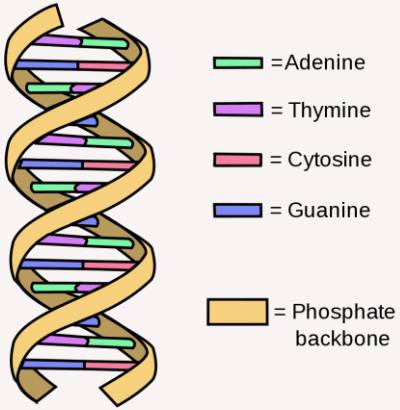
DNA Detectives

Scholars: Suraj, Clare, Vineet, Vyjayanti, Nithin
Instructor: Ayush Pandit



DNA

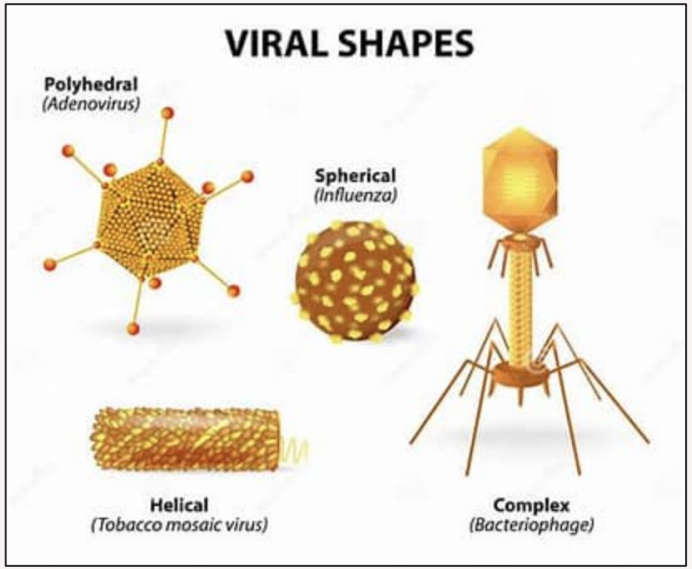
Background

Central Dogma:

- A theory stating that genetic information flows only in one direction

Viruses:

- Not living things
- Diverse shapes and sizes
- Viruses are constantly evolving and adapting to new hosts and environments



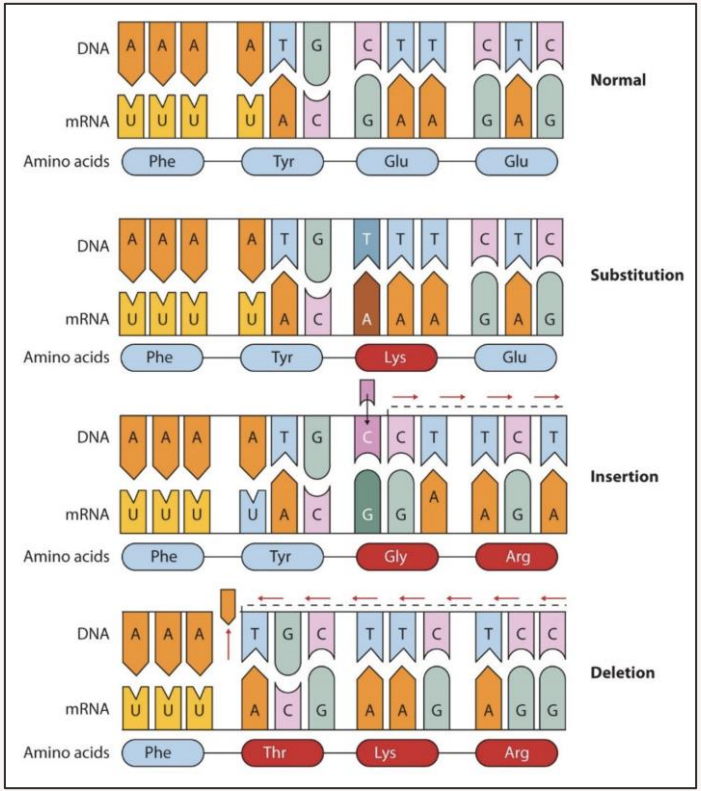
Background

Nucleotides:

