# Comparative genomics for COVID – similarity with other pathogens and also predicting likelihood of getting COVID

## Introduction

Under the genomics hackathon event, gave following 2 problem statements and worked on formulating the EDA as part of this analysis. Also tried to follow Agile methodology if am able to practice on own and follow up with the standards and processes.

Currently vaccines are testing and trials are done on volunteers (takes time but current panic has given lesser time to test). Social distancing, office/schools/colleges shutdown or working on remote are impacting the economy indirectly.

**Problem 1:** If there was a way to know the similarity and differentiation between similar pathogen infected genes, then similar treatment or medication could be applied to suppress the symptoms and buy some time.

**Problem 2:** If there was a way to know the likelihood for a person to be infected with COVID severely/mildly/no symptoms, then the right support could be provided to those susceptible to be infected more and those less likely could help in resuming work phase wise and help boost economy

## Userstory

**As an** "observer (Vaccinologist, Medical Practitioner)",

**I want** "to know the similarity and differentiation between similar pathogen infected genes"

**So that** "The similar trials for medication could be provided to overcome the symptoms of COVID“

**By using** “Comparison results given by the model for SARS/Ebola/Bats/Malaria infections”

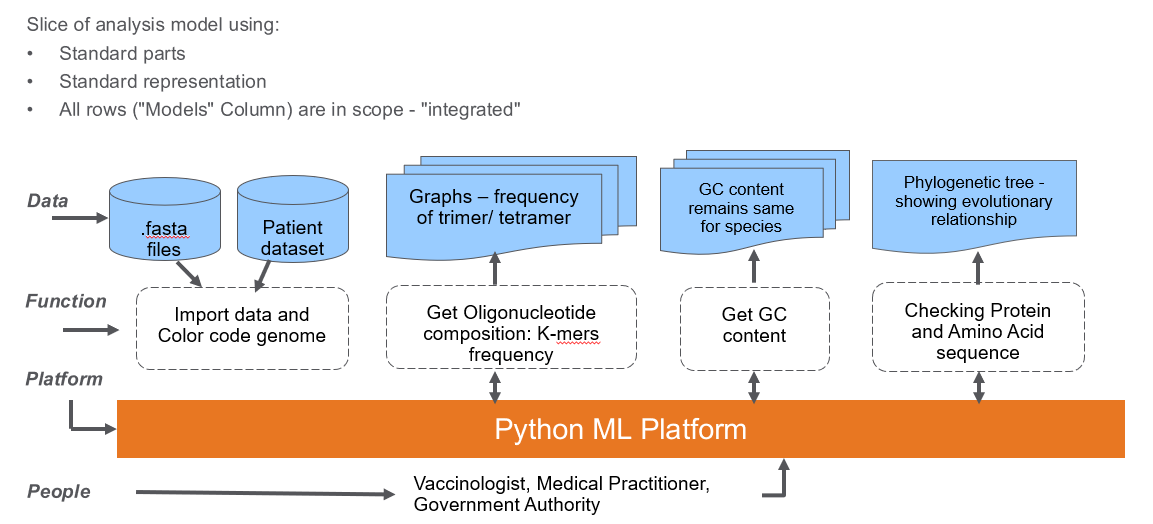
**As an** "observer (Govt Authority)",

**I want** "to know the likelihood for a person to be infected with COVID severely/mildly/no symptoms"

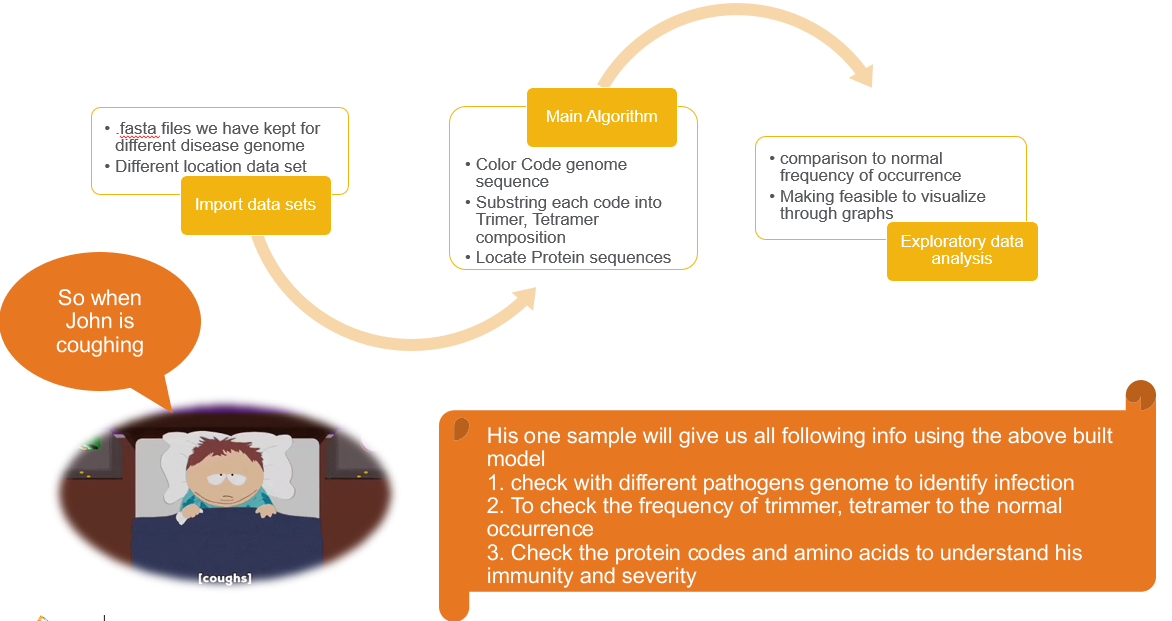
**So that** "The right support could be provided to those susceptible to be infected more and those less likely could help in resuming work phase wise and help boost economy“

**By Using** “Protein analysis for protein (ACE2 Angiotensin converitng enzyme2) in the genes)

## Architechtural Model:



## Steps flow:



## Model Outputs and Benefits:

**Model outputs:**

* + Clustering all COVID-19 patients’ genomes
  + Based on the oligonucleotide compositions, all the genomes from coronavirus family show similar composition (trinucleotide and tetranucleotide composition)
  + Based on GC content, the composition of infected Corona virus is closely similar to COVID-19 genome's GC content.
  + Judging by the number of protein strands, the count in COVID19's genome is very close to those found in infected Bat.

**This analysis can be of used for following:**

* + Phase wise allowing people to resume work - who have good immunity results by the genome study
  + More likely susceptible can be informed under the advisory to be taking precautions of social distancing
  + Same analysis could be used in later times also for further outbreaks of pandemics
  + If all the eatable animals’ genomes could also be studied this data could also help in predicting impacts on seasonal basis, so Government can provide an advisory timely to stop eating those.