Prediction of COVID-19 Deaths using an LSTM Neural Network

# ESC 403 | Introduction to Data Science

**Project proposal | 01 April 2022**

Contributed by

Nikolai Horozov [nikolai.horozov@uzh.ch](mailto:nikolai.horozov@uzh.ch)

Jerome Sepin [jerome.sepin@uzh.ch](mailto:jerome.sepin@uzh.ch)

Vithersan Somasundaram [vithersan.somasundaram@uzh.ch](mailto:vithersan.somasundaram@uzh.ch)

Gabriel Marie Falconer-Stout [gabrielmarie.falconer-stout@uzh.ch](mailto:gabrielmarie.falconer-stout@uzh.ch),

# General topic

The aim of the project is to use a Long Short-Term Memory (LSTM) model to predict future COVID-19 deaths. The model is trained using data from early COVID-19 deaths from around the world. The model will use predictor variables related to the pandemic such as vaccination implementation, closures, and other restrictions. Of particular interest, will be non-COVID-19 related variables; the location of where COVID-19 has occurred will be explored. This will be done via latitude and longitude coordinates to assess the impact of these variables on COVID-19 deaths. The model’s prediction ability will be applied to another validation data set to assess whether such a model would be useful to predict the course of the pandemic.

# Motivation

COVID-19 first appeared in December 2019. Since then, it has caused millions of infections and deaths. With this massive number of casualties, it has become one of the biggest crises in the world.

In addition to the loss of human lives, this pandemic has also caused severe damage to the global economy. Due to lockdowns and distancing strategies, it has also had a negative impact on mental health and our society. Therefore, creating a prediction model is of crucial importance. Knowing the course of the pandemic would allow for better measures to be taken and to limit deaths as well as social disturbance.

# Data

The data with the COVID-19 cases comes from an open dataset from COVID-19 Data Hub. It is compiled from various sources, crunched hourly. Among other things, the data set contains:

* standard COVID-19 variables such as:   
  total population, cumulative number of cases, tests, deaths, recovered, daily number of cases
* geographic coordinates
* external identifiers
* google and apple mobility reports

**Source**:  
*Guidotti, E., Ardia, D., (2020), "COVID-19 Data Hub", Journal of Open Source Software 5(51):2376, doi:*[*10.21105/joss.02376*](https://doi.org/10.21105/joss.02376)*.*

# Data processing

* **Data Cleaning**

The data from COVID-19 Data Hub packages are used to calculate the variables necessary for the analysis. These are:

* Daily cases
* Proportion of daily cases per population
* Proportion per 100K
* Daily deaths
* Latitude and longitude of locations
* **Exploratory Data Analysis (EDA)**
  + Sorting, summarization
  + Visualization
  + Plotting:
    - Stratified daily deaths
    - Important covariates
* **Conducting neural network**
  + LSTM
  + Train (70%) and Test (30%)
  + Validation set – To compare our forecast
  + **Creating features and labels:**
    - specific time length: optimal time lag needs to be determined

# Research Question

Are different locations affected by COVID-19 differently and can region-specific predictors be used to predict the number of COVID-19 deaths using LSTM neural network?

# Analysis Techniques / algorithms

Using Time Series Deep Learning to forecast COVID-19 deaths with Keras Stateful LSTM to answer the primary research question.