Development of a Real-Time Estimate of Flu Activity in the United States Using Dynamically Updated Lasso Regressions and Google Search Queries

A thesis presented

by

Wendong Zhang

To

Applied Mathematics

in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

March 29, 2013

Abstract

Seasonal influenza is a significant public health threat, annually infecting three to five million and killing 250,000 to 500,000 people worldwide. In addition, epidemics such as the pH1N1 outbreak remind the world of even graver dangers if faced with influenza subtypes for which there is little to no preexisting human immunity, with the potential to cause devastating pandemics. On both fronts, improving surveillance of outbreaks before and as they occur can quicken response times and possibly save millions of lives. Google Flu Trends (GFT) was developed as a tool for large-scale, real-time surveillance of influenza-like illnesses (ILI) using Google search queries. Unfortunately, it provided inaccurate estimates during crucial influenza outbreaks, particularly during the pH1N1 outbreak, and also, even after GFT received an update to its database of search queries used in its model, during the most recent 2012-13 influenza season. The data used in GFT was also not made publicly available. However, using open-source tools, we assemble here a proxy dataset of search queries and build a dynamically updated, lasso regression model for ILI surveillance, and successfully outperform GFT on the national level, especially during crucial periods for influenza surveillance. In particular, we achieve Pearson correlation of 0.828 where the original GFT model achieves 0.290 during pH1N1, and we achieve correlation of 0.982 and average relative error of 12.1% from Sep. 2009, when GFT was updated, through Jan. 2013, while the updated GFT model achieves correlation of 0.854 and average relative error of 30.7% during the same period. Our results show that improving the underlying regression model makes substantial and long-term improvements to ILI surveillance using query-based methods, rendering the update to GFT's database of search queries unnecessary.

Acknowledgements

First, and most importantly, I thank my wonderful advisor, Mauricio Santillana, who has been with me in this journey even before it took off, and then with me through to the end. No one could have asked for a better teacher, advisor, and friend.

I am grateful to John Ayers, for helping me frame the initial question that steered me these past eleven months, and for being on calls to answer my questions and provide new ideas. I am also thankful to Ben Althouse and Mark Dredze, for providing R support in this field and invaluable advice. Finally, I would like to thank my readers, John Brownstein and Yiling Chen, for agreeing to comment on my work, and I hope they will enjoy reading about what I, personally, have had tremendous pleasure building during this past year.

Contents

Chapter 1

Introduction

1.1 Motivation

An angry father walked into his local Target and demanded the manager explain why his high school daughter had received ads for baby products. Equally bewildered, the manager apologized, and called the father at home several days later to apologize once more. On the phone, the father instead sounded embarrassed, and actually apologized to the manager for his behavior the other day. It turned out he had finally learned that, indeed, his daughter had become pregnant without his knowing. Therefore, more remarkably, Target's advertisements knew before anyone else, manager or father, had a clue about the real situation [8]. While somewhat chilling, stories such as these highlighted the potential of collecting and analyzing "big data" to answer crucial questions that businesses, institutions, or society at large could have. The wide range of real-world questions one could tackle, the mathematical and statistical methods required for these tasks, and the philosophical implications behind data mining and machine learning all motivated me to spend my past few undergraduate years exploring this field.

This story of the angry father appeared in a New York Times magazine article in February 2012, which profiled Andrew Pole, a statistician, and his work at Target. Pole could estimate, accurate up to a narrow due date of a few weeks, which of Target's female loyalty program customers were pregnant, simply based on their shopping patterns. This information, in turn, provided valuable sales potential to Target. Though the resulting marketing was probably too blunt (and since then Target had introduced subtler advertising), what Pole developed ultimately accounted for several billion dollars in additional revenue for the company [8]. In fact, in all domains, "big data" and the ability to analyze it has stirred a firestorm of research, inventions, and adaptations. Certainly in business, companies increasingly desire to take advantage of swaths of electronic data generated both internally and externally. The book that piqued my interest freshman year to study more statistics and quantitative methods, *Super Crunchers*, mentions a broad range of applications, from using regressions to forecast the quality of upcoming wine vintages to making conclusions about social and family issues, such as the likelihood of divorce through credit card reports [1]. Of course, while it was exciting to read about new applications each year, and to study the methods theoretically in classrooms, I wished to choose a current, significant, real-world problem for my thesis that would require data mining methods, and take it from start to finish in studying, experimenting, and finally in some manner, solving the issue. After a few conversations, one fruitful discussion at the School of Public Health introduced me to influenza surveillance and the possible contributions I could make in this area.

