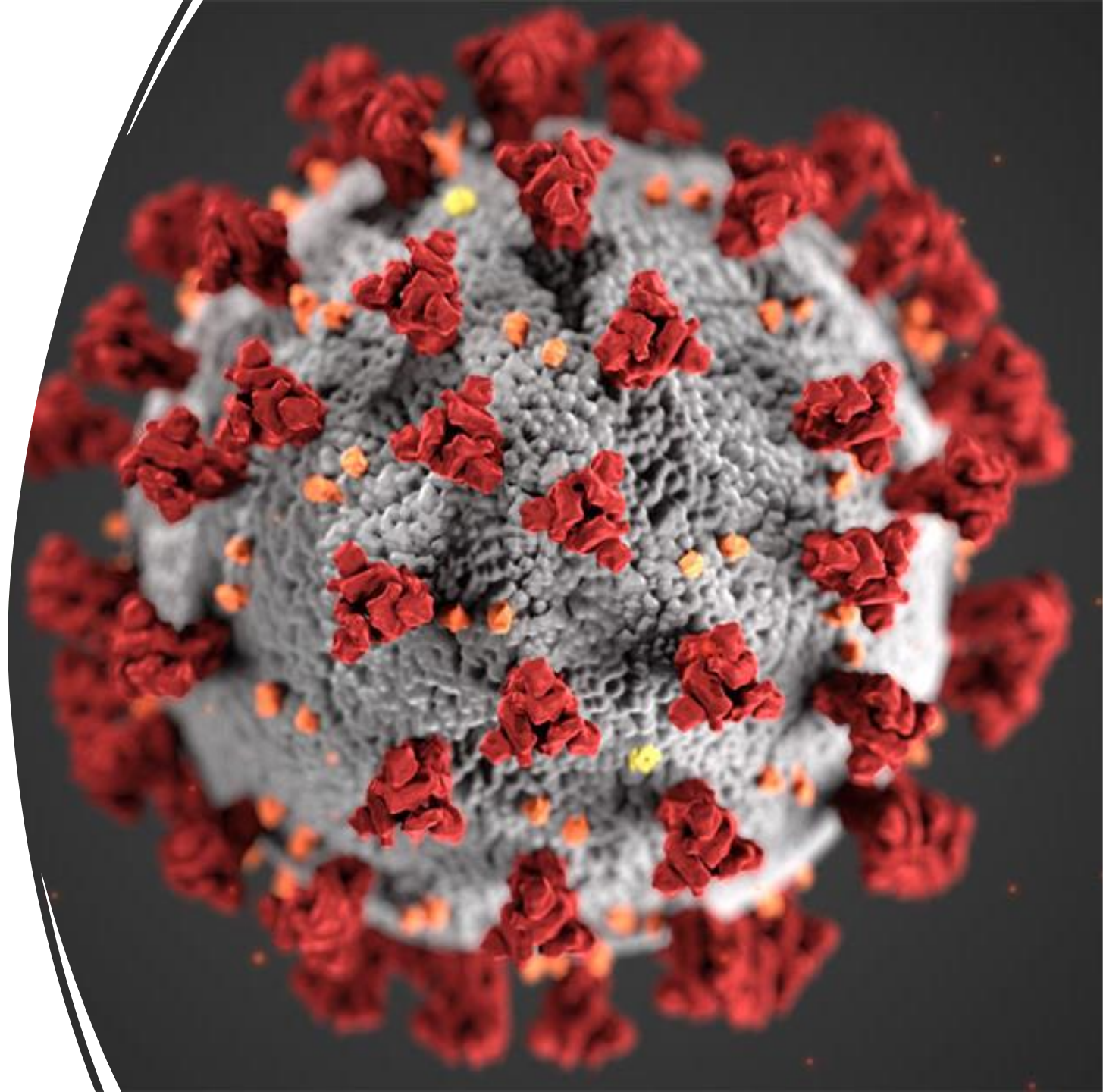


# Effective and Scalable Clustering of SARS-CoV-2 sequences

---

## Authors:

- SARWAN ALI (GSU, USA)
- TAMKANAT-E-ALI (LUMS, Pakistan)
- MUHAMMAD ASAD KHAN (Hazara University, Pakistan)
- IMDADULLAH KHAN (LUMS, Pakistan)
- MURRAY PATTERSON (GSU, USA)



## Table of contents

Sr.No	Topics
1.	Research Motivation
2.	Problem Introduction
3.	Research Applications
4.	Objectives
5.	Dataset
6.	Proposed Approach
7.	Data Visualization
8.	Results
9.	Conclusion
10.	Future Work
11.	References

## Research Motivation

- SARS-CoV-2, like any other virus, continues to mutate as it spreads, according to an evolutionary process and it is important to study the different mutations in order to control their spread.
- Identifying variants is an important part of understanding the evolution of a virus. The number of currently available sequences of SARS-CoV-2 in public databases such as GISAID is several million that cannot be processed by traditional methods, so new and scalable methods will need to be devised in order to analyze the ever-increasing number of viral sequences.
- In order to process the amino acid sequences using machine learning methods keeping the order of amino acids under consideration, the alphabet vectors of the sequence need to be converted to numerical vectors.
- Previous methods like one-hot encoding turn out to be less effective while computing pairwise distance.
- To deal with this problem we need to come up with a different approach to preserve the order of each sequence.

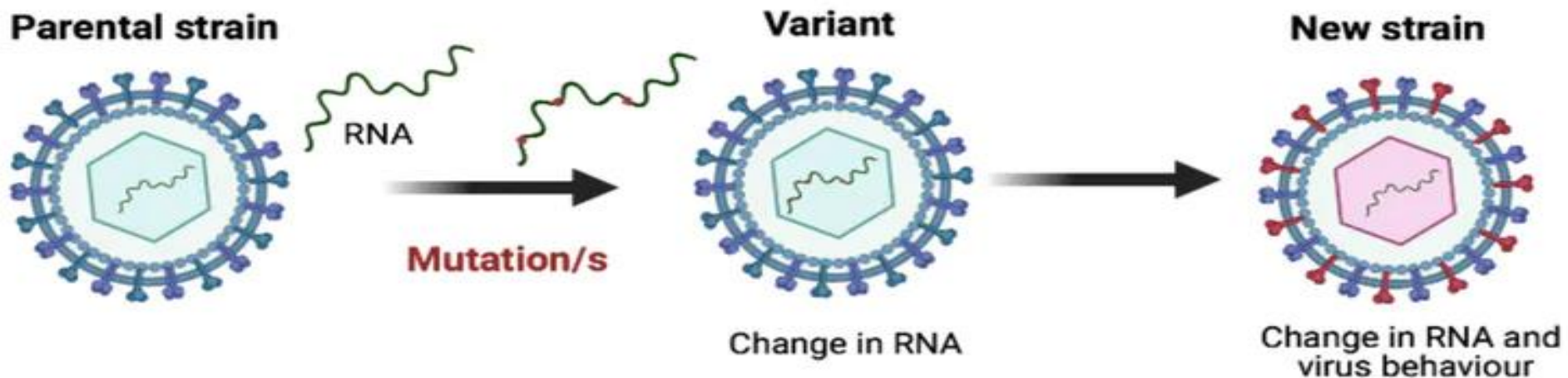


Figure 1: Virus mutation[3]

## Applications

- Identifying the variants can help in COVID vaccine design.
- Observing the distribution of SARS-CoV-2 variants can help in COVID Vaccine distribution decisions.
- Study of COVID variants and their spread help in forming strategies on how to monitor and prevent future outbreaks.