- Form the basic structural unit of nucleic acids such as DNA
- Adenine (A), Guanine (G), Cytosine (C)
- Thymine (T) in DNA and Uracil (U) in RNA

Mutations and Viruses:

- Shows how viruses are related
- Allows the virus to evade the immune system



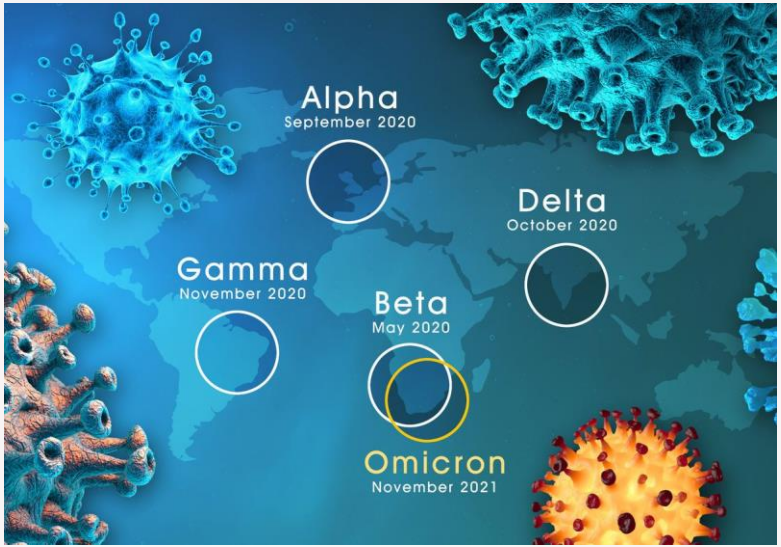
Problem Statement

Background:

- SARS-CoV-2 spreading and mutating
- Variants are more infectious

→ Epidemiologists finding ways to contain the spread

→ Genomic technologies



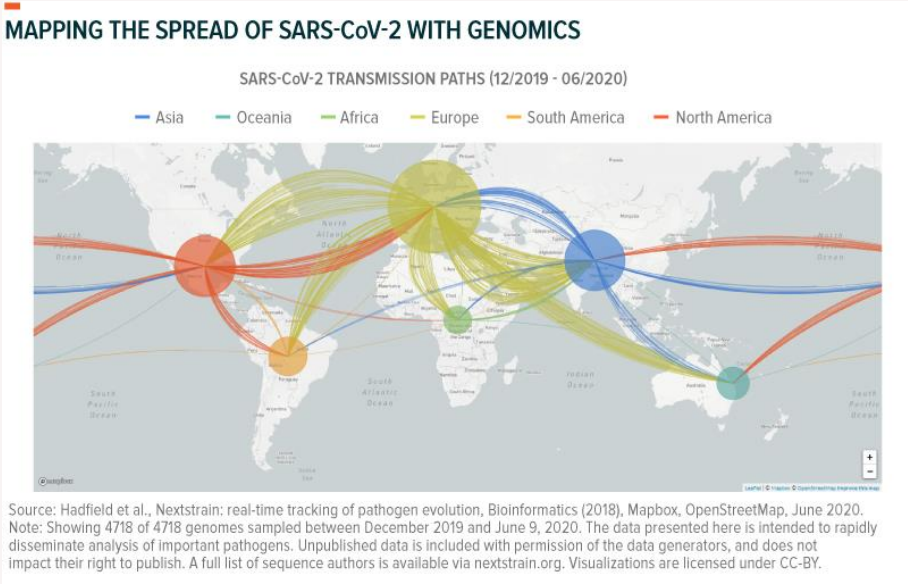
Problem Statement

What Does Our Model Do?