1.2 Influenza and Google Flu Trends

1.2.1 Overview of Influenza and Surveillance Techniques

Influenza poses as a serious public health issue, with the annual seasonal form of flu epidemics resulting in three to five million severe cases worldwide annually, and about 250,000 to 500,000 deaths [25]. Of graver danger, moreover, is unpreparedness in the face of emerging pandemics. Researchers and public health officials maintain surveillance of non-seasonal flu outbreaks, especially from swine or avian influenza for which regular season human influenza vaccines are ineffective. If not treated or contained, an influenza subtype for which there is little or no preexisting human immunity could cause pandemic proportions of illnesses and deaths. For instance, the 1918 Spanish flu pandemic casts a historic shadow over the possible devastation of unpreparedness, with worldwide deaths estimated at 50 million [4]. Therefore, for strains that do not have readily available vaccines, as well as for the seasonal flu, early detection and prevention can play a vital role in minimizing influenza spread and mortality.

To this end in the United States, the U.S. Centers for Disease Control and Prevention (CDC) publishes national and regional data concerning influenza using virologic and clinical data on a weekly basis, but with a one to two week reporting lag. Specifically, among several methods, the CDC reports the number of influenza-like illness (ILI) related physician visits out of total physician visits for a particular week from more than 2,700 outpatient healthcare providers in all 50 states. ILI is defined as fever (temperature of 100F [37.8C] or greater) and a cough and/or a sore throat without a known cause other than influenza [5]. Since whether influenza is actually present in a patient can only be confirmed through laboratory tests, a slower process that makes reporting the true weekly incidence on an up-to-date basis difficult, ILI serves as an accepted substitute of the actual influenza level, and forecasting it prior to the CDC's

lagged reporting has been the aim of several recent novel methods using telecommunications, business, and Internet data. For example, telephone triage data used by hospitals and healthcare systems to direct patients to appropriate medical resources [9] and records of over-the-counter (OTC) pharmaceutical sales [7, 17, 20, 24] have been demonstrated as possible earlier indicators of ILI. Relating to online search activity, queries to a medical website [18], web access of illness-related articles [19], the number of clicks in an influenza-related Google Adsense campaign [10], and the frequency of certain influenza-related terms in Yahoo! search [23] all have demonstrated some ability to detect ILI prior to reports issued by the CDC. Finally, Flu Near You, an ongoing mobile and online crowdsourcing tool, allows the public to register and report their health information directly using a quick weekly survey, and then maps this information to provide local and national views of ILI [16].

1.2.2 Select Method: Google Flu Trends

Another widely-known model that used search activity to provide estimates of ILI was the Google Flu Trends (GFT) model [11], which mined its own massive database of search queries and provided forecasts that were reported online each week on the Google Flu Trends website [13]. This model and its shortcomings served as the springboard for making methodological improvements in ILI surveillance through search activity. To build their original model, Ginsberg et. al. scored the weekly time series data of 50 million of the most common queries in the U.S. according to their correlations to ILI-related outpatient visits data from the CDC. Cross-validated tests showed that an optimal set of 45 most-correlated terms should be utilized in order to construct a simple logit regression model to best predict ILI. Google Flu Trends aimed for nationwide surveillance capabilities, and automated the process for selecting search queries to use in its model, both significant contributions to the field of ILI surveillance.

