# EXL EXCELLENCE QUOTIENT 2021

TEAM- THE GIMMICKS, BIRLA INSTITUTE OF TECHNOLOGY, MESRA

# Understanding Problem Statement

- **Problem statement**: The problem statement by EXL Excellence Quotient 2021 required us to develop a tool for a US public health client to predict daily Covid-19 cases at a county level.

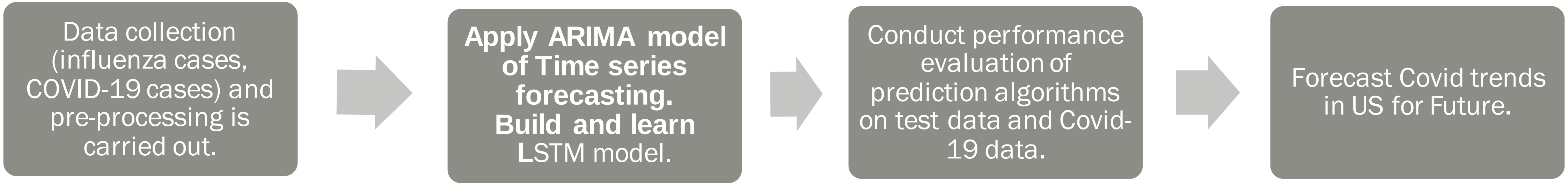
- **Establishing its need for a solution:** As COVID-19 outbreak evolves, accurate forecasting plays an extremely important role in informing policy decisions. Forecasts can help us to contribute to the distribution of medical supplies in the worst effected areas across the country by determining the medical supply need for individual hospitals. We hope that our forecasts and data repository can help improve hospitality and vaccination strategy, guide necessary county-specific decision making and help countries prepare for their continued fight against COVID-19.

- **The desired outcomes and impact on society:** Analysis of the data, predicts the evolution trend of the existing pandemic data. The predicted forecast will enable an **investigation into the connection between socioeconomic and demographic** information with health resource data (number of ICU beds, medical staff ) that will help in gaining a focused understanding of the severity of the pandemic thus, provide guidance about its prevention and control at the county level. This will overcome one of the greatest challenges of equitable and efficient **allocation of scarce resources** and also predict future medical and **public health workforce** requirements according to the Identified geographical areas in which staff shortages are most likely to develop.

- **Approach for the solution:** The data given with information updated to January 2021 is a **time series** as it was collected over a period of time and over a series of constant, regular intervals which are the key characteristics. We have used deep learning technique to apply **ARIMA model of Time series forecasting**. It is best suited over all other methods as it predicts a given time series based on its own past values, utilizing grid search it recognizes a lot of boundaries that delivers the best-fit model for our time series data. It can be used for any nonseasonal series of numbers that exhibits patterns and is not a series of random events.

- **Constraints:** The dataset with which we were provided with contains many NaN values. We dropped these rows which resulted in the loss of data. During the training we ignored this point

# Solution Design



Data collection (influenza cases, COVID-19 cases) and pre-processing is carried out.

**Apply ARIMA model of Time series forecasting. Build and learn** LSTM model.

Conduct performance evaluation of prediction algorithms on test data and Covid-19 data.

Forecast Covid trends in US for Future.

