Infectious Disease Spreading Pattern*

Final Report of The Doctors

GUANZHI WANG, The Hong Kong University of Science and Technology, China
JI LIU, The Hong Kong University of Science and Technology, China
JIAYI YE, Georgia Institute of Technology, China
QINHAN LIU, The Hong Kong University of Science and Technology, China
ZHIHONG CHEN, The Chinese University of Hong Kong, China

ACM Reference Format:
Guanzhi Wang, Ji Liu, Jiayi Ye, Qinhan Liu, and Zhihong Chen. 2018. Infectious Disease Spreading Pattern: Final Report of The Doctors. 13 pages.

1 INTRODUCTION

Nowadays, the spreading of infectious diseases has become a threat to public health because of rapid globalization process. An accurate detection of the spreading trends is crucial for effective initiation of public health intervention measures. However, it requires a lot of efforts to do such detection, and not every country has the public health infrastructure

*CX4242 Project 2018 Spring Georgia Tech

Authors' addresses: Guanzhi Wang, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, 999077, China, gwangaj@connect.ust.hk; Ji Liu, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, 999077, China, jliubk@connect.ust.hk; Jiayi Ye, Georgia Institute of Technology, North Ave NW, Atlanta, GA, 30332, China, jye71@gatech.edu; Qinhan Liu, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, 999077, China, qliuan@connect.ust.hk; Zhihong Chen, The Chinese University of Hong Kong, Shatin NT, Hong Kong, 999077, China, 1155077078@link.cuhk.edu.hk.

to conduct monitoring closely. Thus, this project aims to investigate the GIS(Geographic Information System) of global infectious diseases monitoring and prediction tools in order to assist the public to be aware of or the government to improve the system of prevention, control and treatment on these diseases (Heilmeier 4). The survey of 18 papers has enlightened our team to develop data visualization and applicable prediction tools for future disease detection.Work distribution of each group member is provided for reference. All team members have contributed similar amount of effort.

2 PROBLEM DEFINITION

In the current practice of disease spreading pattern analysis, the disease data focused on a particular district of land, and was plotted against time to show the trend of attributes in a local area. Mathematical models (usually linear regression model) are used to make predictions. (Heilmeier 2) However, such prediction problems are complicated and various factors can affect the spreading trend. Thus, the problem mainly focuses on how the current infectious diseases spread in geospatial time-scale could be displayed and what is the best way to predict future pattern.

3 SURVEY

Several study have been conducted to module and visualize the spread trend of infectious diseases. For the prediction, attempts can be summarized into 3 directions. The first is to use mathematical model to simulate the physical process of infections among the population [Perez and Dragicevic 2007], which can only be implemented in small scale prediction, eg, the urban region of Atlanta. The second is to employ statistical methods to consider various risk factors to evaluate the long term spatial and seasonal spread of diseases [M.C.Nicholson and T.N.Mather 1996] [Han and Drake 2016]. The factors can include climate, population, healthcare condition etc. This method have been successfully implemented in syphilis [Kenyon et al. 2016] and Tuberculosis [Naranbat et al. 2009]. And a systematic explanation can be found [Keeling et al. 2008]. The third method include various way to use web news and search record to predict the outbreak of infectious diseases [Polgreen et al. 2008] [Wilson and Brownstein 2009] [Ginsberg et al. 2009]. Past records can also be a parameter for consideration because the periodic nature of outbreaks. Google Flu Trends (GFT) demonstrated a marble progress in using web information and