These improvements tremendously broadened the relevance as well as the sophistication of influenza surveillance using the Internet, but two major issues hampered GFT. First, Google did not release the specific queries it used in its model, which made research by third parties to replicate exactly the GFT model impossible, and to expand upon the methodology highly difficult. Second, the original GFT was demonstrated to perform poorly during the pH1N1 influenza epidemic, an off-season flu outbreak, achieving a Pearson correlation of only 0.290 during the first critical wave of outbreak, which was one motivation for Google to update the set of search queries used in GFT [6]. The original training period for query selection ended the week of Mar. 11, 2007. The pH1N1 epidemic began the week of Mar. 29, 2009, and Google ended its updated period for query selection the week of Sep. 13, 2009, several months after the pH1N1 epidemic began [6]. Google therefore implied that they believed search behavior changed over time, and needed to update their set of search queries used in their regression to account for this changing behavior. Failing at pH1N1 and making an update after this epidemic began, Google essentially reselected their search queries to take into account terms that might have become relevant during this off-season flu epidemic. Unfortunately, this updated GFT also performed poorly in the most recent flu season, forecasting a peak in the 2012-13 flu season of over 10.5% against an actual peak in ILI of 6.1%. These insufficiencies suggested that GFT, likely due to its simple regression model, lacked the flexibility to adapt to changing conditions. Hence, we attempted to make improvements that not only addressed these insufficiencies, but did so through lesser-quality, open-source data.

1.3 Overview of Improved Methodology

First, to obtain data, we used Google Correlate as a proxy for search queries used in GFT [12]. An open-source tool, Google Correlate allows users to upload their own time series data, and then outputs the time series, standardized with zero mean and divided by sample standard deviation, of the top 100 most highly correlated search queries to the uploaded time series. Hence, by uploading weighted ILI-related outpatients visits at the national level from the CDC, we could obtain a dataset that resembled the top most correlated search queries to CDC data on ILI used by GFT. Second, through this open-source approach, we then applied several multivariate regression methods to dynamically select the appropriate predictive model as new information became available, with the lasso method providing the most accurate initial forecasts through 2009. Once we found that lasso provided substantial initial improvements in predictive performance, we continued to make weekly forecasts after 2009 and compared to published data from updated GFT. We found that, at the national level, and over the past six years when it could have hypothetically operated, lasso forecasted ILI on the whole more accurately than updated GFT, and especially made predictions accurately where they were most important. In contrast to GFT, which required its search queries to be updated in 2009, our model used the same set of queries through Oct. 14, 2012, selected using the original GFT's training period from Sep. 28, 2003, through Mar. 11, 2007. Unfortunately, the last week for which data was available from Google Correlate was Oct. 14, 2012, so to continue our forecasts, we used a separate tool called Google Trends to build a dataset on which we could run lasso. Google Trends outputs the search volume, rescaled as an integer from 0 to 100, of any search term the user provides. Unfortunately, because of this rescaling and the rounding to integers, low search volumes in particular weeks for queries results in an output of 0, and when too many 0's appear, Google Trends condenses weekly output into monthly output for these queries. Of the 100 terms from

Google Correlate, therefore, only 70 had weekly data through Google Trends. Hence, we built a script that augmented these 70 queries from Google Correlate essentially with over 2,700 additional queries we thought might be related to ILI, and obtained volume data from Google Trends. We then correlated these terms against CDC in our own ad-hoc replication of GFT methodology in a more recent correlation period, from Dec. 27, 2009, through Oct. 7, 2012, due to a lack of data in the earlier years for these search queries. The top 100 most correlated terms were used with lasso to forecast weeks subsequent to Oct. 14, 2012. Our patchwork model, compared to our original model, was less refined, as we could not be sure we used the actual top 100 most correlated search queries, nor could we train on the years from 2004 through 2009. Nevertheless, this model, which still used open-source data, managed to forecast the most recent flu season accurately.

In summary, prevention of both seasonal and off-season influenza epidemics would benefit greatly from improved surveillance. Considered in the face of pandemic strains or possible bioterrorism, improved early detection of influenza outbreaks could result in millions of additional lives saved. Officially, while the CDC would publish weekly reports on ILI in the United States, the cost and lag in reporting encouraged research into cheaper, more timely methods of surveillance to supplement existing techniques. To this end, while GFT represented one of the most comprehensive and sophisticated techniques of surveillance to date, several critical improvements could be made to the model to make it more accessible to exploration by the general public, as well as to make it more robust to changing conditions. We offered one possible set of improvements to these standing issues, and hoped future research would develop upon it to create an ever stronger ILI surveillance network.

