# Project: Predicting Epidemics

## 1. Domain Background

Accurate and precise epidemic data is scarce and that which exists is not fully reliable — yet epidemics are a significant issue, if left unchecked they can grow with catastrophic consequences. It would be phenomenal if we could apply deep learning to epidemic prediction, such that a model could be created that was capable of predicting the magnitude of an epidemic, in order to mitigate them in time and hopefully reduce the damage they would cause otherwise. The hypothesis is formed since an epidemic is hardly ever "sudden" though they appear to be so, i.e. an epidemic is present for some time before it's recognized as it's early growth is rather slow. Thus there is a time it is latent in the population before the exponential increase, the outbreak, in infected is observed – and that is the time when it is most crucial to detect the epidemic. But it is inherently difficult to detect the early stages of an epidemic as the changes implied are small. It follows that if these minute and delicate shifts within the initial days of an epidemic can be modeled, then it would be possible that the magnitude of an epidemic could be determined with reasonable accuracy and the future behavior of an epidemic could be modeled, and thus predicted. And predicting the magnitude and behavior of an epidemic before it explodes could very well have significant consequences in mitigating its damage and panic.

## 2. Problem & Solution statement

Problem Statement**:** The goal of this project is to try to predict the nature of an epidemic from it's early onset within a shortest possible time-span so that it's future behavior can be predicted through modeling, and thus the epidemics deleterious effects can be handled appropriately.

Solution Statement: The solution to predicting the nature of an epidemic is to build a recurrent neural network that takes as input features that are *easily observable and trackable* during an epidemic, i.e. people sick, people dead – they are described below. Then we can use the RNN to regress the three key rates that determine the behavior of an epidemic on a population, and hence predict its magnitude. If the model can predict with a reasonable proximity the three key rates based off a short sequence of data of daily observable features, then the predicted three key rates can be used to mathematically model the epidemic and predict it's future behavior.

To clarify:

- T*he nature or magnitude of an epidemic* consists of how the values of the three key rates ultimately determine how an epidemic behaves on a population. It is not an objective nor explicit like Note: these are explained in depth in section **3. Dataset and Inputs.**
- The *shortest possible time-span* describes the aim of creating a model such that it gives reasonable predictions not far from the truth based from the fewest data of sequential step possible, in this case the fewest number of days. One must remember that when it comes to infectious diseases with an exponential growth, such as epidemics, time is of the essence. And a week or a day earlier, can make a significant difference.

- *Predicting the future behavior* means that we can use the three key rates of an epidemic $\beta$, the recovery rate $\propto$, and the death rate $\delta$ to mathematically model the epidemic over time and thus predict it's likely behavior.

3. **Dataset & Inputs**

As said before the precise and accurate epidemic data is scarce, but thankfully mathematicians have studied extensively the mathematical behavior of epidemics and from that have created mathematical models. We will use one of these mathematical models, the SIR model developed by Kermack and McKendrick, to create our dataset.

The implementation of the SIR model will be as follows:

$$N = S_t + I_t + R_t + D_t$$

Where:

- $N$ : represents total initial population (constant)
- $S_t$ : number of people susceptible at time step $t$
- $I_t$ : number of people infected at time step $t$
- $R_t$ : number of people recovered at time step $t$
- $D_t$ : number of people deceased at time step $t$

At every time step, the population dynamics changes as described by these partial differential equations:

- $\partial S/\partial t = -\beta SI/N$
- $\partial I/\partial t = \beta SI/N - \partial R/\partial t - \partial D/\partial t$
- $\partial R/\partial t = \alpha I$
- $\partial D/\partial t = \delta I$
- $\partial N/\partial t = 0$ **no change in total initial population*

Where at each time step $t$:

- $\beta$ : Infection Rate, in terms of probability of being infected if susceptible given proportion of infected
- $\propto$ : Rate of recovery, in terms of probability of recovering if infected at each time step $t$
- $\delta$ : Rate of mortality, in terms of probability of dying if infected at each time step $t$

From the SIR model and its partial differential equations, we will *generate* our dataset. We will do so by running numerous simulations of epidemics (projected 10,000 simulations), in each simulation randomly varying the three key rates of an epidemic with realistic ranges. The independent and random variation of the three key rates within realistic ranges leads to a range of parameter combinations that further lead to a dataset of within all magnitudes of an epidemic

The three key rates are : the infection rate $\beta$, the recovery rate $\propto$, and the death rate $\delta$. Their realistic ranges where the key they will be randomly generated will be as follows:

- Infection Rate $\beta$ range: [0.15 – 0.7]
    - Example: $\beta = 0.3$ can be interpreted as there being a 30% probability that a susceptible person is likely to be infected after contact with an infected person given the proportion of infected in the total population.
- Recovery Rate $\propto$ range: [0.1 – 0.6]
    - Example: $\propto = 0.4$ can be interpreted as there being a 40% probability that a person recover each time step given that the individual is currently infected.
- Mortality Rate $\delta$ range: [0.01 – 0.05]
    - Example $\delta = 0.03$ can be interpreted as there being a 3% probability that a person succumbs to their disease at each time step given that the individual is currently infected.