past records to predict the flu [Olson et al. 2013]. However, as time passed studies show that its algorithms are problematic and could be further improved [Lazer et al. 2014], also this relatively good results in predict flu can not be repeated in other infectious diseases [Pelat et al. 2009]. Besides the three method, it is worth notice that several epidemiology studies in a network perspective [Tatem et al. 2006] [Christakis and Fowler 2009] has been conducted and shown meaningful results. Also several paper pointed out using machine learning could be a promising way to predict the spread of diseases. In this project, we will implement a machine learning model taking past records as inputs to generate future infection trends. For the visualization, Karlsson's paper [Karlsson et al. 2013] has done a meaningful step in visualize disease occurrence per primary care center to help health care industry preparing for their workload. Jones' paper [Jones et al. 2008] provided a classic visualization map for infectious disease spread pattern and outbreak location globally. However, the codes generating the map have not been made into a software nor a webpage. Freifeld's group [Carroll et al. 2014] built a web-based global visualization tool, HealthMap, to monitor the infection news with multi-dimensional filters. As pointed out in Carroll's [Carroll et al. 2014] and Schneiderman's paper [Schneiderman et al. 2013], expectations of a modern web-based visualization tool with multiple dimensions and filters and user-friendly interaction methods were made by the public, which is exactly what The Doctors is made to satisfy.

4 PROPOSED METHODS

The Doctors want to construct a 3D interactive world map to demonstrate the current spreading trends of some contagious diseases and build a prediction model for forecasting (Heilmeier 1). A heatmap is formed on geospatial time-serial data and updated to show the broadcasting patterns. The system is going to be successful because the information is conveyed in a cleaner and more straight-forward way comparing to existing methods(Heilmeier 3). As long as the heatmap is updated as the time varies, we should be able to show the trends as desired. In addition, we have validated our model by comparing its predictions with existing records and prepare a reasonably accurate model(Heilmeier 5).

5 EXPERIMENTS AND EVALUATION

In the progress report, the Tuberculosis disease database from Kaggle [Hena 2016] is used to develop visualization and Google Flu database [goo [n. d.]] to generate machine learning models.

5.1 Database Description

Name	Spatial Scale	Time Scale	Size	Downloaded/Scrapted
Kaggle TB	Global	2007-2014	2MB	Downloaded
Google Flu	Global	2003-2015	2MB	Downloaded
Google Dengue	Global	2003-2015	6KB	Downloaded

Fig. 1. Dataset

In this final report, three databases (Tuberculosis disease database from Kaggle, Google Flu database, and Google Dengue database) are used to develop visualization and generate machine learning models.

- TB dataset: It pertains to Tuberculosis spread across countries from 2007 to 2014.
- Google Flu dataset and Google Dengue dataset: Google gathered the queries of people conducted on Google.com that were related to influenza-like illness from 2003 to 2015.

List of Questions To Answer:

(1) Which attributes are included in the visualization to reflect the trending pattern across countries and time?
(2) How to design a user friendly interface?
(3) Which is the best model to predict future spreading pattern?
(4) How to compare the models?

5.2 Visualization

5.2.1 *Data Cleaning.*

- TB: The data is firstly cleaned to a more formatted way from .csv (Fig.2) to .json (Fig.3) and made consistent with the world map dataset (.tsv and .json files are combined).

Country	Year	Number of deaths due to	Number of deaths due to	Number of deaths due to
Afghanistan	2014	14 000	10 000	18 000
Afghanistan	2013	13 000	9 300	17 000
Afghanistan	2012	12 000	8 700	16 000
Afghanistan	2011	11 000	8 300	15 000
Afghanistan	2010	11 000	8 000	14 000
Afghanistan	2009	11 000	7 900	14 000
Afghanistan	2008	11 000	7 900	14 000
Afghanistan	2007	11 000	7 900	14 000
Albania	2014	17	12	23

Fig. 2. Original format of datasets

Fig. 3. Dataset converted to .json format

In particular, country id attribute is added to the Kaggle dataset to match with that in the world map dataset, and the countries which are not in the world map dataset are set to a fixed number to classify the missing data in later visualization.

- Google Flu and Google Dengue: Similar to the above technique, these two .csv files are restructured to .json. The number of Google queries in the datasets are recorded in the form of date. Thus, in order to to reflect the changes across time within a specific year, the dates attribute is converted to weeks attribute and the year is used as an attribute that user could select.