Chapter 2

Method

2.1 Data Description

Our ultimate goal was to build a model that performed comparably to GFT on the national level, using open-source data and improved methodology. Therefore, our goal in collecting data was to find open-source search query data that resembled search queries used for GFT. As our baseline, ILI data from the CDC could be obtained for every season since the 1997-98 season [3]. GFT mined through 50 million of the most common search queries in the United States, where a query was defined as a complete exact sequence of terms from a user, to find a set of queries most correlated to this data from the CDC. Ginsberg et. al. performed this correlation on the regional level, using 4-fold cross-validation for each term in each of the 9 census regions, therefore producing 36 correlation values for each term, and taking the score of the query's performance as the average of these values. Using a simpler method, Google also built Google Correlate, which would provide the top 100 most correlated search queries at the national level to data a user could upload, as well as these queries' respective weekly time series [12]. We therefore used this tool to obtain an open-source dataset that reasonably substituted for the query data used in GFT,

which Google would not release. We uploaded weighted ILI data from the CDC from Sep. 28, 2003 through Mar. 11, 2007 into Google Correlate and obtained an output of presumably the 100 most correlated search queries. The time series for each of these queries from Correlate was not the actual volume, but the volume subtracted by the mean and divided by the sample standard deviation. The outputted time series also ranged from Jan. 4, 2004 through Oct. 14, 2012, no matter how early in the past or how recent were the uploaded data. Therefore, to compare to the original and updated GFT outputs, which ranged in combined training and forecast periods from Sep. 28, 2003 through the present-day, modifications to comparison periods in the past and tweaks to methodology in the present had to be made. Specifically, beyond Oct. 14, 2012, a separate Google-related tool, Google Trends, was used to build an ad-hoc database of highly correlated search queries on which our improved methodology could be performed.

Google Trends would allow users to enter keywords and would then return the search volume of the keyword over time, scaled in integer values from 0 to 100 [14]. Naturally, yet unfortunately, search volume tended upward over the years since 2004, and since volume remained an integer value between 0 to 100 in Trends, for most queries, volume would be denoted 0 most of the time in earlier years, generally between 2004 to 2009. Moreover, when volume was low by weekly counts for a keyword, Google Trend would aggregate data automatically to return the monthly volume instead for that keyword. All these barriers made obtaining data for all 100 queries outputted by Google Correlate impossible, since 30 of these queries had only monthly output. Recalling our goal at this point was to obtain as high quality data as possible (though never matching the original quality accessible to GFT) to use in our determined model, we compiled a database of approximately 2,700 additional terms with weekly and monthly time series related to ILI symptoms. We used a script to use the root queries {*cough, flu, fever, headache, aches, fatigue, exhaustion, sneezing, sore throat*}

as keywords in Trends, and then used as keywords related queries to these roots suggested by Trends, and then related queries to those searches, and so on, until we had a large database. Queries with only monthly output were discarded. In addition, this dataset also included the 70 queries with weekly time series in Trends from the original 100 queries obtained using Google Correlate. Finally, once this database was compiled, in total 697 queries with weekly data, the top 100 most correlated search queries of this database were obtained, simply by obtaining the correlation of each query to ILI data during the period of Dec. 27, 2009 through Oct. 7, 2012. This period was chosen to avoid poor Trends data from the earlier years, and to find correlations on regular season ILI data, averting the off-season data from pH1N1 during 2009. Once these 100 most correlated Trends queries were obtained, we continued our methodology to make forecasts for the weeks subsequent to Oct. 14, 2012.

2.2 Details on Google Flu Trends and its Update

2.2.1 Google Flu Trends: Original Model

Again, building GFT began with mining through the weekly search volumes for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. A set of 100 top influenza-related queries was chosen using 4-fold cross-validation of correlation between search volumes to regional ILI data from the CDC in each of the nine census regions in the U.S. Search volume here was measured by query fractions, which were proportions of particular search queries to the general pool of queries over time. Specifically, for every search query in a region, its time series was normalized by dividing the count for the query in a particular week by the total number of online search queries submitted in that location during the week, thereby normalizing the volumes across searches [11].

