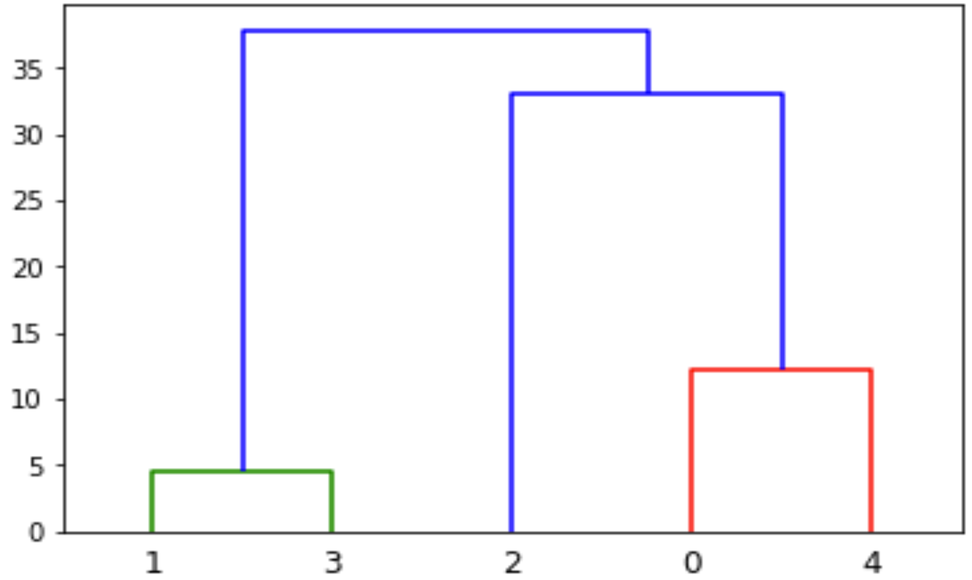
Hierarchical Clustering

Part - 1

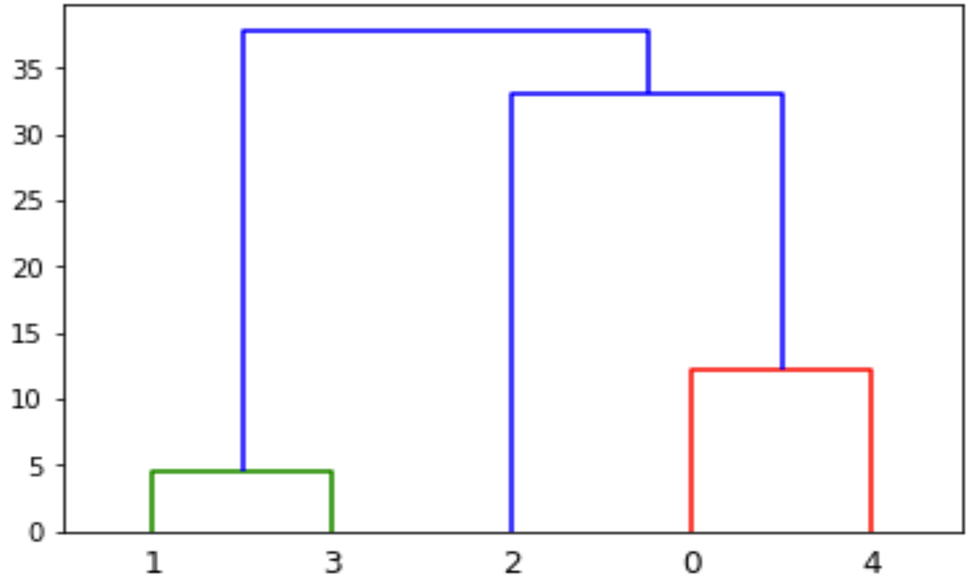
Hierarchical Clustering

Hierarchical algorithms produce a nested sequence of partitions of the data which can be represented by using a tree structure that is called as dendrogram.



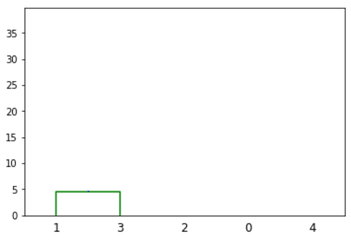
Hierarchical Clustering

There are two types of hierarchical algorithms. One is agglomerative and the other one is divisive.

