# A Deep Learning framework to model Influenza/Flu predictions using aggregated Google Search query data
# Mid-Progress Report
# Team : Kingpins

Sagar Satyanarayana
Mohammad Turab Ali
Garima Silewar
Nikhil Kumar Mutyala

November 7 2019

# Contents

# 1 Introduction

Influenza outbreaks cause up to 500,000 deaths a year worldwide, and an estimated 3,000–50,000 deaths a year in the United States [1]. The ability to effectively prepare for and respond to outbreaks heavily relies on the availability of accurate real-time estimates and the existing methods remains limited [2, 3]. Traditional flu surveillance systems, such as Center for Disease Control and Prevention's (CDC) influenza reports lag behind real-time by one to two weeks, whereas information contained in internet users' search activity is available in near real-time [4].

This project proposes and implements a framework for epidemiological predictions (such as influenza/flu prediction) using internet users' search activity.

# 2 Problem Description

Building a real time framework using online activity data (like Google Search Trends) using Deep Learning and traditional forecasting methods and comparing the results with Google's prediction of flu trends (GFT) in the United States.

# 3 Literature review

## 3.1 Need for internet user's search activity trends in epidemiological predictions

Traditional methods of data collection in epidemiological studies need heavy resources in terms of logistics, time, as well as human and material resources, so leading the way to searching alternative strategies for collecting data [5]. Since internet has increasingly become a meaningful health resource for both laypeople and health professionals, internet-derived information has been recognized as a surrogate tool for estimating epidemiology and gathering data about patterns of disease and population behavior [6]. Internet query platforms, which allows to interact with internet-based data, have been considered a source of potentially useful and accessible resources, especially aimed to identify outbreaks and implement intervention strategies [7]. The US Institute of Medicine (IOM) has also recently acknowledged that the use of internet data in health care research holds promise, and may also "complement and extend the data foundations that presently exist"[8]. Numerous studies have also suggested great potential of these big data sets to detect/manage epidemic outbreaks [9, 10, 11, 12].

## 3.2 Internet query platforms

Big data sets are constantly generated nowadays as the activities of millions of users are collected from Internet-based services. In recent years, methods that harness Internet-based information have also been proposed, such as Google, Yahoo, and Baidu Internet searches, Twitter posts, Wikipedia article views, clinicians' queries, and crowdsourced self-reporting mobile apps such as Influenzanet (Europe), Flutracking (Australia), and Flu Near You (United States). Among them, GFT `https://ai.googleblog.com/2014/10/google-flu-trends-gets-brand-new-engine.html` has received the most attention and has inspired subsequent digital disease detection systems.

## 3.3 Accurate estimation of influenza epidemics (ARGO)

Yang et al., in their paper 'Accurate estimation of influenza epidemics using Google search data via ARGO' [3] have developed an Autoregression model with Google search data (ARGO). ARGO outperforms all previously available Google-search based models including the 2014 Google Flu Trends. ARGO captures the changes in people's online search behaviour over time as well as incorporates seasonality in influenza epidemics. ARGO is self-correcting, scalable, flexible and robust making it a potentially powerful tool which can be used for real-time tracking of other social events at multiple temporal and spatial resolutions.

### 3.4 A Deep Learning Framework in Epidemiological studies

Yuexin et al., in their paper 'Deep Learning for Epidemiology Predictions' [13] derived a deep learning framework to predict epidemiology profiles in the time series perspective. The deep learning framework consisted of two types of neural networks: Recurrent Neural Networks (RNNs) to capture long term correlation in the data and Convolutional Neural Network (CNNs) to fuse information from different data sources. This model was tested against standard Autoregressive (AR) methods and Gaussian Process Regression (GPR) and performed consistently better than the latter two. However, these models were trained and tested using historical data and not real-time search trends.