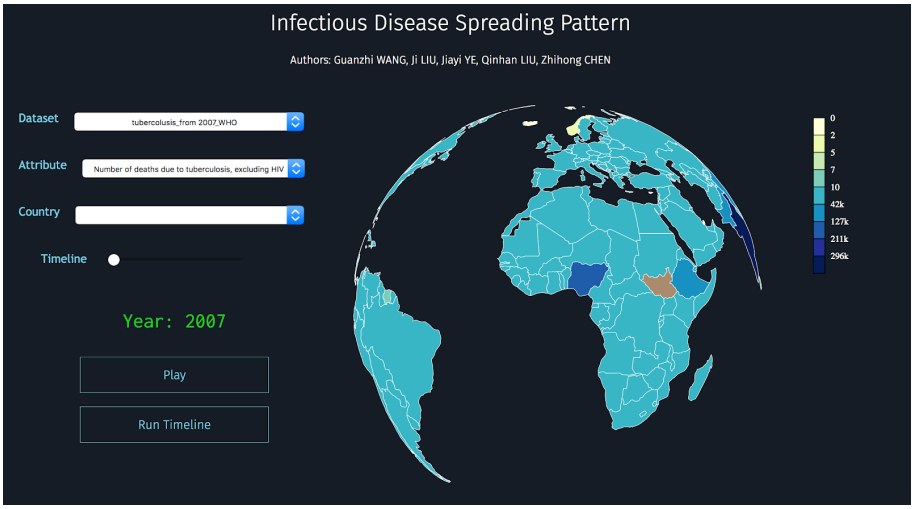

Fig. 4. 3D Earth Visualization

5.2.2 Algorithm/User Interface. For D3 visualization, we applied the similar methodology from Homework 2 Q7 to this project. Beyond that, we have presented the data in a three dimensional globe, implemented the dragging and moving functionalities, enabled the drop-down menu for disease dataset selections, country selection as well as attribute selection, and added dynamic timeline for past data and prediction data. When the Country and Year attributes are not selected, the globe will display heatmap changes in time order if Run Timeline is clicked.When the Dataset and Attribute are selected, the globe will self rotates and go through each country in alphabetical order. The effects are demonstrated below in Fig.4.

5.3 Prediction

In this section, Google Flu and Google Dengue datasets are analyzed, while the TB dataset are discarded due to the lack of enough data to make prediction.

5.3.1 Google Flu and Google Dengue Data. Google gathered the queries of people conducted on Google.com that were related to influenza-like illness from 2003 to 2008, sumed the digits per week as a series of data related to time and tried to use them to predict the actual influenza infection population both state-wisely and country-wisely. The results of their program demonstrated a very promising 0.9 correlation of their predictions comparing to the real infection statistics from Center of Disease. Typically, when the absolute correlation between two variables is higher than a certain number, 0.8 for example, we consider the two variables relating to each other linearly. So in this case, the web activity data from Google is a very good simulation of the actual infection population.
For our prediction model built on the Google data, the output (our predicted value) is thus not the actual population but a value that is strongly related to the population in a linear way.

5.3.2 Data Preparation. The original data is continuous time-series data of the number of Google queries related to the corresponding topics from 2002-12-29 to 2015-08-09, recorded on every Sunday.
Given the data, our target is to predict the number of sensible queries in the next week on every Sunday. On a Sunday, in order to predict the sum of queries of next week, knowing only the data of this week will not be sufficient, simply because there is no information about the history and trend of the number. But if all the records of the last 5 weeks are given, it would be somewhat enough to make a reasonable prediction for the case of flu spreading. So the task of our model is set to be that predicting the number of queries knowing the number of queries in the previous 5 weeks.
Furthermore, the differences of the number of queries between weeks matter in a sense that it carries the information about how the number evolves as time passes by. Adding them up, we should have enough features (in total 10, 5 are records of the past 5 weeks, 5 are the record differences) for building the machine learning model.