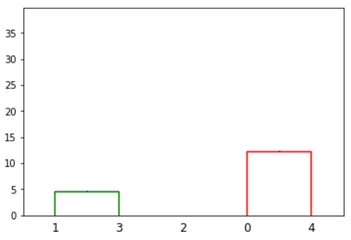


Agglomerative Hierarchical Clustering

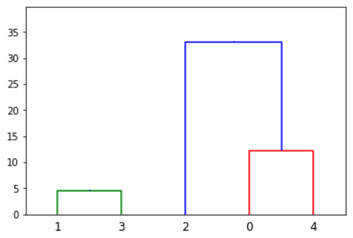
- Each observation is considered as a cluster. At each levels, the most similar pair of clusters are merged to form a new cluster in the next level.
- This reduces the size of the partition by one.
- This is a bottom-up approach. This may be represented by a up-side down tree structure



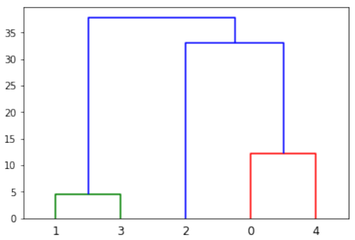
step 1



step 2



step 3



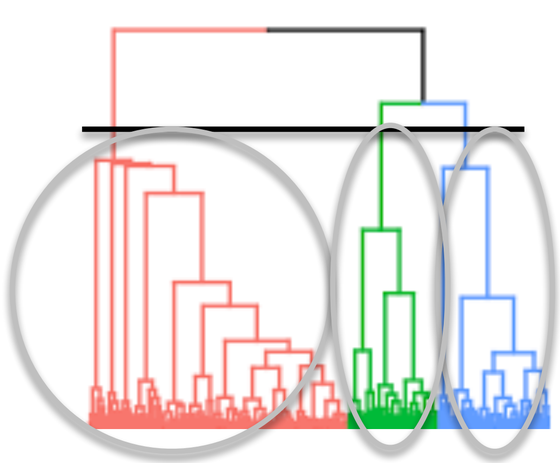
step 4

Divisive Hierarchical Clustering

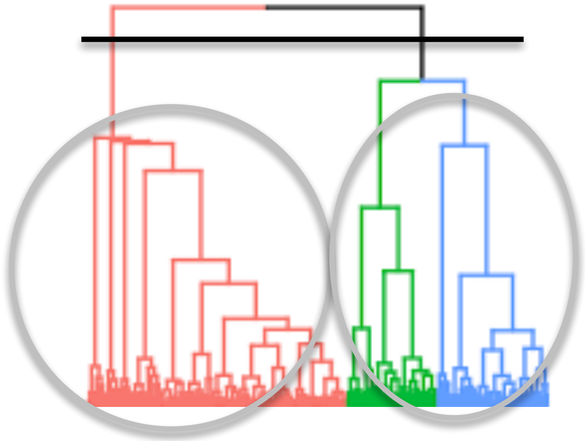
- Here each observation is considered to be in a single cluster.
- At each step a cluster is split; this process continues until we end up with each observation in a cluster or a collection of singleton clusters.
- It’s a top-down approach.
- Divisive Hierarchical clustering is not a very popular technique.

Hierarchical Clustering

No need to decide the number of clusters at the beginning of analysis. The clusters are formed by following a hierarchical structure or dendrogram and merging two clusters at each level of the structure.



3 clusters



2 clusters

Similarity or distance matrix is used for merging two clusters at a level.

This is not suitable for large dataset.

Hierarchical Clustering

Similarity Or Dissimilarity Matrix

The total number of observations ’n’ forms the order of the matrix.

Observations are arranged across the row and column.

Each cell C_{ij} contains the distance or difference between the i^{th} row observation and j^{th} column observation.

Dissimilarity matrix is a symmetric matrix of order n.

Let us assume we have 5 currency notes of denominations Rs 10, Rs 20, Rs 50, Rs 100 and Rs 200.