Once query fractions were obtained for each of the top 100 most correlated queries, these fractions were sequentially added from most correlated to least correlated into a single independent variable and cross-validated for best in-sample performance in a final simple linear (logit) model. Specifically, Google fit a linear model using the log-odds of an ILI physician visit versus the log-odds of an ILI-related search query:

$$logit(\mathbf{p}) = \beta_0 + \beta_1 * logit(\mathbf{q}) + \boldsymbol{\epsilon} \tag{2.1}$$

where $\mathbf{p}$ was the percentage of ILI physician visits, $\mathbf{q}$ was the ILI-related query fraction, β_0 was the intercept, β_1 was the multiplicative coefficient, and $\boldsymbol{\epsilon}$ was the error term. In essence, the GFT model regressed the proportion of ILI-related outpatient visits against this single variable obtained by summing the proportion of queries. Through in-sample cross-validation, Ginsberg et. al. determined that the summed query fractions of the top 45 most correlated search queries provided the optimal results. This in-sample period during which top correlated queries were obtained and trained in the GFT model ranged from Sep. 28, 2003 through Mar. 11, 2007. This model was finally verified with out-of-sample performance ranging from Mar. 18, 2007 through May 11, 2008. Subsequent results from GFT were also made publicly available at the GFT website, but the results likely stemmed instead from the model after its update in 2009 with an extended queries selection period.

2.2.2 Google Flu Trends: Update

During the spring of 2009, the influenza A (pH1N1) virus emerged and spread quickly to the United States, occurring out-of-season from annual seasonal influenza. Specifically, the complete "pH1N1 period" was defined to have occurred from Mar. 29, 2009 through Dec. 31, 2009. We followed Cook et. al.'s naming for further divisions, designating the time period from Sep. 28, 2009 through Mar. 29, 2009 as the "pre-

H1N1" period; within pH1N1, we defined Mar. 29, 2009 through Aug. 2, 2009 as "Wave 1", and Aug. 2, 2009 through Dec. 31, 2009 as "Wave 2" [6]. Two weeks, beginning Apr. 27, 2009, and May 3, 2009, were excluded from consideration due to high media attention. During the pH1N1 period, the original GFT model performed poorly in forecasting ILI levels compared to its out-of-sample forecasts for seasonal ILI. Especially in Wave 1, the original GFT model achieved a Pearson correlation of only 0.290 [6]. Cook et. al. mentioned that Google planned an update of the search queries used in GFT in 2009, and this update occurred using an extended correlation period from Sep. 28, 2003 through Sep. 13, 2009 during which correlations of search queries to ILI data were determined, as well as using an expanded pool of candidate search queries. Since this new training period occurred partially during pH1N1, Cook et. al. essentially implied that the set of queries used in GFT needed to be manually updated to account for changing search behaviors, such as during an unexpected epidemic like pH1N1, and that GFT in itself empirically did not appear flexible enough to adapt to changing conditions. Once this updated model was completed, Cook et. al. used it to produce both prospective estimates of ILI from Sep. through Dec. 2009 and retrospective estimates from Jul. 2003 through Sep. 2009 [6]. Finally, the estimates from the GFT website definitely reflected this update during the pH1N1 epidemic, and likely reflected this updated model (or other updated models thereafter) for what was shown and available for download after 2009 as well. As far as we know, however, the underlying simple linear (logit) regression model that Ginsberg et. al. used to train GFT remained unchanged in Cook et. al.'s update and in wherever output is taken from for present-day forecasts on the GFT website. In fact, when we completed our own methodology, we concluded that improvements to this underlying model would be sufficient to making substantial improvements to ILI forecasts, rendering this update of search queries that Cook et. al. pursued unnecessary.

2.3 Our Multivariate Approaches to ILI Forecasts

First, because Google did not release any of the sets of search queries along with their query fractions used in GFT, we needed to use a substitute dataset in order to build models to estimate ILI. These substitute datasets were obtained via Google Correlate and Google Trends and have already been described in detail. ILI data could be obtained through the CDC, while the estimates from updated GFT could be downloaded directly at `http://www.google.org/flutrends/`. Hence, once we built our model and made our forecasts, we had adequate benchmark data with which to evaluate our performance.