## Three types of coronavirus vaccines in development

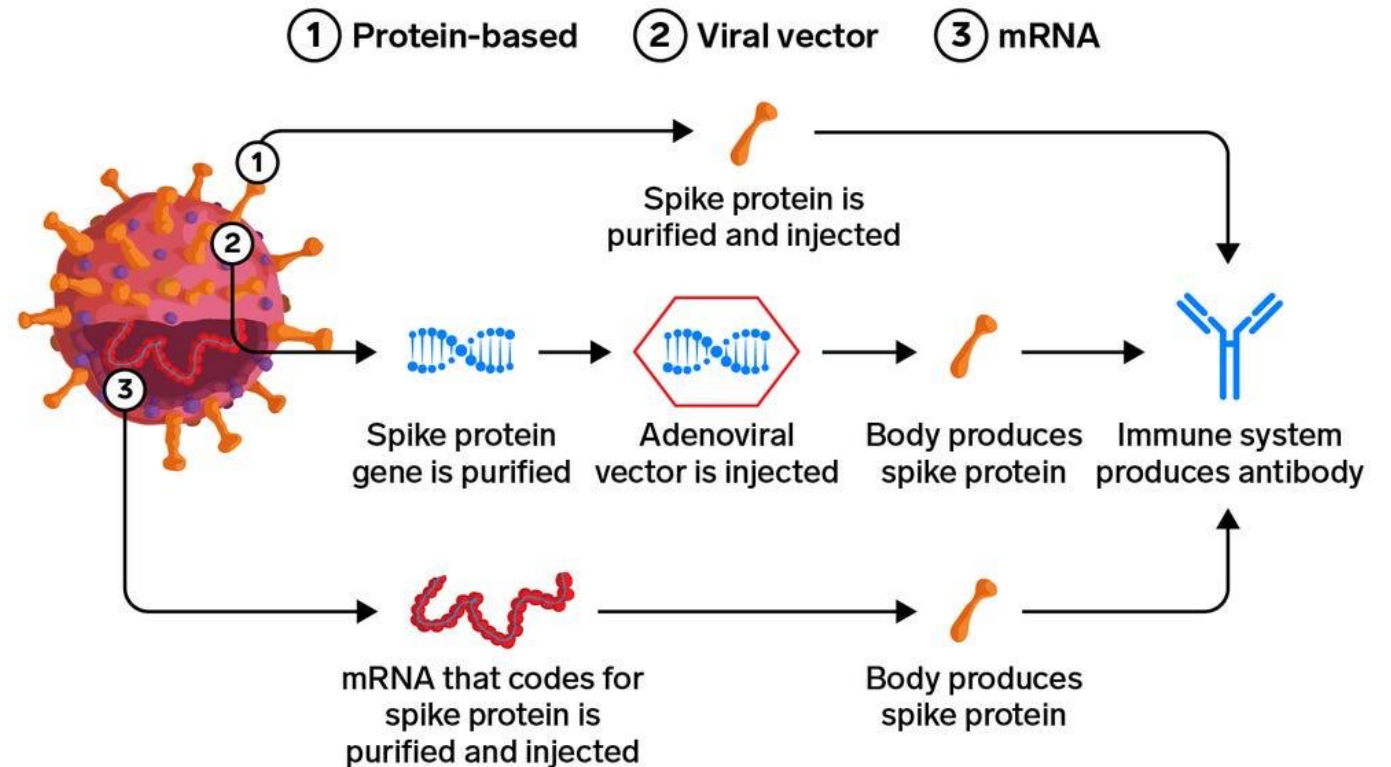
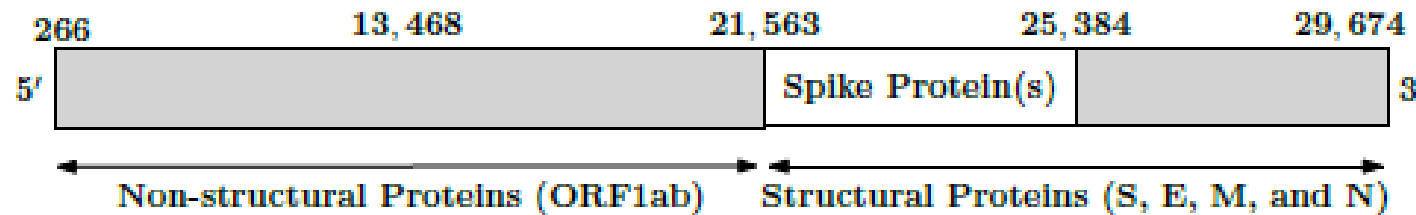


Figure 2: Types of Vaccine[2]

# Problem Introduction

- We propose an approach based on clustering sequences to identify the current major SARS-CoV-2 variants.
- The majority of the variation in the SARS-CoV-2 genome takes place in the spike region.
- The spike region encodes the spike protein an important function of the virus.



- Each sequence is first converted into length k substrings (called k-mers). These k-mers successfully preserve the order of each sequence, that can be crucial to the performance of the classification/clustering tasks.



# Objectives

- Propose a method based on k-mers for efficient SARS-CoV-2 sequence clustering. The following figure shows K-mers with different values of K:

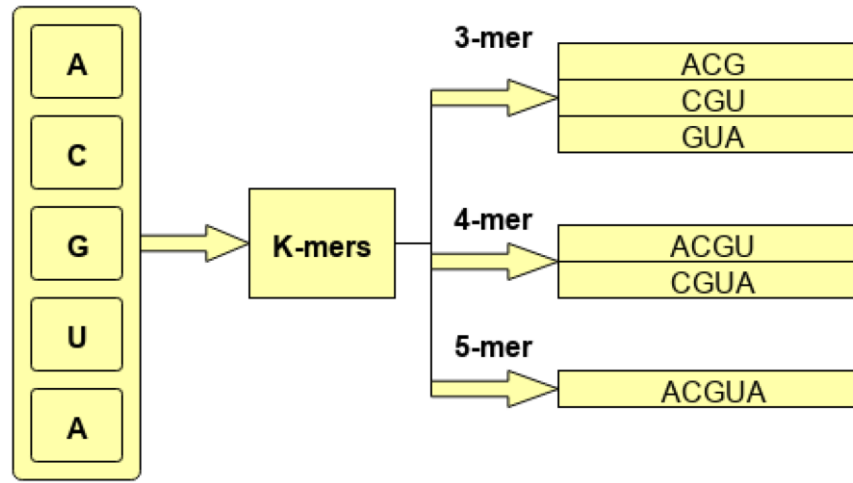


Figure 3: K-mers[1]

- Analyze the spread pattern of different variants of Covid over time, which can guide vaccination design and distribution strategies.
- Compute the importance of each amino acid position in distinguishing a variant.

## Dataset

- We used the (aligned) amino acid sequences corresponding to the spike protein from the largest known database of SARS-CoV-2 sequences, GISAID.
- The Variants information and distribution in the dataset is as follows:

Pango Lineage	Region	Labels	Num. Mutations S-gene/Genome	Num. of sequences
B.1.1.7	UK [19]	Alpha	8/17	13966
B.1.351	South Africa [19]	Beta	9/21	1727
B.1.617.2	India [43]	Delta	8/17	7551
P.1	Brazil [30]	Gamma	10/21	26629
B.1.427	California [44]	Epsilon	3/5	12784

- In our dataset, we have 5 most common variants known to date. After preprocessing data (removing missing values and truncated sequences), we end up with 62,657 amino acid sequences.

# Proposed approach

## k-mers Computation

- The first step is to compute all  $k$ -mers of each sequence to map it to a fixed length vector, while allowing its order to be preserved. Given a sequence, the total number of  $k$ -mers that can be generated is  $N - k + 1$ , where
- $N$  is the total number of elements in the sequence (1274 amino acids in our case)
- $k$  is a parameter for the size of each mer (In our case we use  $k = 3$  based on standard validation set approach)

## Frequency Vectors Generation

- The second step is to generate the numerical representation of these vectors. For this purpose, we design frequency vectors that contain the counts of each  $k$ -mer in the corresponding sequence.

## Protein sequence with 3-mers

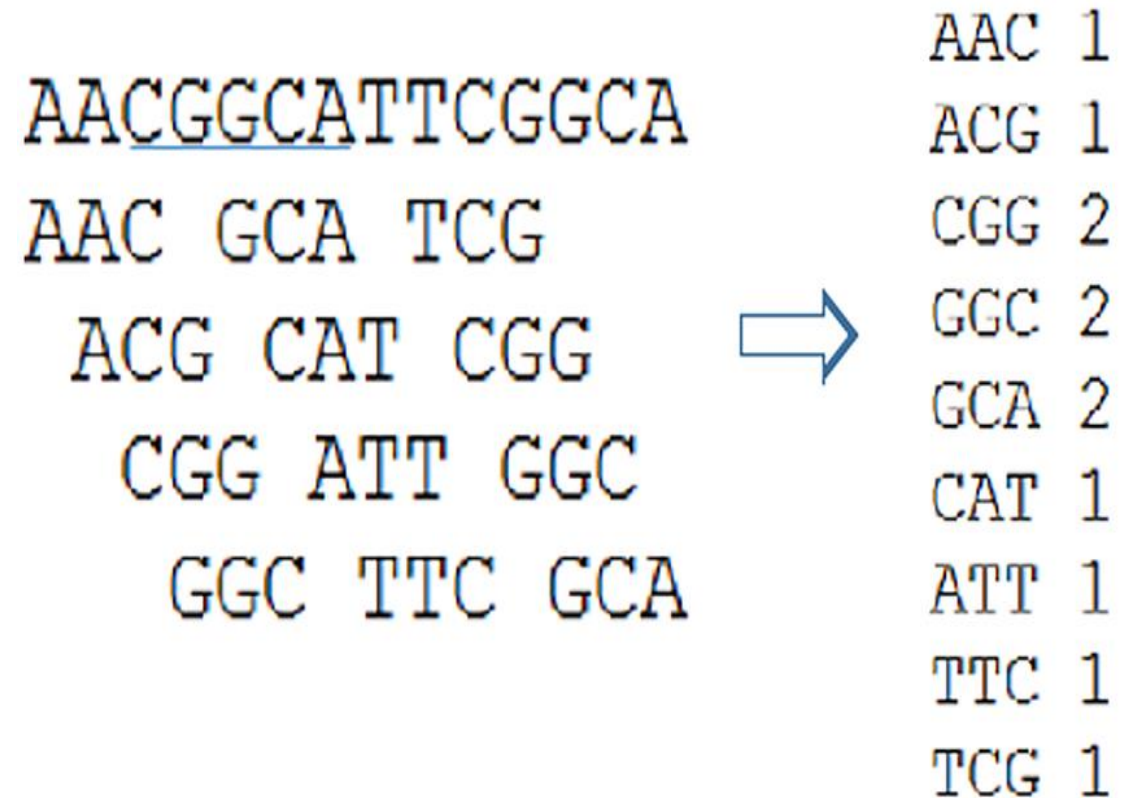


Figure 4: Protein Sequence k-mers counting ( $k=3$ )[5]



## Feature Selection

- In the third step we select the most important features in the feature vectors using two supervised feature selection methods:
  - **Ridge Regression**
    - It works by introducing a Bias term
      - The goal is to increase the bias to improve the variance (generalization capability)
    - It works by changing the slope of the line (can reduce the slope close to zero)
    - $\text{Min}(\text{Sum of square residuals} + \alpha * \text{slope}^2)$
  - **Lasso Regression**
    - It works by introducing a Bias term
      - The goal is to increase the bias to improve the variance (generalization capability)
    - It works by changing the slope of the line (can reduce the slope exactly equals to zero)
    - $\text{Min}(\text{Sum of square residuals} + \alpha * |\text{slope}|)$

## K-means based Clustering

- In this step we use the  $K$ -means clustering algorithm to cluster the data. Since we have 5 variants in our data, we used  $K = 5$  in  $K$ -means.