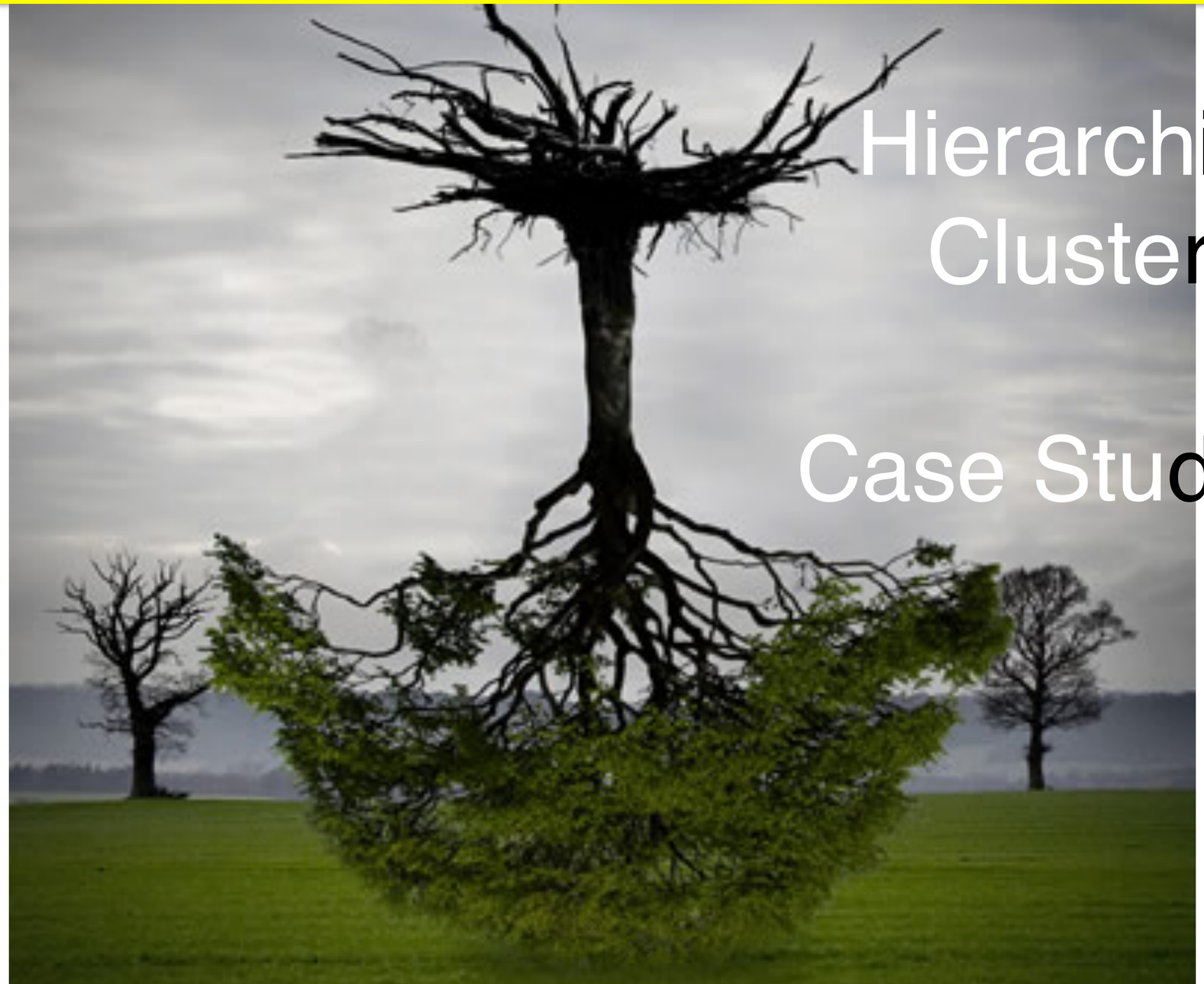
Denominations		10	20	50	100	200
	Observation	0	1	2	3	4
10	0	0				
20	1	20-10=10	0			
50	2	50-10=40	50-20=30	0		
100	3	100-10=90	100-20=80	100-50=50	0	
200	4	200-10=190	200-20=180	200-50=150	200-100=100	0

Agglomerative Clustering Algorithm

Basic algorithm:

Compute the distance matrix between the input data points
Every observation is a cluster

Repeat
    Merge the two closest clusters
    Update the distance matrix
Until only a single cluster remains



Hierarchical Clustering

Case Studies

Applications Of Hierarchical Clustering - Healthcare

Did you know hierarchical clustering even has a use case in a court case in the US?

A gastroenterologist was convicted of attempted second-degree murder by injecting his former girlfriend with blood or blood-products obtained from an HIV type 1 (HIV-1)-infected patient under his care. Phylogenetic analyses of HIV-1 sequences were admitted and used as evidence in this case, representing the first use of phylogenetic analyses in a criminal court case in the United States. *Source: pnas*

Can hierarchical clustering be used to find out where a viral outbreak originated?