5.3.3 Prediction Models. Our team were not familiar with predicting time-series data at the beginning. Inspired by most popular models in sklearn, we chose a four models to implement and compared their performance:

```
def nn(X_train,y_train):
    # model = MLPClassifier(activation='relu', alpha=1e-05, batch_size='auto',beta_1=0.9, beta_2=0.999,
    early_stopping=False,epsilon=1e-08, hidden_layer_sizes=ARCHI, learning_rate='constant',learning_rate_init=0.02, max_iter=200,
    momentum=0.9,nesterovs_momentum=True, power_t=0.5, random_state=1, shuffle=True,solver='lbfgs', tol=0.0001,
    validation_fraction=0.1, verbose=False,warm_start=False)
    model = MLPRegressor(hidden_layer_sizes=(100,), activation='relu', solver='adam', alpha=0.0001, batch_size='auto', learning_rate
        ='constant', learning_rate_init=0.001, power_t=0.5, max_iter=200, shuffle=True, random_state=None, tol=0.0001, verbose=False
        , warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=0.9, beta_2
        =0.999, epsilon=1e-08)
    model.fit(X_train,y_train)
    label = "Neural Net-L=0.02,L" + str(ARCHI)
    return model,label
```

Fig. 5. Neural Network

```
def knn(X_train,y_train):
    # model = KNeighborsClassifier(n_neighbors=N_NEI)
    model = KNeighborsRegressor(n_neighbors=10, weights='distance', algorithm='auto', leaf_size=30, p=2, metric='minkowski',
        metric_params=None, n_jobs=1)
    model.fit(X_train,y_train)
    label = "KNN-Neighbors " + str(N_NEI)
    return model,label
```

Fig. 6. KNN Regression

```
def svr(X_train,y_train):
    model = SVR(kernel='rbf', degree=3, gamma='auto', coef0=0.0, tol=0.001, C=1.0, epsilon=0.1, shrinking=True, cache_size=200,
        verbose=False, max_iter=-1)
    model.fit(X_train,y_train)
    label = "SVM Regression"
    return model,label
```

Fig. 7. SVM Regression

```
def lr(X_train,y_train):
    model = LinearRegression(fit_intercept=True, normalize=False, copy_X=True, n_jobs=1)
    model.fit(X_train,y_train)
    label = "Linear Regression"
    return model,label
```

Fig. 8. Linear Regression

(1) Neural Network Regression shown in Fig.5.
(2) K Nearest Neighbor (10) Regression (Readings 1) shown in Fig.6
(3) Support Vector Machine Regression (Readings 1 2) shown in Fig.7
(4) Linear Regression shown in Fig.8

5.3.4 Metrics Used to Measure Performances. Then these values are computed based on the predicting values of the model and the actual values to help evaluate the performance of models:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Explained Variance Score (EVS)
  - It measures the proportion to which a mathematical model accounts for the variation (dispersion) of a given data set. The best possible score is 1, lower values are worse.
- R2 Score (R2)
  - The coefficient of determination, denoted R2 or r2 and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s). Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse).

5.3.5 Model Experiment and Evaluation. To train and test each model, the dataset is equally separated to 5 time-continuous portions. 4 of them are used to train each model, the remaining 1 is for later model testing. The exact portion to be tested on is chosen at random in the experiment.
The following is a example for the flu data of United States that is used to do a simple test on these 4 potential models.
As shown in Fig.9 and Fig.10, Neural Networks, K Nearest Neighbor and Linear Regression stand out as they achieve reasonably good predictions. But that is just for the flu data of United States.
A full cross-validated test on data of all countries on Google Flu Dataset (Fig.11) confirms with the result and further suggests that Linear Regression is generally the best model amongst all with R2 of 89.2%.
According to Fig.12, the performance of Neural Networks on Dengue data does not meet our initial expectation. For KNN, it does not have better performance than Linear Regression for both datasets. Hence, Linear Regression is the model that we are going to use for predictions.