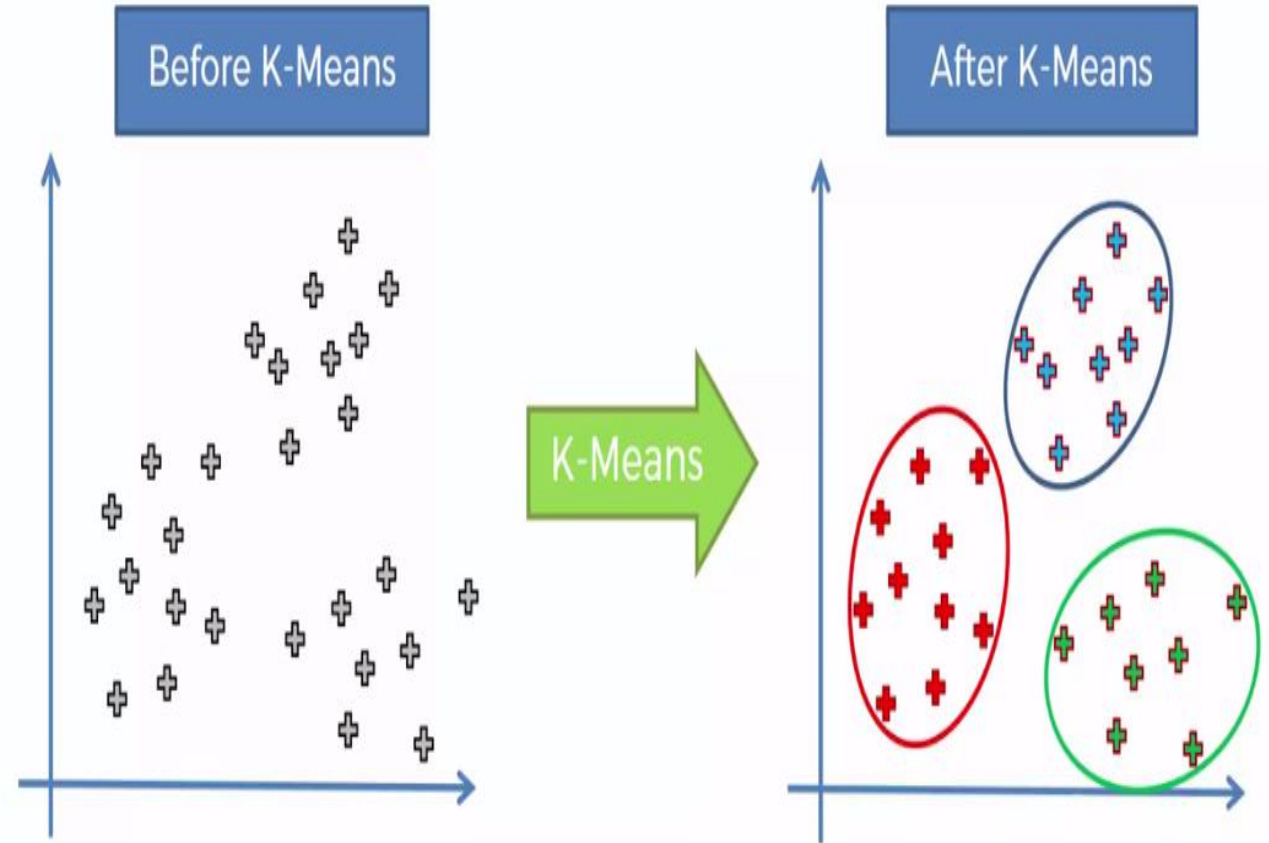
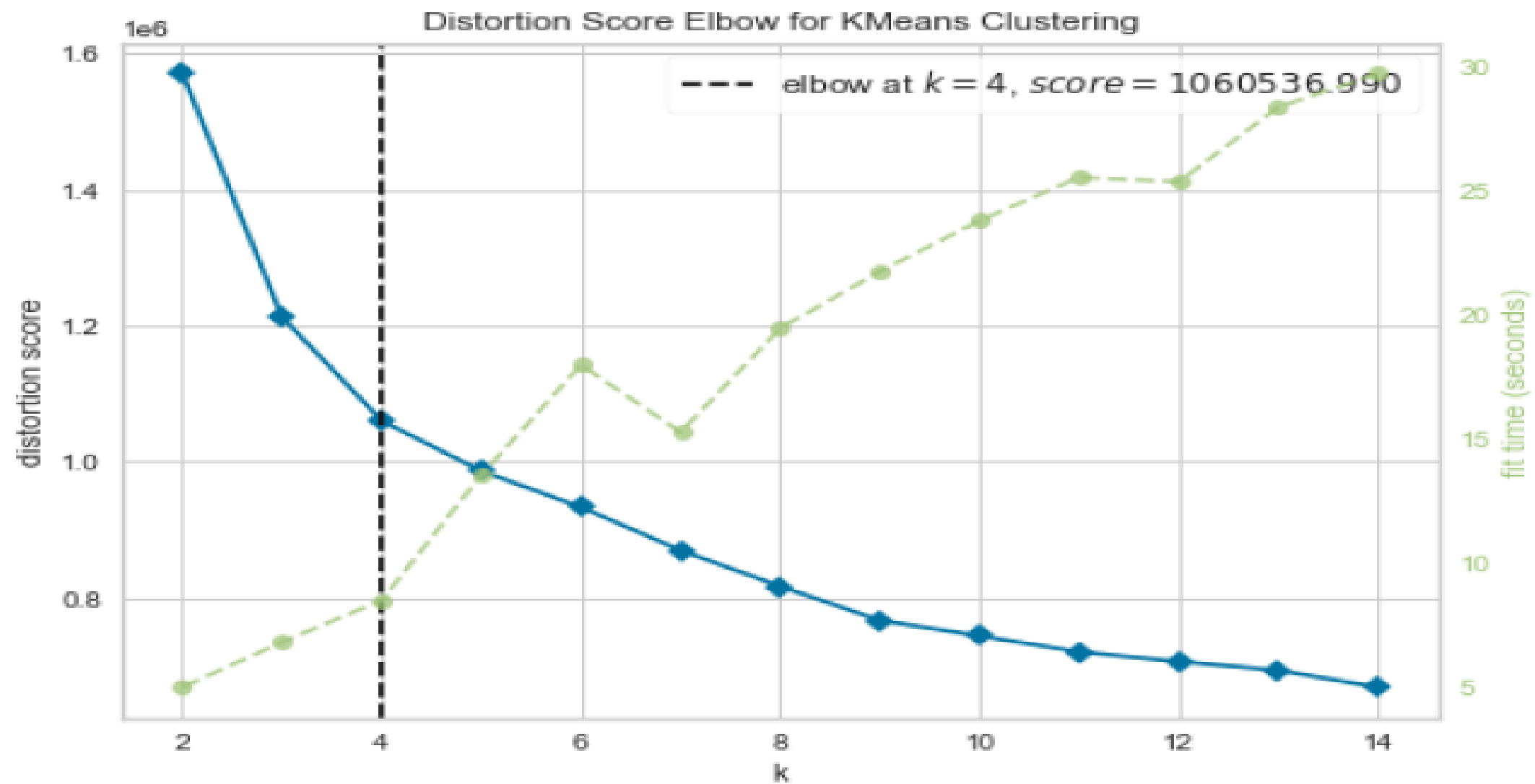





