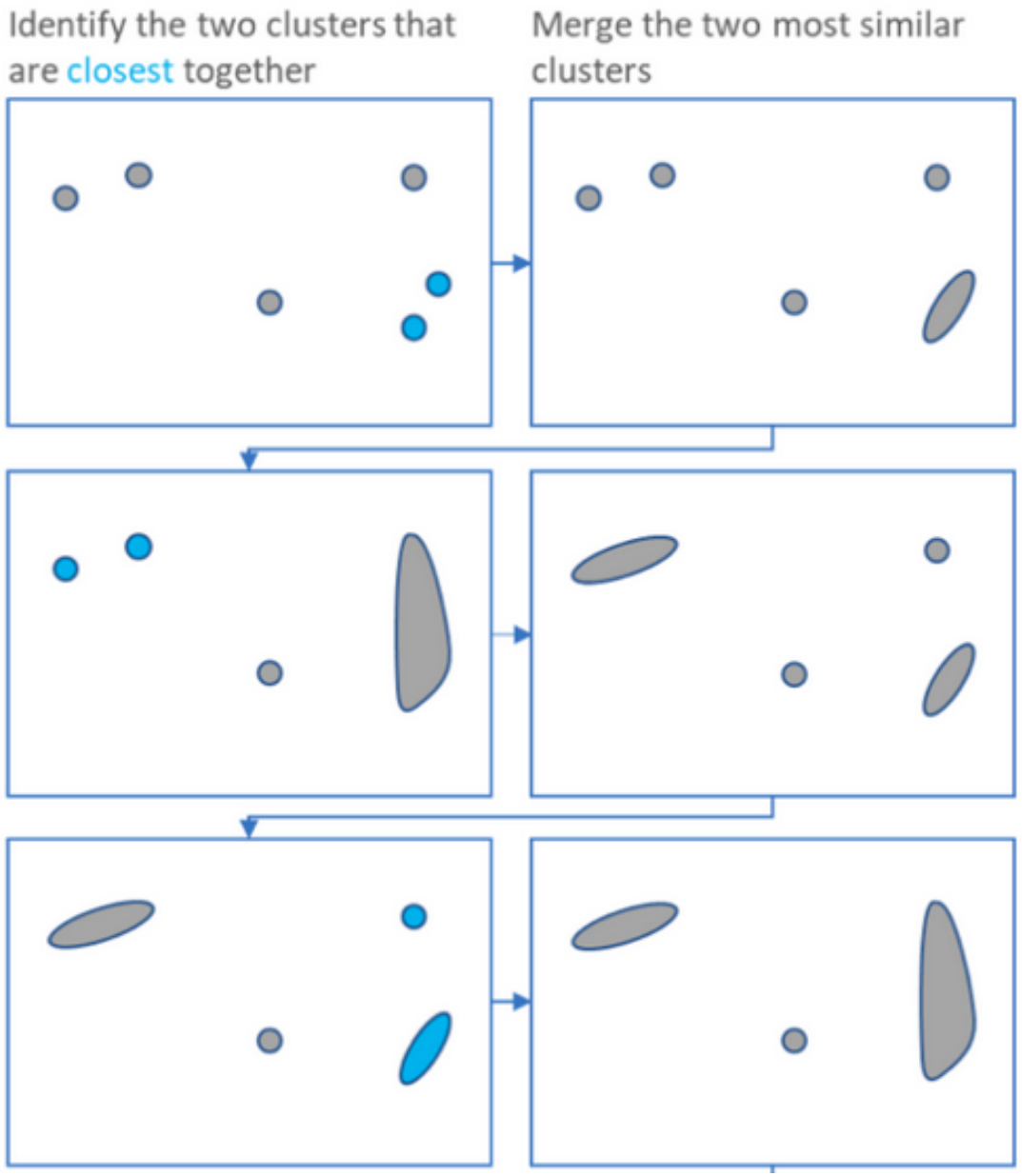
- Time-consuming to find the optimal number of clusters.
- Time-consuming feature engineering (features must be numeric and normalized)

Hierarchical Clustering

Hierarchical clustering*,* also known as *hierarchical cluster analysis,* is an algorithm that groups similar objects into groups called *clusters*. The endpoint is a set of clusters*,* where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