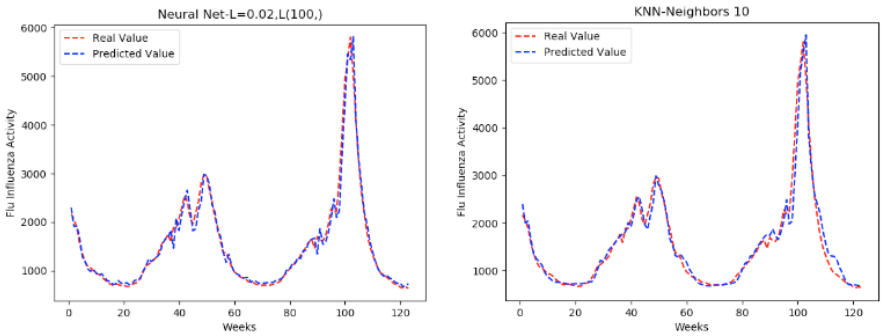

Figure-2 *Figure-3*

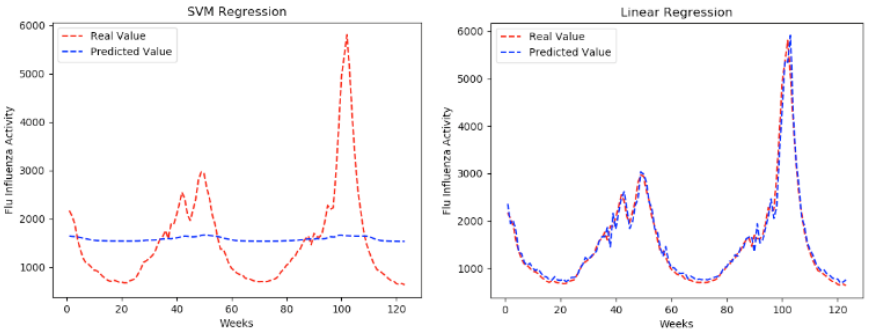

Figure-4 *Figure-5*

Fig. 9. Performance of 4 different models

United States	MAE	MSE	ESV	R2
Neural Network	184.037	111315.513	0.927	0.92
K Nearest Neighbor	225.502	169059.566	0.889	0.885
Support Vector Machine	762.867	1454199.847	0.036	0.007
Linear Regression	179.370	95464.810	0.935	0.935

Fig. 10. Metrics of models on flu data

5.3.6 Making Predictions. With a reasonably good model as described above, we can finally try to predict the future values using Linear Regression.

Notice that the model is trained for predicting the value of next week given the data of

Cross-Validated Over All Countries in Flu Dataset	MAE	MSE	ESV	R2
Neural Network	44.900	13099.092	0.882	0.877
K Nearest Neighbor	57.422	30176.453	0.857	0.846
Support Vector Machine	168.298	221195.485	0.307	0.149
Linear Regression	44.202	12050.025	0.897	0.892

Fig. 11. Metrics of models on flu data of cross validation

Cross-Validated Over All Countries in Dengue Dataset	MAE	MSE	ESV	R2
Neural Network	0.069	0.022	-0.632	-1.625
K Nearest Neighbor	0.028	0.004	0.761	0.715
Support Vector Machine	0.069	0.009	0.503	-6.224
Linear Regression	0.018	0.001	0.886	0.869

Fig. 12. Metrics of models on Dengue data

last 5 weeks, we can only make the predictions week by week and, more importantly, the newly predicted value will be utilized to compose the features for later prediction. The problem is, inevitably, that all the predicted values contain some error and using the predicted value as input to the model will just keep adding error to the next prediction. In fact, this is the same reason why the weather forecast can not be so accurate in a long term scale. Hence, here we only iterate the predicting process for 3 times to predict for the following 3 weeks (2015-08-16 to 2015-08-30) to reduce the effects of recurring precision error.

The generated prediction data are put back into the Google Flu and Google Dengue datasets as the same format with previous data. Therefore, the predicted results could also be visualized on the interactive map.

6 CONCLUSIONS AND DISCUSSION

In 2 months, a prototype of this project has been finished. It fully demonstrates the interaction flow of our desired system and the prediction model achieves a good accuracy. The fully functional system may need 2-5 more years on gathering practical data and

polishing predicting strategy (Heilmeier 5).
Visualization: Our system provides a cheaper and cleaner solution to solve the problem we defined in the beginning. For example, people could use our model to prevent the risk of choosing areas of infectious diseases as vacation destinations. The insurance industry could apply this model on risk analysis of insurance products (Heilmeier 4).
Prediction: In the future,integrate population density and other environmental parameters could be considered as factors to see whether a more clear correlation can be observed.