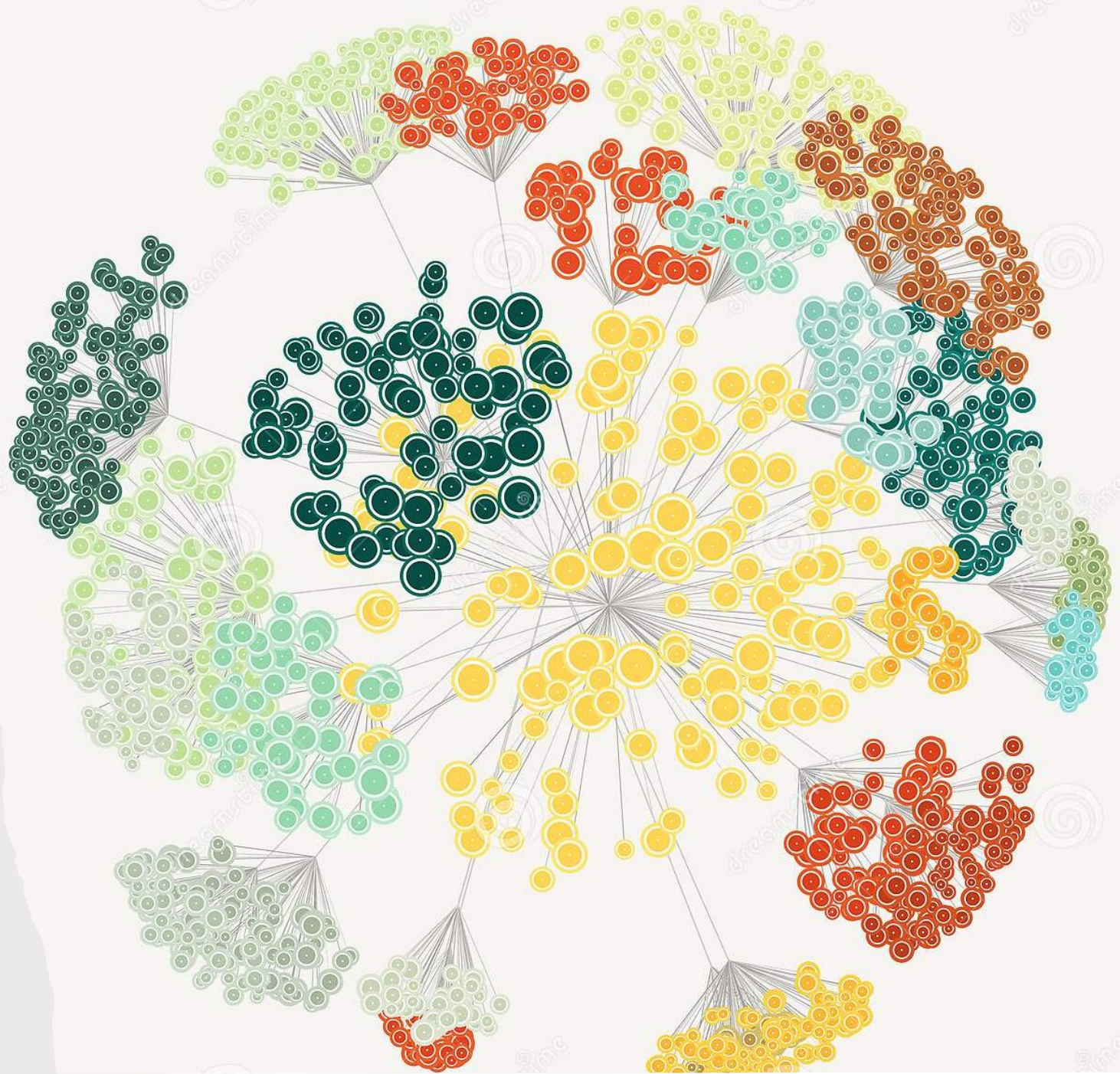
Figure 4: K-Means Clustering [4]



Key:

-  Optimal number of cluster computed using Elbow Method
-  The runtime (sec) for different number of clusters
-  Distortion score

# Data Visualization





## Data Visualization using t-SNE plot

- To see if there is any (hidden) clustering in the data, we mapped the data to 2D real vectors using the t-SNE approach.
- t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique used for visualizing high-dimensional data which helps analyzing the arrangement of data.
- t-SNE is incredibly flexible and can often find structure where other dimensionality-reduction algorithms are not successful, the algorithm also makes all sorts of adjustments that tidy up its visualizations.