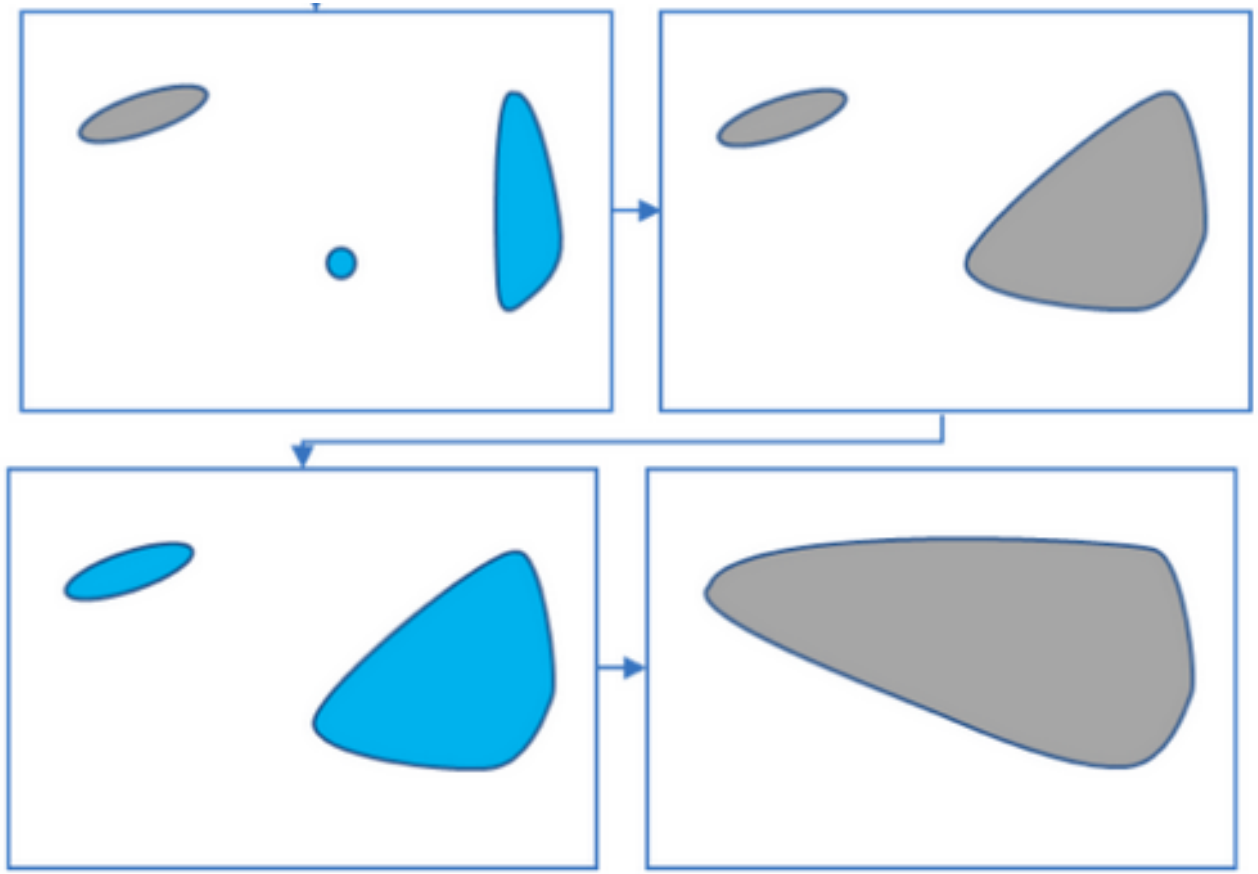
Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This continues until all the clusters are merged together.

Hierarchical Clustering



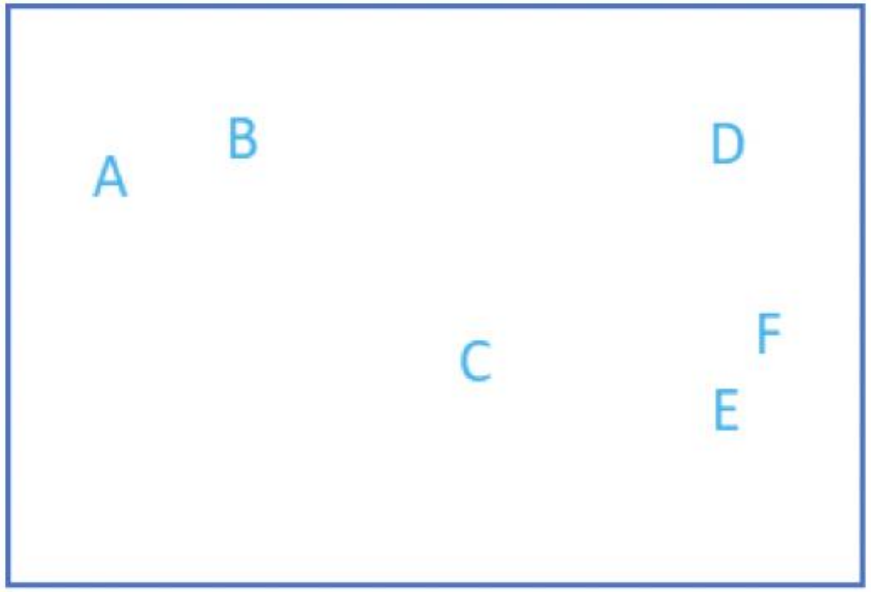
Hierarchical Clustering

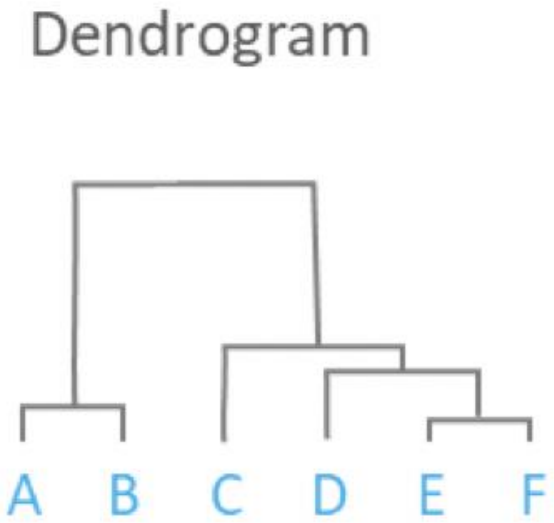
Continued..



Hierarchical Clustering

The main output of Hierarchical Clustering is a *dendrogram*, which shows the hierarchical relationship between the clusters:





Measure of Distance

In the example above, the *distance* between two clusters has been computed based on length of the straight line drawn from one cluster to another. This is commonly referred to as the *Euclidean distance.* Many other *distance metrics* have been developed.

The choice of distance metric should be made based on theoretical concerns from the domain of study. That is, a distance metric needs to define similarity in a way that is sensible for the field of study. For example, if clustering crime sites in a city, city block distance may be appropriate (or, better yet, the time taken to travel between each location). Where there is no theoretical justification for an alternative, the Euclidean should generally be preferred, as it is usually the appropriate measure of distance in the physical world.