REFERENCES

[n. d.]. ([n. d.]). https://www.google.org/flutrends/about/

Lauren N Carroll, Alan P Au, Landon Todd Detwiler, Tsung-chieh Fu, Ian S Painter, and Neil F Abernethy. 2014. Visualization and analytics tools for infectious disease epidemiology: a systematic review. *Journal of biomedical informatics* 51 (2014), 287–298.

Nicholas A Christakis and James H Fowler. 2009. Social network visualization in epidemiology. *Norsk epidemiologi= Norwegian journal of epidemiology* 19, 1 (2009), 5.

Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012.

Barbara A Han and John M Drake. 2016. Future directions in analytics for infectious disease intelligence: Toward an integrated warning system for emerging pathogens. *EMBO reports* (2016), e201642534.

Hena. 2016. (Aug 2016). https://www.kaggle.com/henajose/determine-the-pattern-of-tuberculosis-spread/data

Kate E Jones, Nikkita G Patel, Marc A Levy, Adam Storeygard, Deborah Balk, John L Gittleman, and Peter Daszak. 2008. Global trends in emerging infectious diseases. *Nature* 451, 7181 (2008), 990.

David Karlsson, Joakim Ekberg, Armin Spreco, Henrik Eriksson, and Toomas Timpka. 2013. Visualization of infectious disease outbreaks in routine practice.. In *MedInfo*. 697–701.

Matthew James Keeling, Pejman Rohani, and Babak Pourbohloul. 2008. Modeling infectious diseases in humans and animals. *Clinical Infectious Diseases* 47, 6, Article 864-865 (2008).

Chris Richard Kenyon, Kara Osbak, and Achilleas Tsoumanis. 2016. The Global Epidemiology of Syphilis in the Past Century–A Systematic Review Based on Antenatal Syphilis Prevalence. *PLoS neglected tropical diseases* 10, 5 (2016), e0004711.

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.

M.C.Nicholson and T.N.Mather. 1996. Methods for Evaluating Lyme Disease Risks Using Geographic Information Systems and Geospatial Analysis. *Journal of Medical Entomology* 33, 5, Article 711-720

(1996).

Nymadawa Naranbat, Pagbajabyn Nymadawa, Kurt Schopfer, and Hans L Rieder. 2009. Seasonality of tuberculosis in an Eastern-Asian country with an extreme continental climate. *European Respiratory Journal* 34, 4 (2009), 921–925.

Donald R Olson, Kevin J Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen. 2013. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS computational biology* 9, 10 (2013), e1003256.

Camille Pelat, Clement Turbelin, Avner Bar-Hen, Antoine Flahault, and Alain-Jacques Valleron. 2009. More diseases tracked by using Google Trends. *Emerging infectious diseases* 15, 8 (2009), 1327.

Liliana Perez and Suzana Dragicevic. 2007. An agent-based approach for modeling dynamics of contagious disease spread. *International Journal of Health Geographics* 8, 2, Article 50 (2007).

Philip M Polgreen, Yiling Chen, David M Pennock, Forrest D Nelson, and Robert A Weinstein. 2008. Using internet searches for influenza surveillance. *Clinical infectious diseases* 47, 11 (2008), 1443–1448.

B Schneiderman, C Plaisant, and BW Hesse. 2013. Improving health and healthcare with interactive visualization methods. *HCIL Technical Report* 1 (2013), 1–13.

Andrew J Tatem, David J Rogers, and SI Hay. 2006. Global transport networks and infectious disease spread. *Advances in parasitology* 62 (2006), 293–343.

Kumanan Wilson and John S Brownstein. 2009. Early detection of disease outbreaks using the Internet. *Canadian Medical Association Journal* 180, 8 (2009), 829–831.