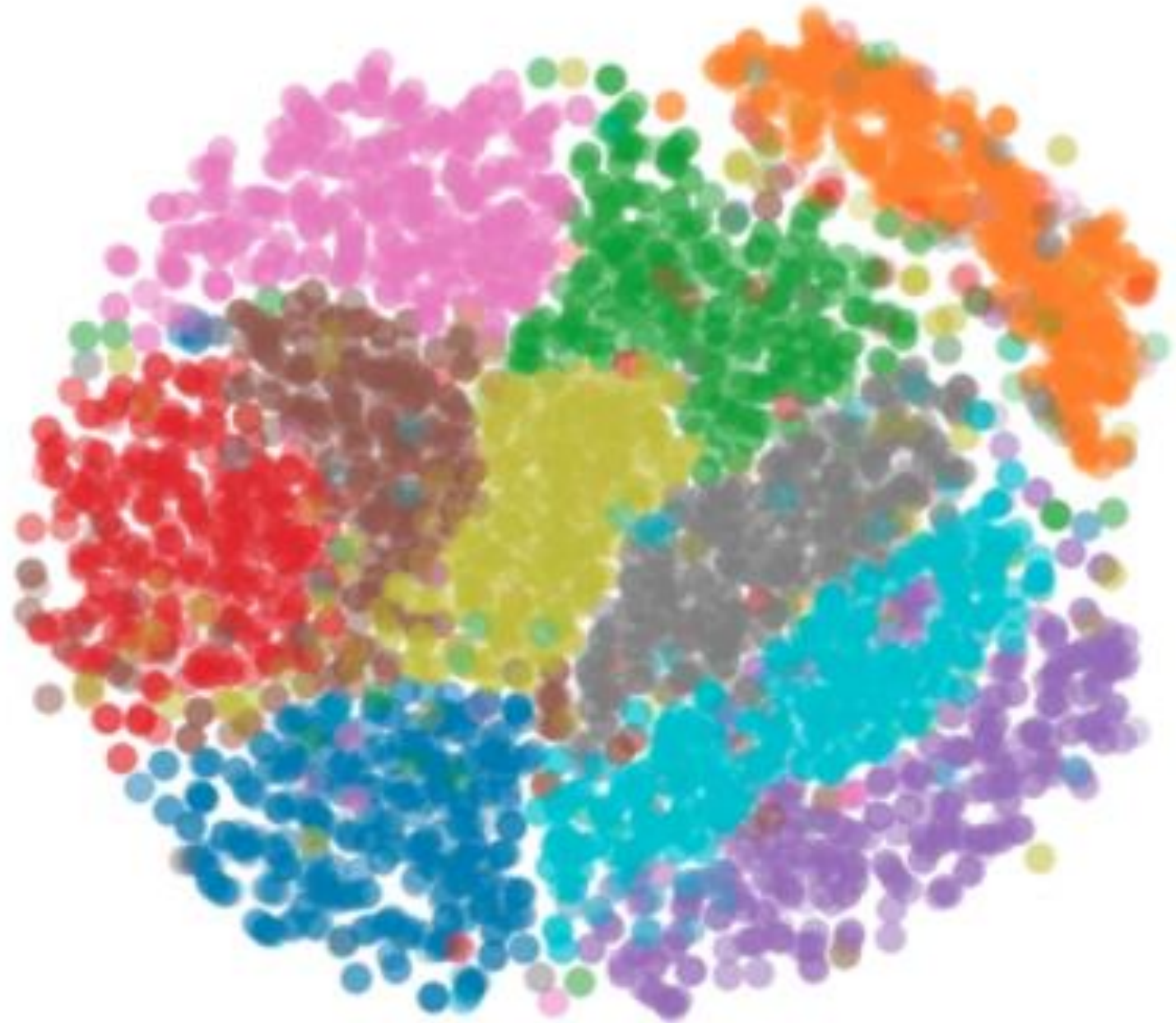
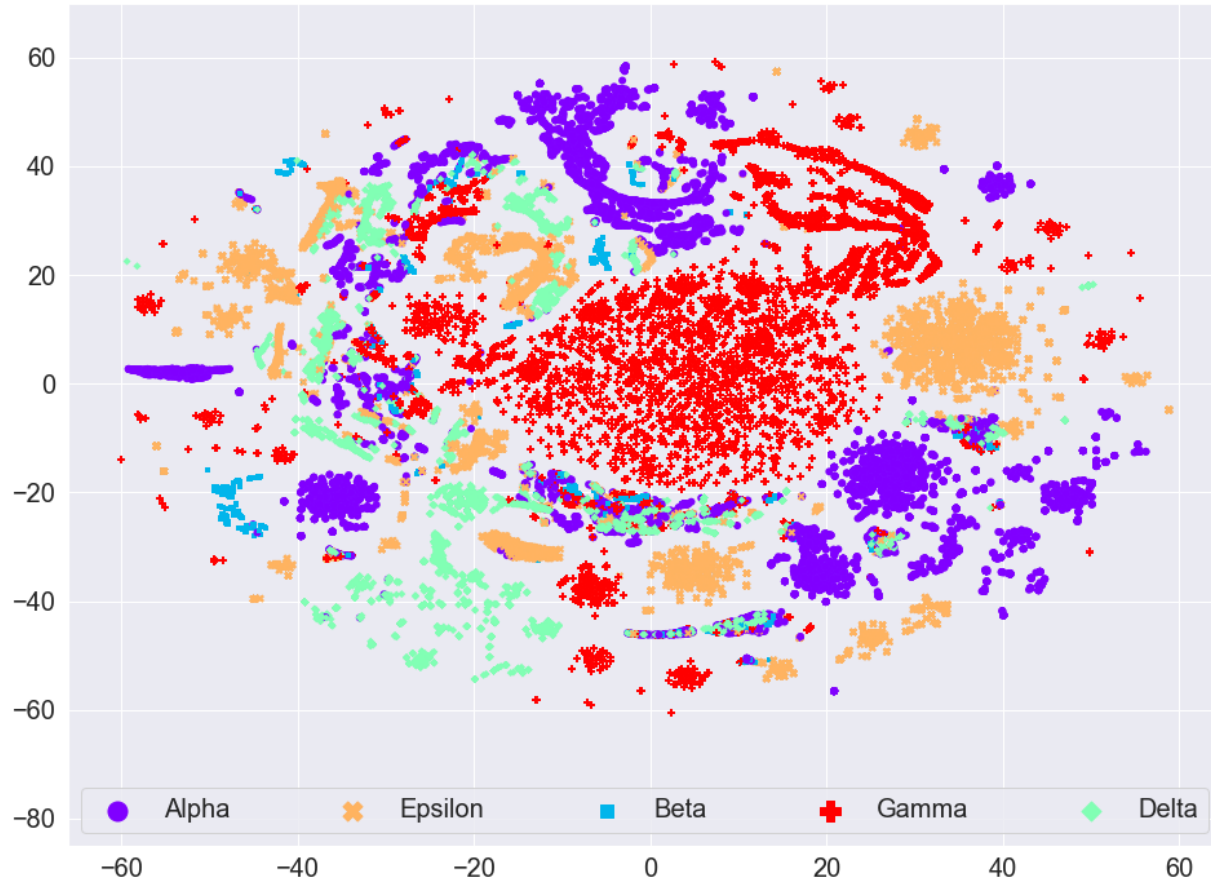
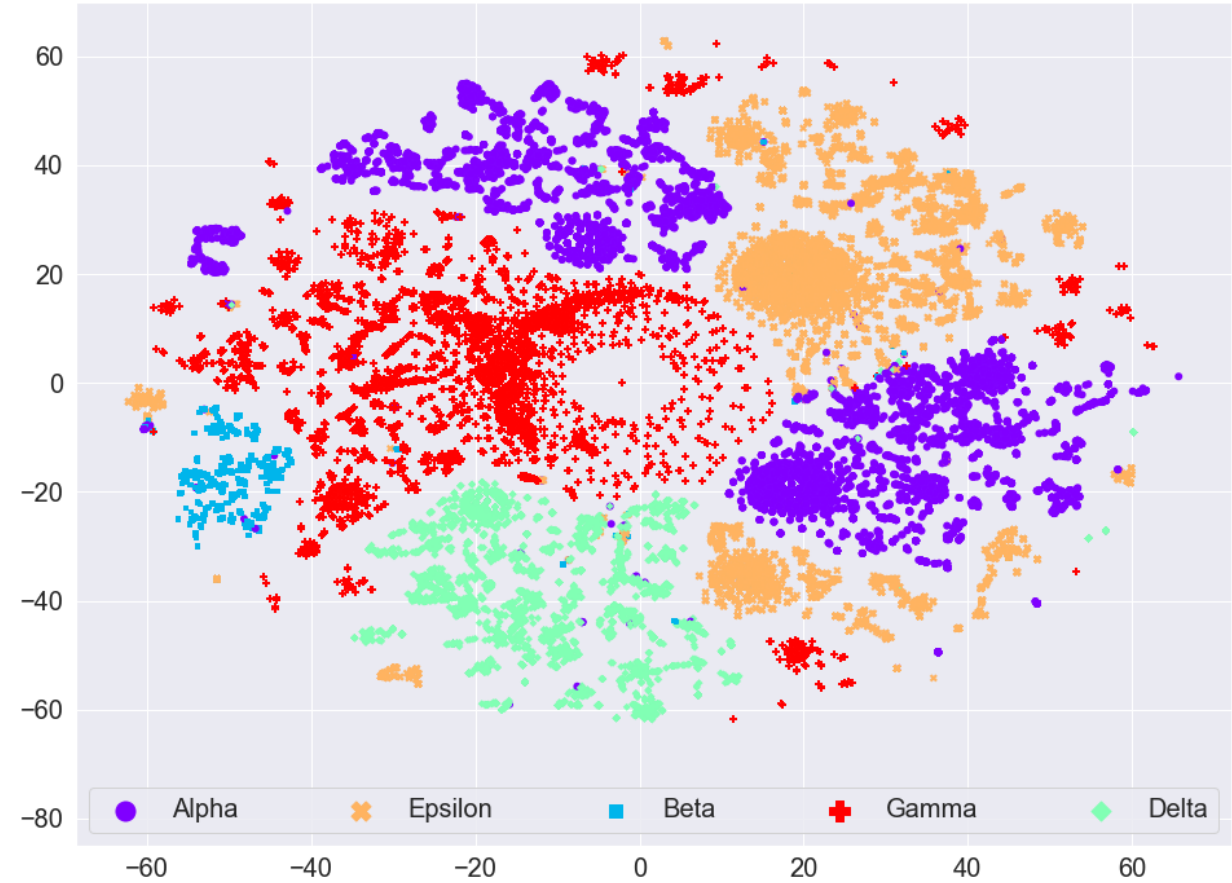


Figure 5: MNIST t-SNE plot[6]