We have chosen features such that can be easily and realistically be observed during the onset of an epidemic and that as well can be feasibly and realistically tracked throughout time:

1. 'IS_ratio' - ratio of number infected over number susceptible in 1000's
2. 'RI_ratio' - ratio of number recovered over number infected
3. 'ID_ratio' – ratio of number infected over number deceased
4. 'RS_ratio' – ratio of number recovered over number susceptible 1000's
5. 'DS_ratio' - ratio of number deceased over number susceptible 1000's
6. 'RD_ratio' - ratio of number recovered over number deceased
7. 'num_inf' - number of people that are currently infected
8. 'num_scspt' – number of people that are susceptible
9. 'num_recov' – number of people that are recovered
10. 'num_dcs' – number of people that have deceased
11. 'delta_inf' - daily change in number infected
12. 'delta_scspt' - daily change in number susceptible
13. 'delta_recov' - daily change in number recovered
14. 'delta_dcs' – daily change in number deceased

For each epidemic simulation that is randomly selected we will save the three key rates that were randomly generated to simulate the epidemic and use those as the targets of the recurrent neural network. These three key rate values will each correspond to a matrix containing a feature in each column and its value at each time step in the corresponding row -- which will be the sequential inputs of our recurrent neural network.

It should be noted that simulations are known to be rather "perfect" and thus unrealistic. That is, the behavior of a mathematical model is hardly ever a clean, smooth partial differential equation. Thus we will incorporate a noise factor to our model (to be determined), such that the underlying model is the same, but the feature values vary slightly randomly at each iteration that helps make the model be more akin to reality, which is inherently chaotic.

## 4. **Evaluation Metrics & Benchmark Model**

The model will regress of three variables: the infection rate $\beta$, the recovery rate $\propto$, and mortality rate $\delta$. Given that this is a regression problem, it is hypothesized that a mean squared error or root mean square error are the best metrics with which to train the model.

While we will train the model with the MSE and RMSE metrics for optimizing weights, it is not ideal to use these same metrics to *evaluate* our model's performance. Given that our three regressors take values between 0 and 1 i.e. $0 < \beta, \propto, \delta < 1$ it would be deceptive to present the squared error between the prediction and the ground truth values as the evaluation metric since the square of a decimal is significantly smaller than itself i.e. 0.1 ^ 2 = 0.01. As well the mean absolute error would also be deceiving as the nature of the problem is to regress extremely low values, all less than one, thus all errors present in absolute terms will be relatively small even though a small difference may be hugely significant.

Because of this, instead we will evaluate the results of our model off the mean absolute percentage error *MAPE*. As such the error of our model's predicted result will be presented as percentage difference from the ground truth value, this seems most appropriate and least deceiving way to present the error difference for the decimal values of our regressors.

The benchmark model will be a naïve model that always predicts the average of the three variables. The benchmark model's skill will as well be presented in terms of the MAPE between the predicted and the ground truth values of the three key rates (regressors).

## 5. **Project Design & Flow**

Process 1: Generate the synthetic data by running simulations of the epidemic model randomly selecting the three key rates
$\beta, \propto$, and $\delta$ bounded by their respective realistic ranges that were previously specified. The other parameters such as the initial infected, $I_0$, will be randomly generated between 5 –15 and total population, $N_0$, will be instantiated at a set 100,000. The initial number susceptible, $S_0$, will be calculated by the difference between total population, $N_0$, and initial infected, $I_0$. After instantiating each epidemic simulation, we record at each time step and save the values of the intended input features at each time step (with a degree of noise incorporated for realism) and as well save the three key rates corresponding to the particular simulation. It is intended to run 10,000 of these simulations or more such that our dataset is complete with a variety of different combinations of all three key rates.

Process 2: Catalog the exploratory analysis and preprocessing of the data to create the analytical base table, ABT, for subsequent model training. Train different recurrent neural networks experimenting with different architectures, layers and hyper parameters to find the model with the best skill..

Process 3: Evaluate and interpret the best model's skill, i.e. how well the model can predict an epidemic. Evaluate the model's weaknesses and strengths, what ranges in each of the regressors it has highest and lowest levels of confidence. Synthesize the models result and create visualizations such that the model's skill and results for different and better understanding of the model's ability. Conclude with the suggestions of further work.