## 4 Accomplished milestones from original plan

### 4.1 Data Collection

Weekly Influenza Like Activity data is obtained from CDC `https://www.cdc.gov/flu/weekly/`. This data contains weighted and unweighted cases of Influenza Like Infections (ILI) from 2010 to 2019 for every state in the U.S. CDC complies actual number of cases of Influenza from clinics and labs and release a weekly report which usually runs two weeks behind real-time. This time series data is used to find correlated search queries on google using the trends in the time-series. These search queries correlating to the influenza trends will be the independent variables in our model. (Google correlates gives search queries that follow the same time-series as influenza cases in the U.S).

Eventually, we will try to develop an automated data extraction pipeline to extra data from CDC ILI weekly reports and find and extract corresponding search queries.

## 5 Approaches

The weighted influenza cases will be our target values (labels) $y_t$. The vector of log transformed of Google Search Queries at time $t$ will be the independent variables $X_t$. This is under the assumption that $X_t$ depends only on the ILI activity at the same time, this follows the intution that flu occurrence causes people to search flu-related information online. With these, we intend to build two models;

- **Deep Learning Model:** Our model is selected with three objectives;
  - It should be able to handle and learn from multiple input variables: We propose a LSTM model to acheive this as they can seamlessly model multivariate problems.
  - It should be able to learn temporal dependencies in the time-series: We propose using a Gated Recurrent Unit, which is a variation of Recurrent Neural Network with fewer parameters which allows it to learn patterns even at data-deficient cases.
  - We also propose another layer of Convoluted Neural Networks to capture local dependencies.

  So, our proposed model will be of three parts; two recurrent networks (LSTM and Gated recurrent unit) and a CNN.

- **Autoregressive Model:** We consider the logit-transformed CDC's weighted ILI activity level $p_t$ at time $t$ and $X_{i,t}$ the log transformed google search frequency term $i$ at time $t$. The autoregressive model is given by

$$y_t = \mu_y + \sum_{j=1}^{N} \alpha_j y_{t-j} + \sum_{i=1}^{K} \beta_i X_{i,t} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} N(0, \sigma^2)$$

### 5.1 Model result evaluation against GFT and ARGO results

We propose to build an ARGO model to evaluate our deep learning model. ARGO is the gold standard of influenza predictions after it outperformed Google Flu Trend. So, we will be building ARGO along with our

Deep Learning Model (peviously, ARGO has not been modeled at state-level but was modeled just at the country-level) and compare our Deep learning model with results from ARGO and Google Flu Trends.

# 6 Difference/Novelty

Traditionally epidemiological predictions are modeled from time-series perspective using Autoregressive (AR) methods and its variants like Gaussian Process Regression (GPR) methods. These methods make use of historical data to make usually a linear (or a pre-defined non-linear kernel) predictions to capture spatio-temporal patterns. This method is popular due to the reason that epidemiological predictions are usually weekly sampled statistics which provides limited training instances. But, as the data available grows in size and diversifies its sources, obtaining training instances is not an issue. This project is novel in a way that we propose use of Neural Networks to model multi-variate time-series data using real time google search queries instead of traditional AR methods.

Yuexin et al., in their paper 'Deep Learning for Epidemiology Predictions' [13] have used a **univariate** Deep learning framework to model influenza trends just from historical trends data. We would model a **multivariate** Deep learning framework with Google Search Query's as independent variables with instances being their frequencies at time $t$.

This project also models predictions at state-level in the U.S, which has not been done using Google Search Query data. Although, Google Flu Trends models prediction at state-level their model is a black-box to the public.

# 7 Difficulties or problems that you are experience (if you have)

- We are facing some trouble while comparing the .csv files: us-state data and us-weekly data in google correlate (`https://www.google.com/trends/correlate/#`) we are getting error:500 internal error.

- Conversion of year and week into date became difficult while pre-processing.