# Variant data visualization using t-SNE plot



Before Lasso Regression



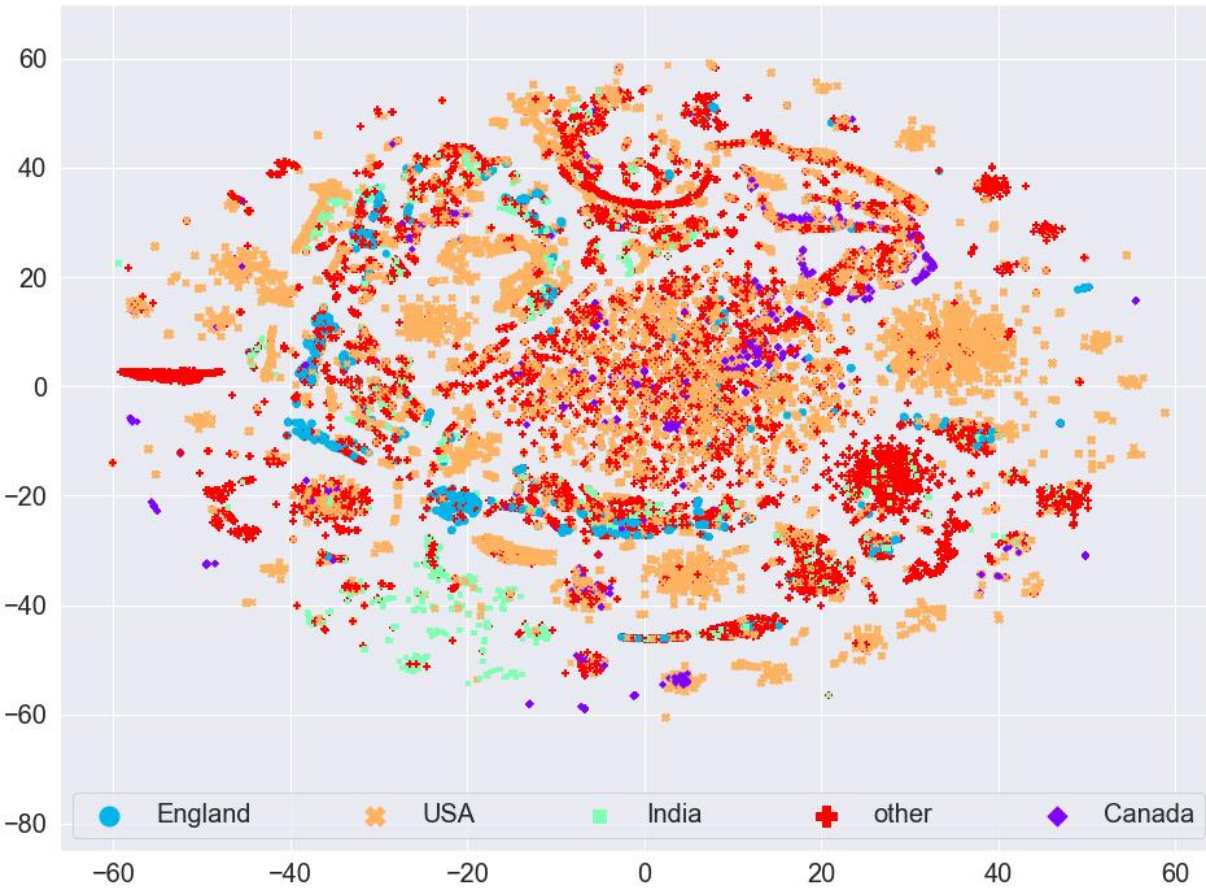
After Lasso Regression

## Observation:

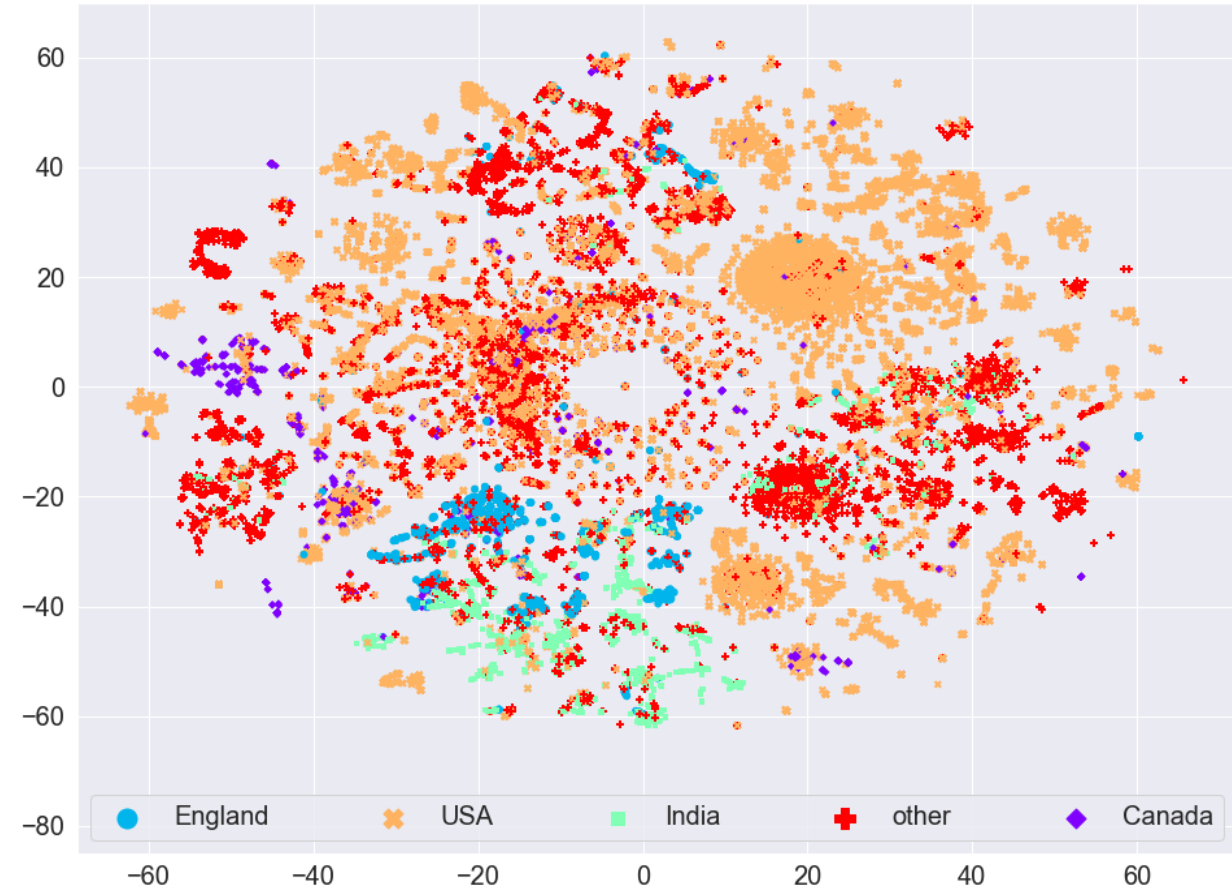
The clusters are more clearly visible after applying Lasso Regression feature selection as compared to clustering on frequency vectors without Lasso Regression.



# Country data visualization using t-SNE plot



Before Lasso Regression



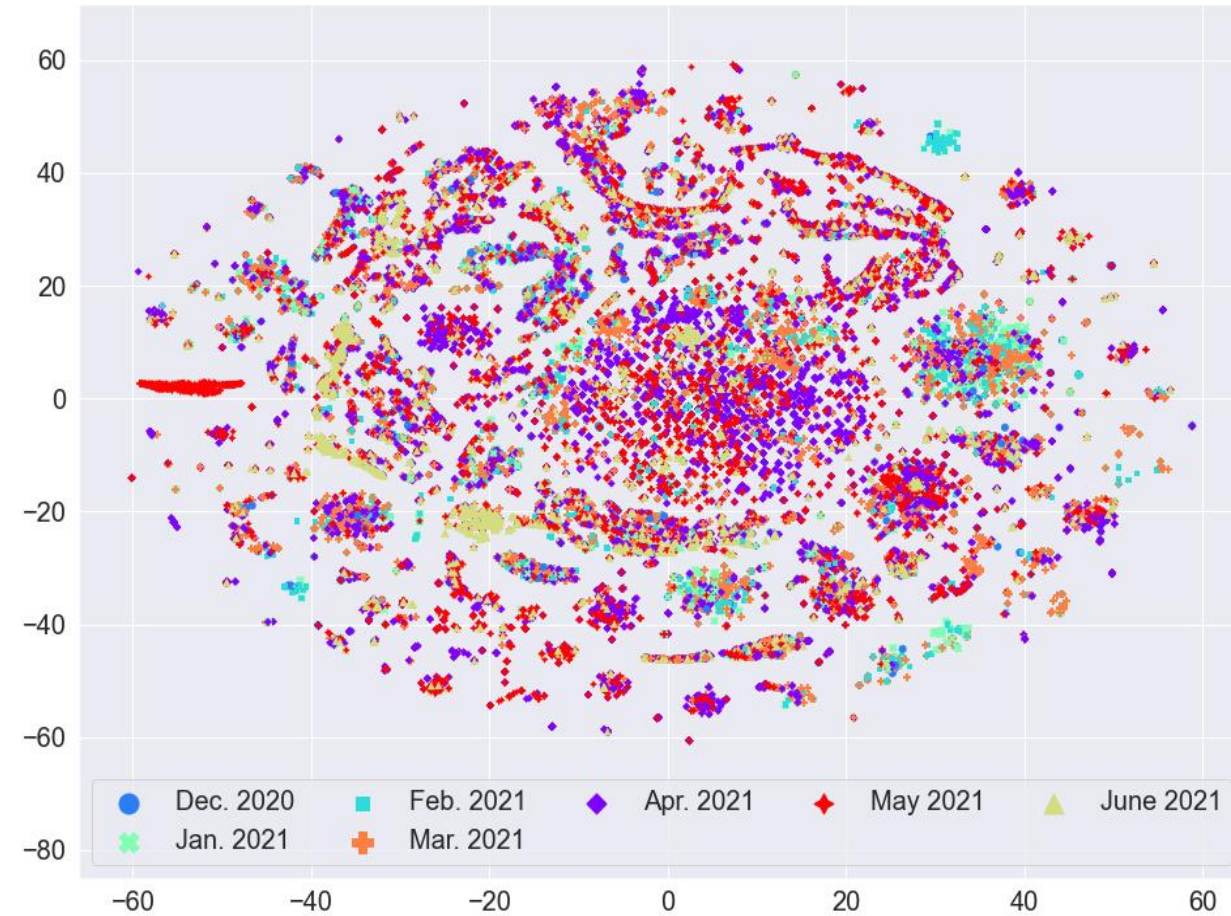
After Lasso Regression

## Observation:

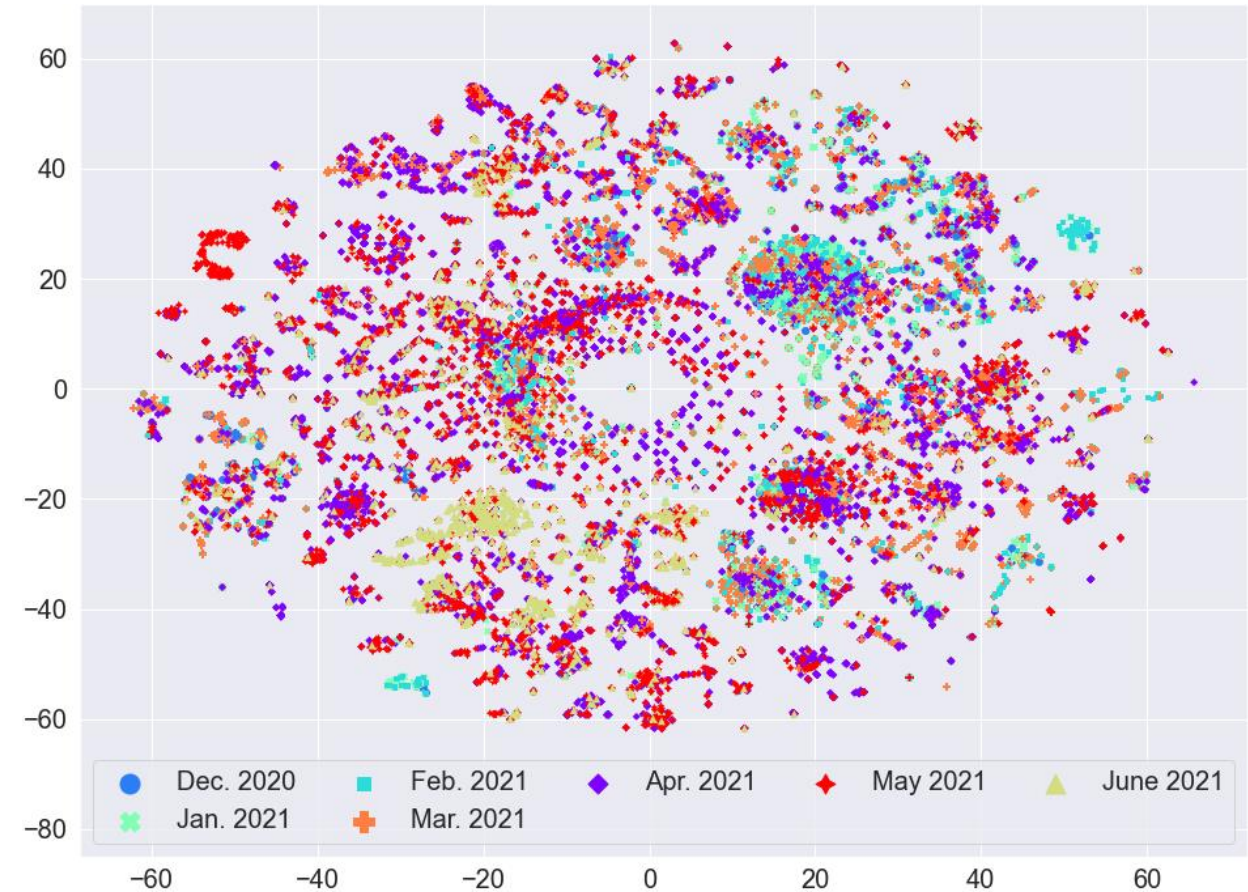
There is no clear cluster for any country even after applying Lasso Regression feature selection method. A country can have patients affected by different variants of the coronavirus so we cannot expect any sort of clustering in this case.