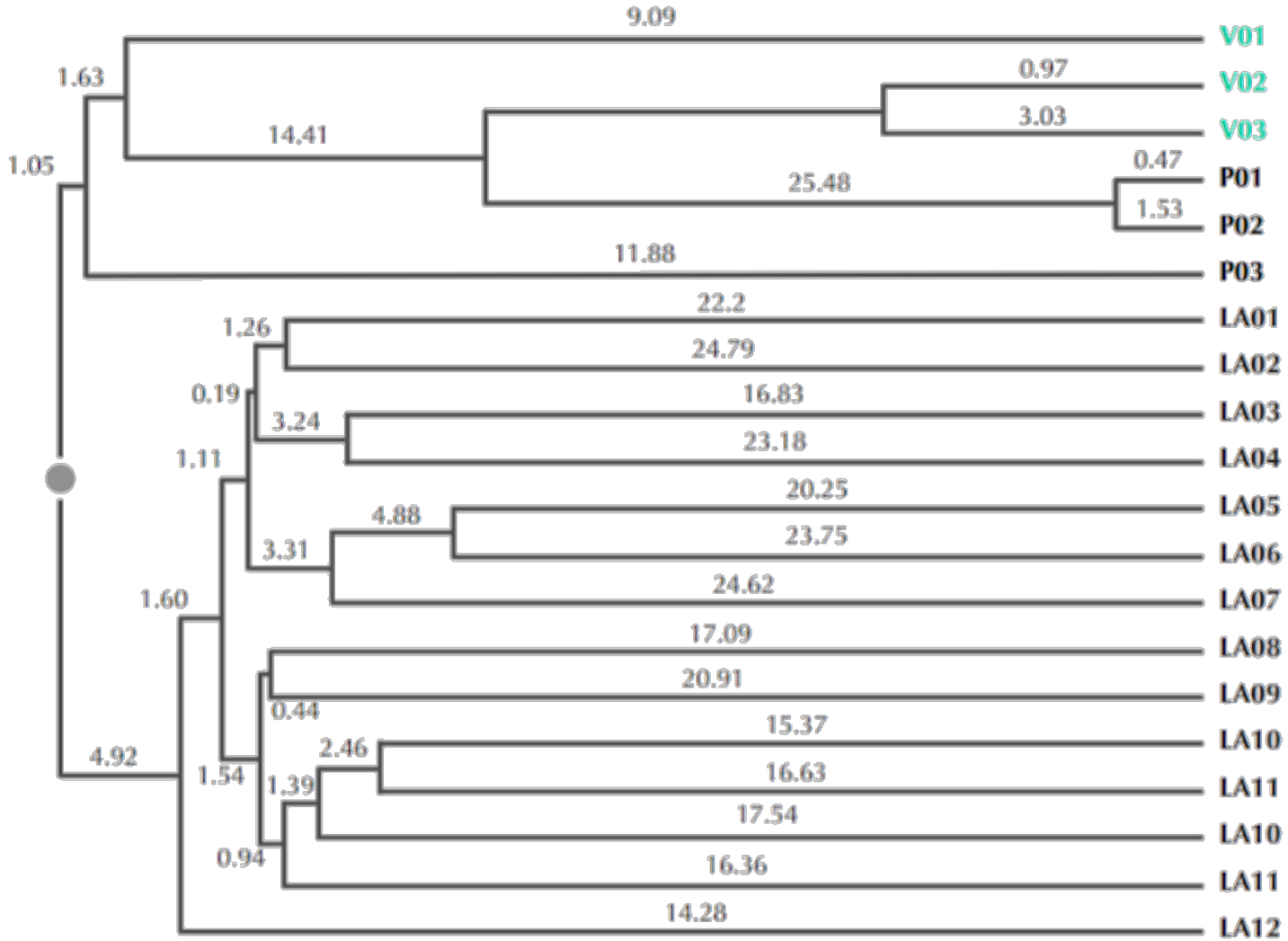
Applications Of Hierarchical Clustering - Healthcare

One of the major challenges in healthcare is tracking the source of viral outbreaks. This will help the scientists to treat the cause at the source. Also it will give data to them about why and how the outbreak began so that they can come up with a preventive solution, which will save millions of lives.

Viruses such as HIV have high mutation rates. This means the similarity of the DNA sequence of the same virus depends on the time since it was transmitted. This can be used to trace paths of transmission.

In the court case we talked earlier this method was used as an evidence. The victim's strand of HIV was found to be more similar to the given patient's strand, compared to a control group.