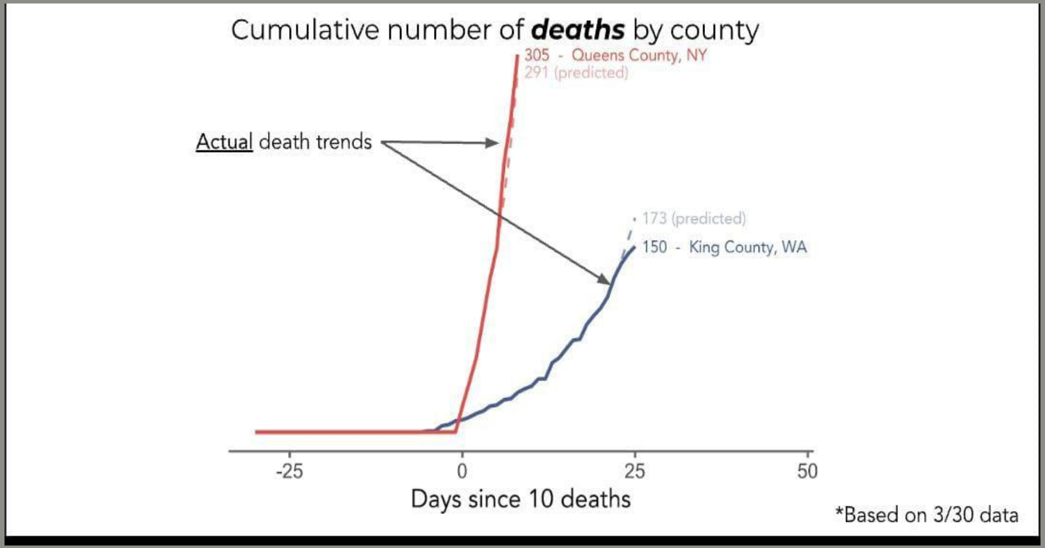
# Forecasting/Modelling Tool Development Process

- The provided dataset with information updated to January 2021 to predict daily Covid-19 cases, is a **Time series**. Time-series forecasting is widely used for such dataset as models are capable to predict future values based on previously observed values.
- Taking the stationarity test by **Augmented Dickey-Fuller test**, the time series exhibited trending behaviour or **non-stationarity** in the mean. This step is very important because the whole results of the regression might be fabricated.
- **Long short-term memory (LSTM)** is RNN architecture, used in the field of deep learning. While using an LSTM model we are free and able to decide what information will be stored and what discarded using ‘gates’. These models are able to store information over a period of time which is extremely useful when we deal with Time-Series Data. Hence, we have built a multi-layer LSTM recurrent neural network to predict the predict daily Covid-19 cases from a sequence of values. The modules used are : Keras, TensorFlow, Pandas, Scikit-Learn & NumPy.
- To boost the performance, the next step in the process was to **normalize the data** before model fitting. We did this by implementing **Min-Max scaling** in Python using scikit-learn and then split the data into training and test sets to avoid overfitting and to be able to investigate the generalization ability of our model.
- We next built a neural network architecture, a **RNN-LSTM architecture** was trained on our training data. After the model was trained, we assigned 1 neuron in the output layer for predicting the the number of confirmed cases on county level. To achieve good results fast, we have used **MSE loss function and the Adam stochastic gradient descent optimizer** as it overcomes redundancy by performing a parameter update for *each* training by one update at a time.

# Validation Results of Forecasting/Modeling Tool

**Performance Evaluation Criteria – Mean Square Root Error**

We have used Root Mean Square Error (RMSE) as a standard statistical metric to measure model performance by calculating difference between the value predicted by the model and the value observed in the actual environment.



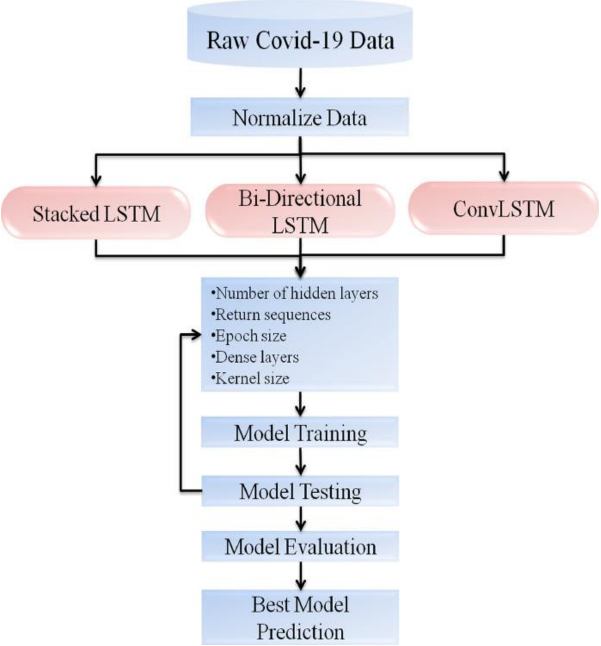
*Based on 3/30 data

**Validation Results**: The mean squared error (MSE) is close to zero.

# Top drivers of the model

**ARIMA model**: We have used ARIMA (Auto-Regressive Integrated Moving Average), a class of models that has the principal objective to correctly recognize the stochastic mechanism of the time series i.e., its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