# Months based data visualization using t-SNE plot



Before Lasso Regression




After Lasso Regression

## Observation:

There is no clear clustering for any month in the data even after applying Lasso regression feature selection method. We can say that the spread of coronavirus has nothing to do with different seasons over the years.





# Results

## Clustering quality computation

- To compute the quality of a clustering, we use the (weighted) F1 score. We label each cluster using the variant which is in majority in the cluster.
- Using these assigned labels, we compute the F1 score for each variant separately. The variant wise (weighted) F1 scores for different clustering methods are as follows:

Methods	F1 Score (Weighted) for Different Variants				
	Alpha	Beta	Delta	Gamma	Epsilon
K-means	0.3598	0.1070	0.6110	0.6908	0.6527
K-means + Ridge	0.9992	0.0058	0.8643	0.9998	0.7748
K-means + Lasso	0.9987	0.2705	0.9991	0.9998	0.9704

### Observations:

- F1 scores for  $K$ -means + Ridge are better than simple  $K$ -means, however, they are not better than  $K$ -means + Lasso for the majority of the variants.
- There is a low F1 score for the Beta variant in all methods, this is due to the much smaller number of sequences available for the Beta variant.



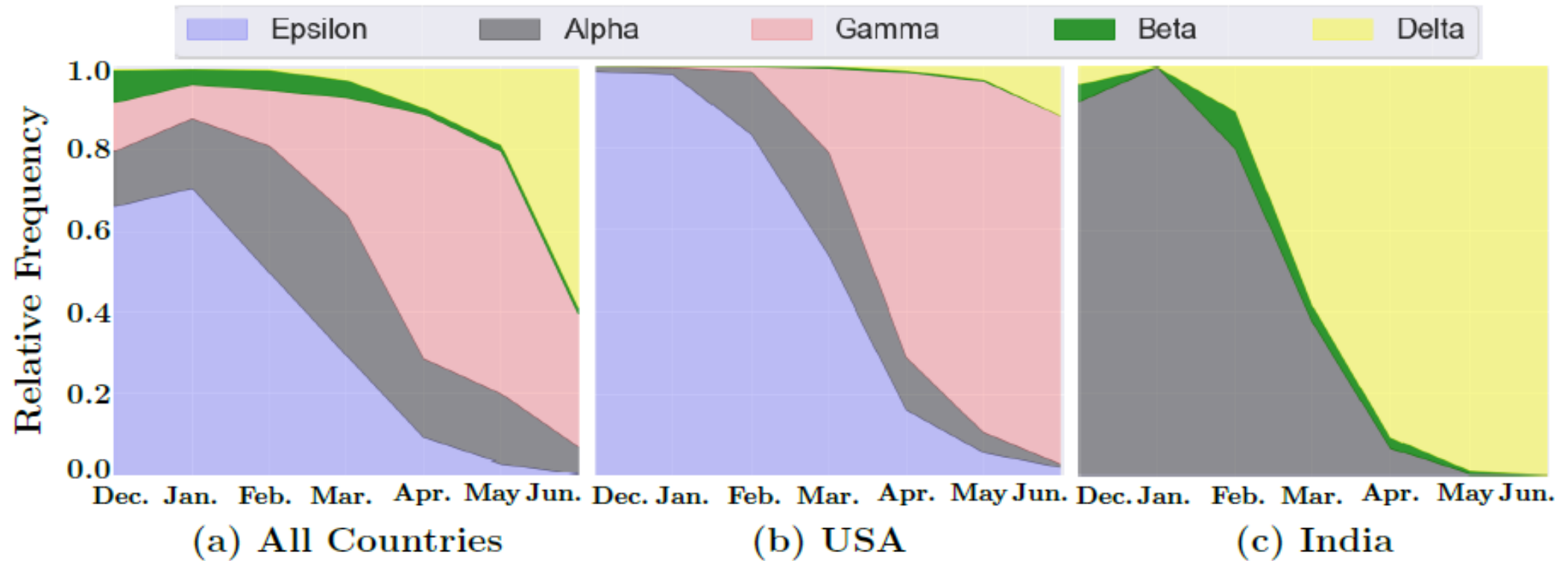
# Contingency tables of variants vs clusters

## Observations

- We can see that the variants are not clearly clustered into separate groups if we only use  $K$ -means without any feature selection method.
- With the feature selection methods (Ridge and Lasso Regression), we can clearly observe that different variants are grouped into their respective clusters.
- We can also observe that Lasso Regression based feature selection gives us more accurate clusters than Ridge Regression based feature selection.
- Although Lasso Regression gave us 964 features out of 4977 as compared to 1242 features given by Ridge Regression, those 964 features are a more accurate representation of the original data

Variants	K-means (Clust. IDs)					K-means + Ridge (Clust. IDs)					K-means + Lasso (Clust. IDs)				
	0	1	2	3	4	0	1	2	3	4	0	1	2	3	4
Alpha	1512	8762	86	680	2926	6	11415	310	344	1891	6	74	13622	258	6
Beta	295	601	33	172	626	1	4	1596	109	17	4	37	10	1673	3
Epsilon	956	7848	187	638	3155	0	1	8688	654	3441	0	4076	1	8705	2
Delta	2706	2605	30	868	1342	0	0	3126	3996	429	0	111	0	45	7395
Gamma	682	22140	50	741	3016	26426	13	16	147	27	26566	26	24	12	1

# Spread of Variants over time



**Technique:** We computed the relative frequencies of each variant for different months separately.

**Observations:** This study helps us analyze which variants are increasing in number over time.

- Initially, Delta variant was very rare. However, we can observe that it is spreading very quickly.
- The behavior of the Epsilon variant is opposite to the Delta variant. Initially Epsilon variant was high in numbers in USA but with time as the vaccination process speeds up, we can see the drop in the Epsilon variant.
- It can be observed that Beta variant is not spreading at an alarming rate.
- From February 2021 onward, the Gamma variant is spreading at an alarming rate.
- Initially in India the Alpha variant was the variant of concern. However, after January 2021, Delta becomes the dominantly spreading variant.

