Prediction of COVID-19 cases using LSTM neural network

# ESC 403 | Introduction to Data Science

**Project plan | 01 April 2022**

Contributed by

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

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

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

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

# General topic

The aim is to use Long Short-Term Memory (LSTM) Models to forecast new COVID cases. The model is trained using data from early COVID cases from various regions worldwide except Switzerland. The model will use predictor variables that are not related to the pandemic such as, but not limited to, humidity, temperature, and population density to assess the impact of these variables. Furthermore the model's prediction ability is applied to Switzerland to assess whether such a model would be useful to predict the pandemic course.

# Motivation

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

# 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 (total population, cumulative number of cases, tests, deaths, recovered, daily number of cases, . . . )
* geographic information
* external identifiers
* google and apple mobility reports

*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)*.*

**climate** R package: meteorological and hydrological data from publicity available:

<https://cran.r-project.org/web/packages/climate/vignettes/getstarted.html>

# Data processing

* **Data Cleaning and data tidy**

These data from COVID-Packages is used to calculate the variables necessary for the analysis. These are Daily cases, Proportion of daily cases per population, Proportion per 100K and daily deaths.

* **Exploratory Data Analysis (EDA)**
  + Visualization
  + Validation
  + Sorting, summarization
  + Plotting: Daily cases per country , Temperature, humidity etc.
* **Splitting data in to training (70%) and test set (30%)**
  + Validation set: to compare our forecast
* **Creating features and labels**

specific time length 🡪 optimal time lag needs to be determined

* **Scaling/Normalization**

# Research Question

**Primary research question:**   
How accurately is the prediction of COVID19 deaths in the Switzerland case

**Secondary research question:**

Visual timeline of cases, deaths in relation to demographic variables

# Analysis Techniques / algorithms

Using Time Series Deep Learning to Forecasting Covid deaths with Keras Stateful LSTM to answer the primary research question. The second research question will be answered with help of a Visualization.

* PCA (Principal component analysis) ?