Applications Of Hierarchical Clustering - Healthcare



Source: towardsdatascience

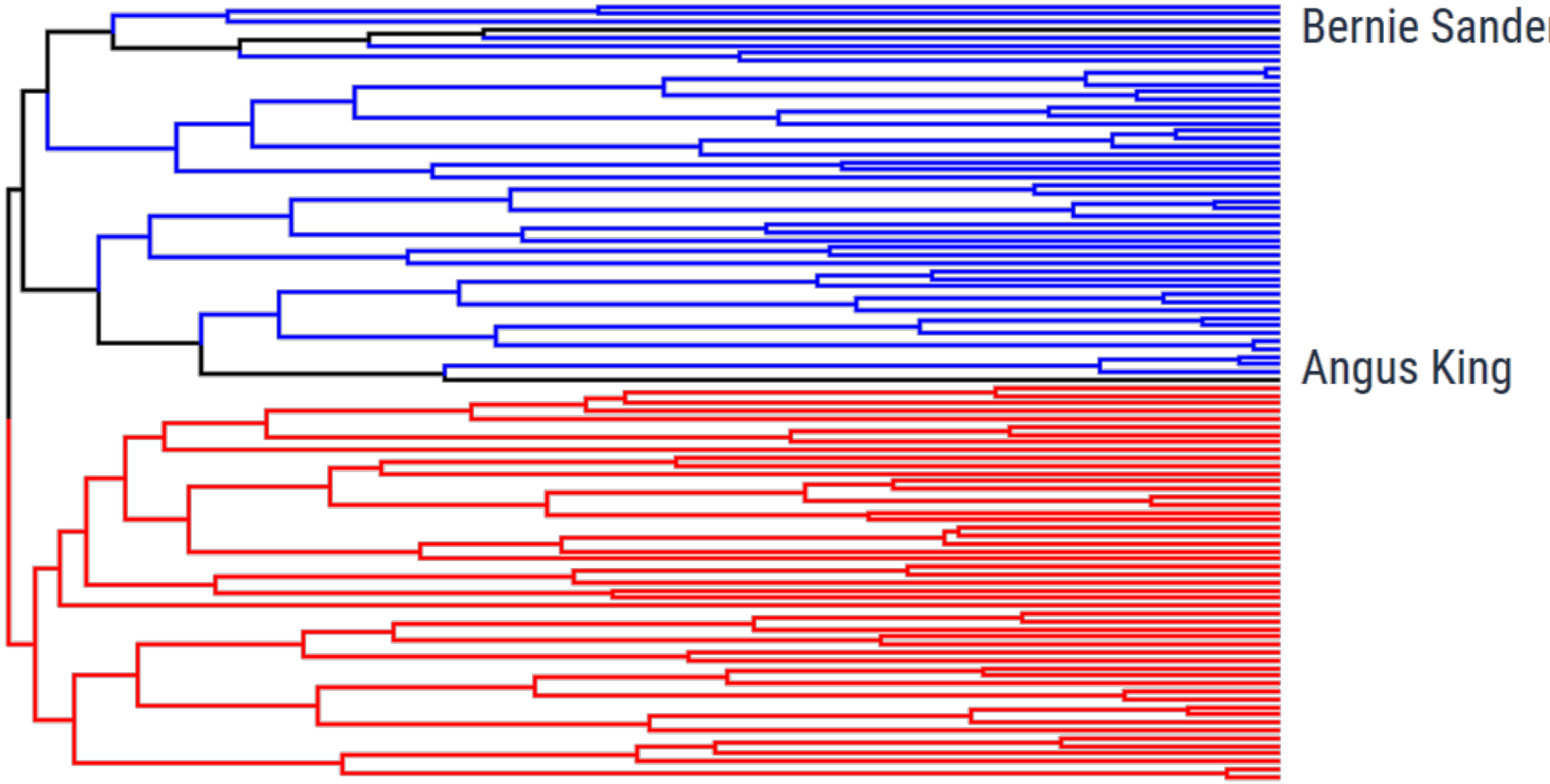
Applications Of Hierarchical Clustering - Social Network

I found a great use case for hierarchical clustering at towardsdatascience. They used Twitter to cluster US senators into their respective parties. This means analysing ones social media presence we can understand his or her political alliance. Interesting isn't it?

They did a simple analysis. They only looked at which senators follow which senators. Those who are familiar with Twitter can easily understand what following means in Twitter. They came up with a graph structure with senators as the nodes and follows as the edges.

They used Walktrap algorithm by Pons et al. on this graph. The algorithm takes a random walk through the graph. They give similarity value to a senator by the number of times the algorithm comes to a senator starting from a senator traversing the edges.

Applications Of Hierarchical Clustering - Social Network



In order to measure the performance of their clustering with greater clarity, coloring the results helped a lot. We can see that Democrats and Republicans are very clearly split from the top in the social space.

Source: towardsdatascience

Reds are Republicans, **Blues** are Democrats, **Blacks** are independent

That’s all for today

#HappyLearning