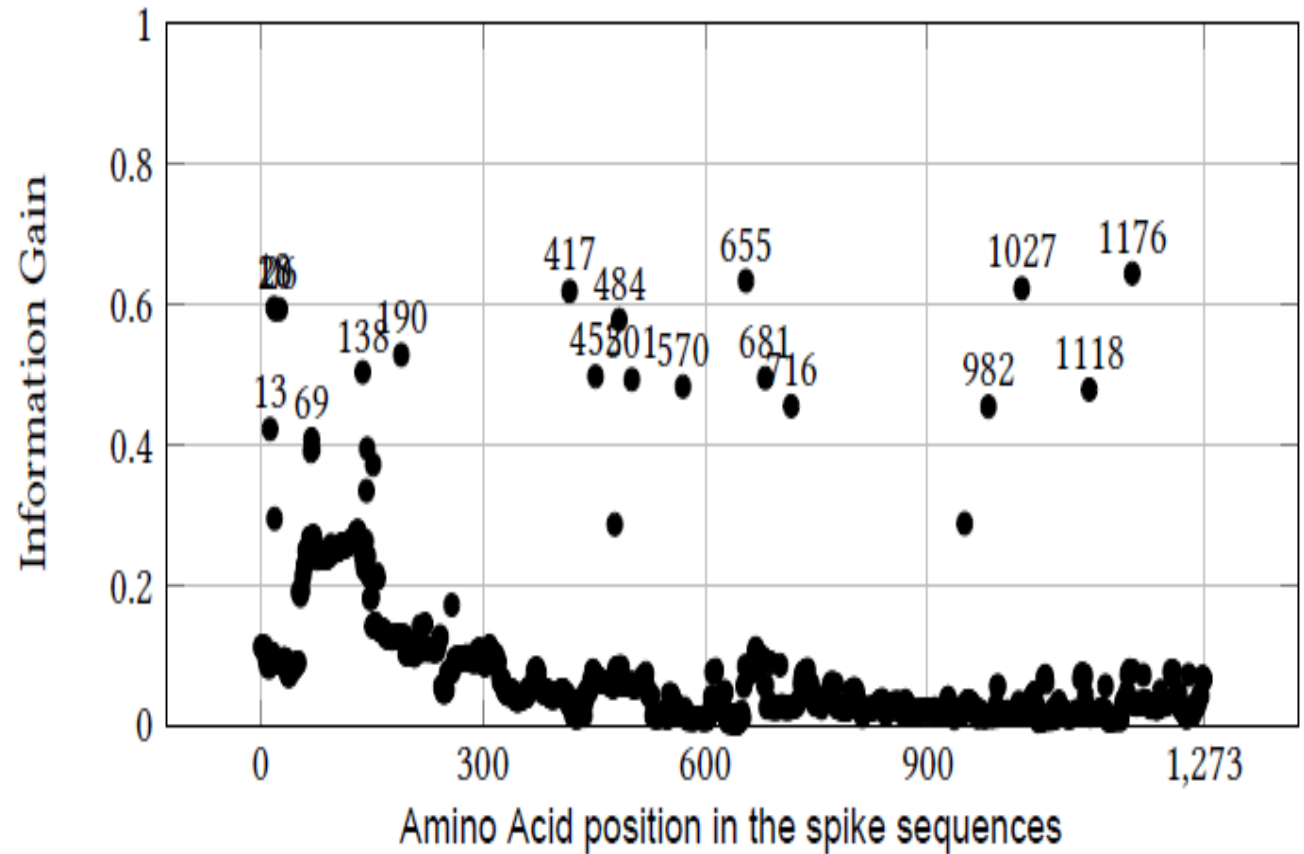
# Importance of each amino acid

## Technique:

- We compute Information Gain (IG) between each attribute (amino acid position) and the class (variant).

$$\text{IG}(\text{Class}, \text{Position}) = H(\text{Class}) - H(\text{Class} | \text{Position})$$

- The USA's Centers for Disease Control and Prevention (CDC) declared mutations at certain positions from one variant to the other. We use their mutation information to compare them with the attributes having high IG values in the figure above.

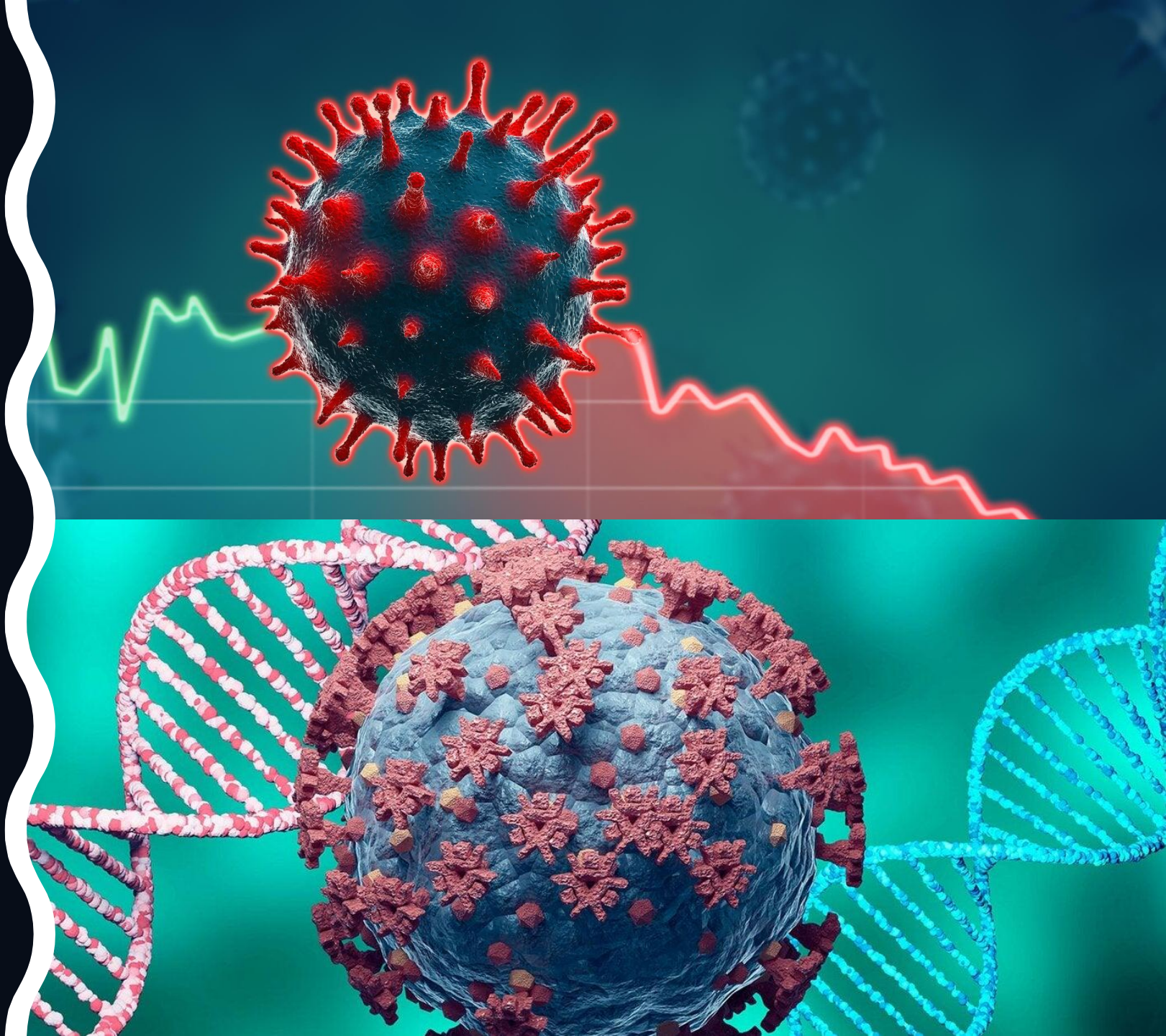


**Observation:** Our high IG value attributes are the same as given by CDC.



# Conclusion

- We propose a clustering-based approach to effectively cluster the variants of SARS-CoV-2 using amino acid sequences corresponding to the spike region.
- Our results show that efficient feature selection methods such as Ridge Regression and Lasso Regression greatly help improve the performance of K-means clustering.
- The rate of spread of each variant with time in different countries is shown.
- The importance of each amino acid position is shown by computing information gain and compare them with the CDC's provided information for these same positions.





## Future Work

- In the future, we will work towards using more SARS-CoV-2 sequences with more variants, as the data continues to accumulate.
- We will try using unsupervised feature selection methods.
- We will analyze the spread of the variants in larger number of countries.
- We will analyze the effect of vaccination on further mutations.



# External Resources

**Github:** [https://github.com/sarwanpasha/Visualization\\_covid\\_data](https://github.com/sarwanpasha/Visualization_covid_data)

**Dataset:** <https://www.gisaid.org/>

Questions!!

# References

- [1] Asim, Muhammad N., Muhammad I. Malik, Christoph Zehe, Johan Trygg, Andreas Dengel, and Sheraz Ahmed. 2020. "MirLocPredictor: A ConvNet-Based Multi-Label MicroRNA Subcellular Localization Predictor by Incorporating k-Mer Positional Information" *Genes* 11, no. 12: 1475.  
<https://doi.org/10.3390/genes1112147>
- [2] Bendix, A. (2020, November 9). *Pfizer's coronavirus vaccine relies on a new, unproven technology. A diagram shows how it differs from other candidates.* Business Insider. Retrieved September 22, 2021, from <https://www.businessinsider.com/leading-us-coronavirus-vaccines-how-they-work-compare-2020-10>.
- [3] Lara Herrero Research Leader in Virology and Infectious Disease, & Eugene Madzokere PhD Candidate in Virology. (2021, May 25). *What's the difference between mutations, variants and strains? A guide to COVID terminology.* The Conversation. Retrieved September 22, 2021, from <https://theconversation.com/whats-the-difference-between-mutations-variants-and-strains-a-guide-to-covid-terminology-154825>.
- [4] Jain, T. (2021, April 9). *K-means clustering.* Medium. Retrieved September 22, 2021, from <https://medium.datadriveninvestor.com/k-means-clustering-ac3ff1d3463d>.
- [5] Chu, Chong & Wu, Yufeng. (2016). REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads. *PloS one*. 11. e0150719. 10.1371/journal.pone.0150719.
- [6] *An introduction to T-Sne with python example.* KDnuggets. (n.d.). Retrieved September 22, 2021, from <https://www.kdnuggets.com/2018/08/introduction-t-sne-python.html>.