Problem: [REDACTED] a public-private biotechnology company, uses historical and medical data, analytics and artificial intelligence to identify which virus families are most likely to cause major outbreaks of new viruses and develop new antiviral drugs before they are needed. When a new virus infects humans, our immune systems are powerless to stop it from replicating. Viruses in the same family share inherited methods for hijacking cells to essentially become virus factories (host-cell machinery essential for virus infection and replication). [REDACTED] scientists identify those common factors within virus families and exploit them as a viral Achilles heel.

1. Identifying at risk viral families based on prior patterns of past pandemic viruses. Most of the problem is identifying what data to use in order to make a prediction that can be empirically supported, and I will go into this further in the rationale section.
    a. Given:
        i. Ratio of year to year occurrences based on population (5 years prior to each viral pandemic, if applicable), viral mode of transmission, method for hijacking of host-cell machinery, attachment protein of virus, cell receptor to which the protein binds, binary value of if disease resulted in a pandemic or not (1 yes/ 0 no). We will also need the same data for every virus in the last 5 years (excluding viruses that have already resulted in a pandemic) in order to classify them as high-risk or not.[6]
    b. Use:
        i. Logistic Regression
    c. To:
        i. Obtain a response value per virus that will detail 1 if the virus is high-risk, and 0 if not.

Rationale: In this problem we are going to be analyzing past records of pandemics and determining if there is any pattern or correlation that can be used to identify high-risk pandemic-creating viral families. By doing this, it is possible to predict which viruses are threats, and thus create a countermeasure before a potential pandemic event occurs.[4] This contributes to [REDACTED]'s overall goal of preparedness for a pandemic by identifying which virus families

are most likely to cause major outbreaks of new viruses. We will need transmission data (ratio of year to year occurrences per population, 5 years prior to each viral pandemic after 1900, if applicable). The year 1900 was chosen so that we can ensure there is suitable data for tracking purposes. We are also excluding the years it was considered a pandemic because we are looking for patterns before a virus becomes a full blown pandemic. In addition if no data is present we will use the closest virus in the family for which we have data (ex. COVID-19 is of the family coronavirus, which has been around for a very long time). We will also need the mode of transmission for each virus (sexually, airborne, faeco-oral, etc.), method for hijacking of host-cell machinery (endocytic or non-endocytic), attachment protein of virus (what the virus uses to initiate the viral life cycle), and cell receptor to which the protein binds.[5] We will also need the same data for every virus in the last 5 years (excluding viruses that have already resulted in a pandemic) in order to classify them as high-risk or not. Finally, a binary response variable that dictates if a virus resulted in a pandemic or not. Logistic regression will then be used to classify each of the viruses as at risk for pandemic or not. Furthermore, a threshold value will be set conservatively so that only the most at-risk viruses can be categorized as pandemic causing. This is due to how expensive research and development is to combat viral infections, so we want to make sure that the budget is spent in the right places.

2. Minimize error in the logistic regression model to be certain that the resources given are spent adequately.
    a. Given:
        i. Aforementioned logistic regression model
    b. Use:
        i. Gradient Descent
    c. To:
        i. Minimize error of cost function, and thus ensure the model can more accurately predict potential pandemic threats.

Rationale: As mentioned above, the cost of research and development into developing an antiviral medicine is extremely costly. Because of this, we must ensure that when we are developing something that can be of use. As a result, I determined that gradient descent should be used to minimize error of the logistic regression model.[2] Similarly to the rationale for the logistic regression above, this step of the process aligns with [REDACTED]'s goals of identifying which virus families are most likely to cause major outbreaks of new viruses. This portion of the project is simply an extension of the previous step to ensure that the company would be heading in the right direction.

3. Identify common structures in the viral families selected that can be exploited as a viral Achilles heel.
    a. Given:
        i. Genomic Sequences of the attachment protein based on results from earlier logistic regression analysis.[3]
    b. Use:
        i. A Clustering algorithm (DBSCAN, OPTICS).
    c. To:
        i. Find a grouping of similar attachment protein sequences to target for drug development.

Rationale: In this problem, we are dealing with the first stages of drug development, identifying a target. We will take all of the high risk viruses from the selected viral families above and determine the similarity to target a common factor among them. Luckily viral genomes are not very large, so analysis of them can be done in a relatively quick time (compared to that of a human genome, for example).[7] I propose a clustering algorithm based on Hamming's distance of attachment proteins in the viral genome. Having done work in the computational biology field dealing with a similar problem (clustering unannotated genomes to determine their function), I know that this procedure has a scientific basis and will work well to classify the proteins. We will be using the clustering algorithms of DBSCAN, and OPTICS because of the way they cluster the figures in the following link: https://scikit-learn.org/1.5/modules/clustering.html (Note: Spectral also clusters in a similar way but it requires to set the number of clusters beforehand, which is something the other two algorithms do not require). The reason I decided to choose to target the attachment protein is because of the recent studies done on combating COVID-19. The method that gained popularity worldwide targeted the spike protein (i.e attachment protein) that allowed the virus to attach and begin its life cycle inside of the cell. By creating a harmless spike protein in the vaccines, the human body was able to create its own antibodies to counteract the actual spike proteins of COVID-19.[1] Based on the effectiveness of this protocol, I believe this to be the best solution for the problem at hand. To align with [REDACTED]'s goals, by determining which viruses in each family have this protein as a common factor, it is possible to exploit them as a viral Achilles heel.

Discussion: Overall, I believe this method could give an unbiased start to how [REDACTED] could begin their venture towards using analytics to develop a broad-spectrum antiviral. Although there can be some computational methods used in the field, these are outside the scope of this class and thus were not considered when creating this analytics model. Furthermore, given the fact that the logistic regression utilizes gradient descent to minimize its error, I would not think that any of these models would need to be run more than once to find a decent output. In terms of where the data could be found, the government has entire databases hosted through the National Institute of Health, Center for Disease Control, and the National Archives and Records Administration that can provide the data outlined in this document.

References:

1. https://www.cdc.gov/covid/vaccines/how-they-work.html
2. https://web.stanford.edu/~jurafsky/slp3/5.pdf
3. https://pmc.ncbi.nlm.nih.gov/articles/PMC6743247/
4. https://www.jci.org/articles/view/170236
5. https://pmc.ncbi.nlm.nih.gov/articles/PMC9318756/
6. https://pmc.ncbi.nlm.nih.gov/articles/PMC8492370/
7. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1602-3