- Reads a set of SARS-CoV-2 genomes
- Predicts which region the strain is from

Allows scientists to:

- Identify where outbreaks may occur
- Contain the spread of the virus



Dataset

Overview:

- 1538 NCBI pre-aligned sequences
- Label → country of variant
- Libraries: Numpy, Pandas,

  Scikit Learn

Dataset - Sequence Alignment

Sequence Alignment:

- Based on origin sequence
- *Maximizing* number of aligned nucleotides
- Incorrect pairs → ‘indels’

Dataset - Input Processing:

Sequences as FASTA file

Store sequences in numpy array

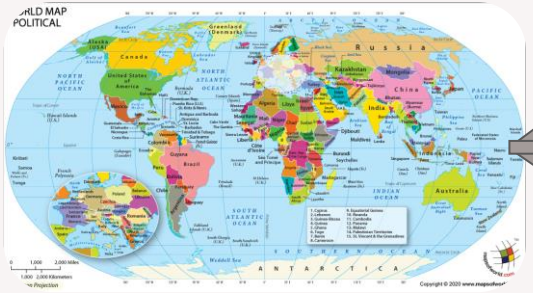
Position_Nucleotide

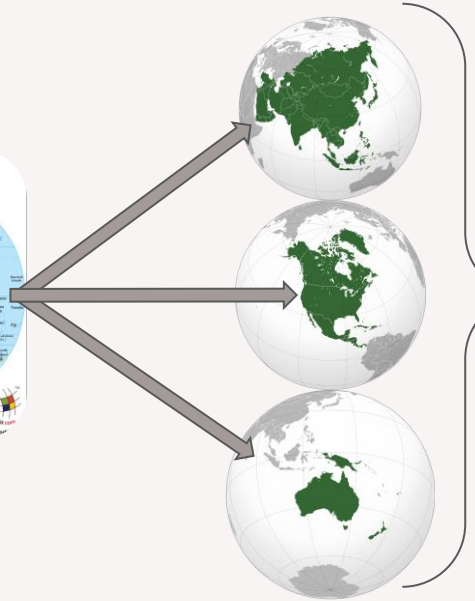
Sequence Number

	0_A	0_T	0_G	0_C	0_-	1_A	1_T	1_G	1_C	1_-
41	0	0	0	0	1	0	0	0	0	1
42	0	0	0	0	1	0	0	0	0	1
43	1	0	0	0	0	0	1	0	0	0
44	1	0	0	0	0	0	1	0	0	0
45	1	0	0	0	0	0	1	0	0	0

“One-hot encoding”

Dataset - Label Processing:





152
Sequences
per
Region

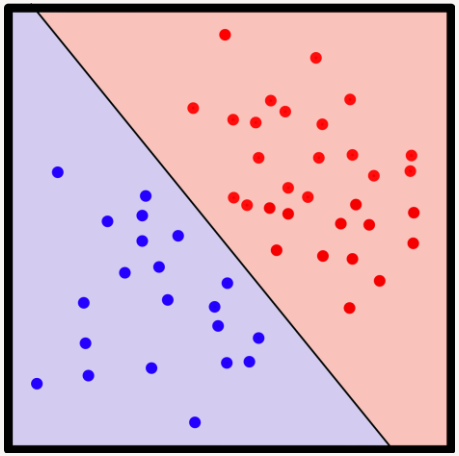
```
Number of samples in each region:
Counter({
    'North America': 152,
    'Asia': 152,
    'Oceania': 152
})
```

The Model

Multinomial Logistic Regression:

Downside:

1. Does not work as well with a smaller data set
2. Variables are presumed to have a linear relationship



Advantages:

1. Smaller SE
2. More efficient training

Final Accuracy = 96.6%

Conclusion

Limitations:

Limited Data

```
# Compute accuracy.
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy: %", accuracy)

# Compute confusion matrix.
confusion_mat = pd.DataFrame(confusion_matrix(y_test, y_pred))
confusion_mat.columns = [c + ' predicted' for c in lm.classes_]
confusion_mat.index = [c + ' true' for c in lm.classes_]

print(confusion_mat)
```

```
Accuracy: % 0.961038961038961
                Asia predicted  North America predicted  Oceania predicted
Asia true                   20                        0                  2
North America true           0                       25                  0
Oceania true                 1                        0                 29
```

Final Accuracy: ≈ 96%

Conclusion

Applications and Significance:

Identifying Regions

Understanding Spread of Virus

Identifying Future Variants