**RNN Modules**: We developed an ensemble framework that combines multiple RNN-based deep learning models using LSTM algorithms that show high performance in time series forecasts and are capable to learn long term dependencies by replacing the hidden layers of RNN with memory cells.



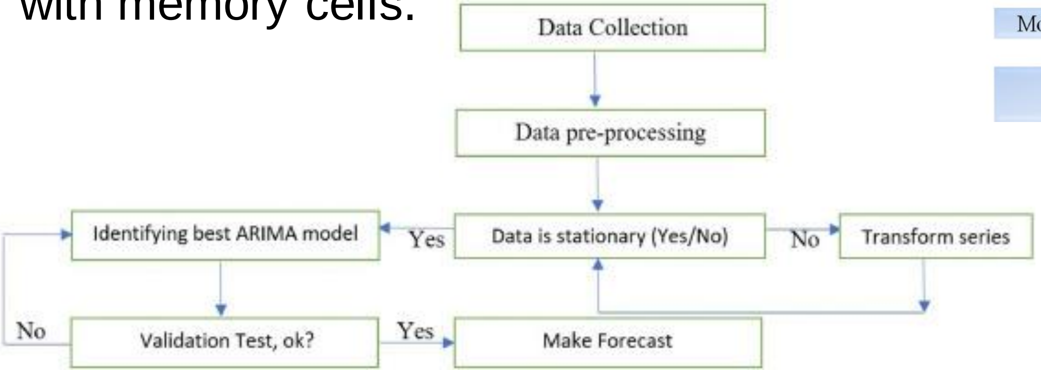


*Fig. Methodology to apply the ARIMA model and LSTM algorithm for forecasting.*

# Additional Analysis

## 1. To Identify top 5 States expected to be worst affected by COVID-19 pandemic by 30th April 2021 considering the population demographics.

- **Data Source :** We have used data of COVID-19 cases in the United States as of March 11, 2021, by state available publicly at https://www.statista.com/statistics/1102807/coronavirus-covid19-cases-number-us-americans-by-stat
- **Statistical models used :** We consider the **exponential model, the logistic model, and the Susceptible Infectious Susceptible (SIS) model** for COVID-19 pandemic prediction at the state level. These models have already been used to predict epidemics like COVID-19 around the world, including China, Ebola outbreak in 2014

  *The exponential model* will estimate the upper bound of the total number of infected people by 30th April.

  *The logistic model*-based prediction will capture the effect of preventive measures that have already been taken by the respective state government.

  *The SIS model* will reflect the effect of the major preventive measure like recent crowd gatherings for protests.
- **State-wise Analysis and Prediction Report:** Dependent on inputs from the exponential, logistic, and SIS model along with daily infection-rates for each state, we interpret the results from different models jointly to conclude :