# 8 Project Plan/Timeline

| Task | Assigned to | Date |
|---|---|---|
| Data Collection | Sagar | 10/01/2019 |
| Exploratory Data Analysis | Ali, Nikhil | 11/01/2019 |
| Planning | Garima | 10/02/2019 |
| Project Proposal | Sagar, Ali, Garima, Nikhil | 10/03/2019 |
| Data Preprocessing | Sagar | 11/1/2019 |
| Feature Extraction | Nikhil, Ali | 11/3/2019 |
| Model Selection, Evaluation and Testing | Sagar | 11/4/2019 |
| Feature Selection | Ali, Garima | 11/4/2019 |
| Mid-progress Report | Sagar, Ali, Garima, Nikhil | 11/7/2019 |
| Model Training | TBD | TBD |
| Model Evaluation and Tuning | TBD | TBD |
| Final Report | TBD | TBD |
| Poster | TBD | TBD |

# 9 Response to the feedback

- How is our model different from ARGO?
  We are proposing the use of Neural Networks to model multi-variate time-series data using real time google search queries instead of traditional AR methods which is used in ARGO.
- Which deep learning model are we going to use?
  Our proposed model will be using the following Deep Learning Models: LSTM (recurrent networks) Gated recurrent unit(recurrent networks) CNN.
- Specific Plan for our deep learning model?
  For our Deep learning Model we are planning to follow the below plan:
  - It will handle and learn from multiple input variables by using LSTM model.
  - It will learn temporal dependencies in the time-series: by using a Gated Recurrent Unit.
  - We will also try to add another layer of Convoluted Neural Networks to capture local dependencies.

# References


[1] World Health Organization, "Influenza (seasonal) (world health org, geneva), fact sheet 211." https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal), October 1, 2019.

[2] J. Shaman and A. Karspeck, "Forecasting seasonal outbreaks of influenza," *Proceedings of the National Academy of Sciences*, vol. 109, no. 50, pp. 20425–20430, 2012.

[3] S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using google search data via argo," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14473–14478, 2015.

[4] S. Yang, M. Santillana, J. S. Brownstein, J. Gray, S. Richardson, and S. C. Kou, "Using electronic health records and internet search information for accurate influenza forecasting,"

[5] A. Ekman and J.-E. Litton, "New times, new needs; e-epidemiology," *European Journal of Epidemiology*, vol. 22, pp. 285–292, May 2007.

[6] J. S. Brownstein, C. C. Freifeld, and L. C. Madoff, "Digital disease detection — harnessing the web for public health surveillance," *New England Journal of Medicine*, vol. 360, no. 21, pp. 2153–2157, 2009. PMID: 19423867.

[7] M. Salathé, L. Bengtsson, T. J. Bodnar, D. D. Brewer, J. S. Brownstein, C. Buckee, E. M. Campbell, C. Cattuto, S. Khandelwal, P. L. Mabry, and A. Vespignani, "Digital epidemiology," *PLOS Computational Biology*, vol. 8, pp. 1–3, 07 2012.

[8] G. Cervellin, I. Comelli, and G. Lippi, "Is google trends a reliable tool for digital epidemiology? insights from different clinical settings," *Journal of Epidemiology and Global Health*, vol. 7, no. 3, pp. 185 – 189, 2017.

[9] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data,"

[10] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein, "Using Internet Searches for Influenza Surveillance," *Clinical Infectious Diseases*, vol. 47, pp. 1443–1448, 12 2008.

[11] Q. Yuan, E. O. Nsoesie, B. Lv, G. Peng, R. Chunara, and J. S. Brownstein, "Monitoring influenza epidemics in china with search query from baidu," *PLOS ONE*, vol. 8, pp. 1–7, 05 2013.

[12] M. J. Paul, M. Dredze, and D. a. Broniatowski, "Twitter improves influenza forecasting.," *PubMed*, vol. 6, 2014.

[13] Y. Wu, Y. Yang, H. Nishiura, and M. Saitoh, "Deep learning for epidemiological predictions," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, (New York, NY, USA), pp. 1085–1088, ACM, 2018.