  **The top 5 states worst affected by April,30 are:**

  **1. California: As of today California has 4,740 confirmed cases on average in one week. The growth rate is -5%.**

  **2. Texas: As of today Texas has 6,190 confirmed cases on average in this week. The growth rate is -5%.**

  **3. Florida: As of today Florida has 5,136 confirmed cases on average in this week. The growth rate is -5%.**

  **4. New-York: As of today New-York has 7,279 confirmed cases on average in this week. The growth rate is +5%.**

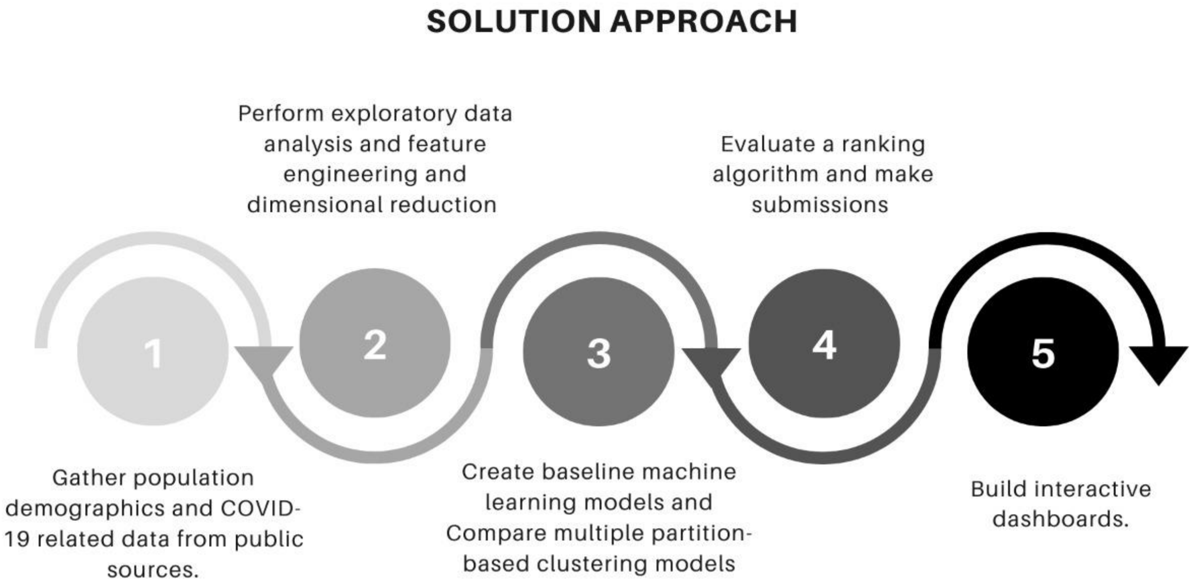
  **5. Illinois: As of today Illionis has 6,190 confirmed cases on average in this week. The growth rate is -5%.**

- **Discussion:** The number of confirmed cases alone is not a good metric for estimating risk of exposure because it doesn't take population density into account whereas **Incidence Rate** controls for population density to give number of confirmed cases per **100,000 people.**

  Drawbacks for predicting the trend of COVID-19 disease and analysing state's conditions to explain the situation, there is lack of information about social and political measurements and reactions to illness spread, state categorized data is not available for the public completely, many algorithms do not work in this restricted dataset due to its time-series nature.

# Vaccination Strategy to optimize overall cost of transport/distribution to minimize the number of new cases

Aim of the strategy is to save lives and reduce the chain of transmission **with AI enabled prioritization** for a fast and effective vaccine delivery. The solution prioritizes the population, considering the limited availability of vaccines initially and at the same time, limiting the spread of COVID-19.



SOLUTION APPROACH

1. Gather population demographics and COVID-19 related data from public sources.
2. Perform exploratory data analysis and feature engineering and dimensional reduction
3. Create baseline machine learning models and Compare multiple partition-based clustering models
4. Evaluate a ranking algorithm and make submissions
5. Build interactive dashboards.

## Expected Results

Prioritization of vaccine delivery for human wellbeing will help various sectors of community like:

► Health workers at high or very high risk

► Employment categories unable to physically distance

► Groups living in dense urban neighbourhoods

► Groups living in multigenerational households

It will help to reduce societal and economic disruption:

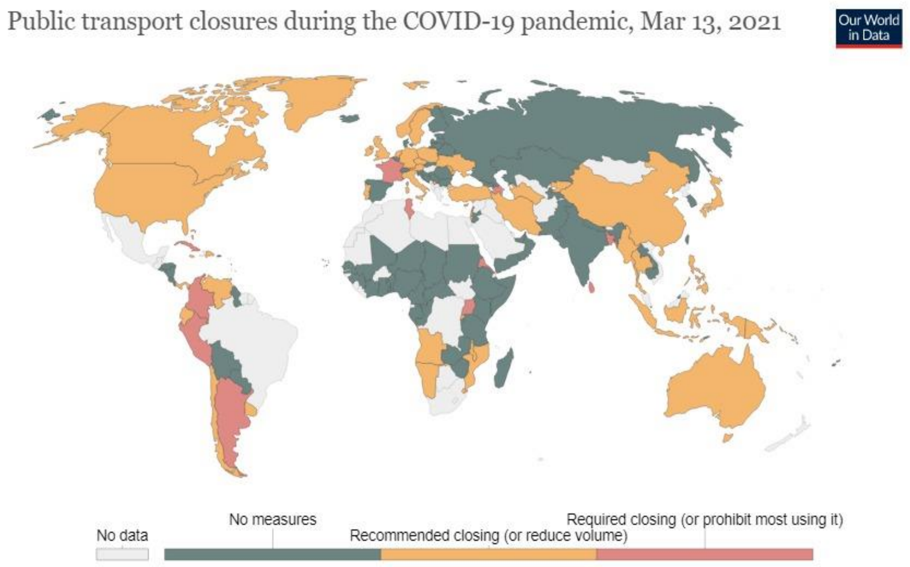
► School-aged children to minimize disruption of education and socioemotional development

► Workers in non-essential but economically critical sectors, particularly in occupations that do not permit remote work or physical distancing

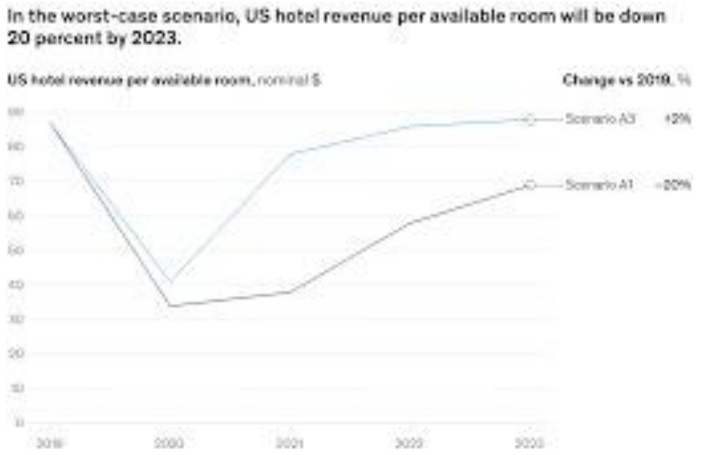
# Additional variables that could have be used or explored to improve the model prediction accuracy

- Bus operators in the UK and Europe have announced a halt to services, while other operators worldwide are running reduced services as a result of the latest wave of COVID-19.
- Many of the London bus drivers who died from coronavirus were suffering with underlying health conditions particularly high blood pressure.
- Since, the population diversity is related to the number of confirmed cases. For eg: Europeans and American are more prone to COVID-19 than Asians. The coronavirus (COVID-19) pandemic has revealed deep-seated inequities in health care for communities of color and amplifies social and economic factors that contribute to poor health outcomes. Recent news reports indicate that the pandemic disproportionately impacts communities of color, compounding longstanding racial disparities.
- FirstGroup, the UK's largest bus company and operator of rail franchises, has suggested it may not be able to continue trading because of **the impact of COVID-19 on its passenger numbers**. Relaxing the social distance rule from two meters to one meter allows an underground train to accommodate more people, it was still limited to 208 people, said Ron Kalifa, board member of Transport for London.

Through the above facts and reports ,it's evident that **transport and mobility data** and population diversity i**s very much correlated to the number of COVID-19 confirmed cases**. Therefore, transport and mobility data can be used to make our model perform better.


Public transport closures during the COVID-19 pandemic, Mar 13, 2021

Our World in Data
No data
No measures
Recommended closing (or reduce volume)
Required closing (or prohibit most using it)


In the worst-case scenario, US hotel revenue per available room will be down 20 percent by 2023.