2.3.1 Initial Model: Unregularized Ordinary Least Squares

Once we obtained search queries and their standardized weekly volumes, our first, naive approach was to apply an ordinary least squares (OLS) multivariate regression on all of our search queries, with a different beta coefficient for each query, against ILI data. Namely, our model was

$$\mathbf{y} = \boldsymbol{\beta}^T \mathbf{X} \tag{2.2}$$

where $\mathbf{y}$ was the column vector of ILI data over time, $\boldsymbol{\beta}$ was the column vector of beta coefficients, including the intercept, and $\mathbf{X}$ was the matrix with its first row filled with 1's, then each row thereafter as one of the search queries over time.

We performed all of our experiments using RStudio v0.97.316, which ran R v2.15.2. To perform OLS regression, we used an initial training period from Jan. 4, 2004 (the first week in which Google Correlate output was actually available) through Mar. 4, 2007, regressing weighted ILI data during this time as the dependent variable against the 100 search queries as independent variables. After obtaining this model, we forecasted the ILI rate for the week of Mar. 18, 2007 using Google Correlate data

from this week in the trained model. Once this estimate was obtained, to forecast the subsequent week beginning Mar. 25, 2007, we incremented our training period to include the week of Mar. 11, 2007 and retrained our model. We continued this process until the end of 2009. Hence, our methodology accomplished two purposes: first, by training two weeks behind the week of interest for ILI, we simulated the the reality in which the most recently available ILI estimates from the CDC usually would provide information for up to two weeks behind the present week; second, by updating our trained model dynamically, we would be able to take advantage of new information as it appeared into our model.

Of course, OLS would only be valid given certain conditions, but we were not as concerned whether a similar method to OLS, perhaps generalized least squares or the application of some other transformations, would have been more statistically or otherwise valid. Instead, our purpose was entirely predictive, so we chose OLS as our initial, naive model to represent a multivariate approach that attempted to minimize the error between the model and the observed data. We felt taking a multivariate approach would perform better than Google's simple regression model because the separate beta coefficients for each of the search queries in OLS would give the model more flexibility in adapting to changing conditions over time. We also noted that (2.1) could equivalently be seen as fixing the number of search queries to the first 45 queries, summing across them (and transforming them through logit), and then weighing this value by a common factor, giving each of these 45 queries essentially the same beta coefficient. In contrast, OLS would have much more flexibility in choosing different beta coefficients for different queries, and using as many as necessary to adequately fit the data.

2.3.2 Addition: Principal Component Analysis

While our OLS model accomplished our broad goals of introducing a multivariate approach and dynamically updating the model as new information became available, we needed to address multicollinearity among the search queries. In fact, GFT managed to avoid this issue by summing across query fractions. That is, since we found the search queries to have as high correlation to ILI data as possible, multicollinearity among at least some of the 100 queries would likely be strong. One way to tackle this issue was to use principal component analysis (PCA), which broke apart a dataset with correlated vectors into uncorrelated components, ordered such that each would capture as much of the variance of the original data as possible.

We provide the mathematical derivation of PCA in the appendix. Once principal components of the search query data from Correlate were obtained, OLS regression of ILI data against the components was again performed dynamically in a similar fashion to our previous method, with an initial training period from Jan. 4, 2004 through Mar. 4, 2007, and with forecasts made through Dec. 27, 2009.

2.3.3 Final Model: Lasso

While PCA addressed multicollinearity, we were also concerned about overfitting. Overfitting is a phenomena that occurs when a model fits minor variations in the training data too well, therefore likely to model a lot of noise relative to true signals [22]. An obvious instance when this phenomena could occur would be when the number of independent variables equaled or exceeded the number of data points in the training period. Fortunately, this situation was not the case here. However, even if this situation did not occur, certain conditions could arise in training that could cause the model to overfit. Several principles of modeling can deal with overfitting, and in our case, we chose to consider ways that would reduce the number of independent variables used, that is, to reduce the complexity of the model.

In addition to this technical consideration, in our exploration, we were also interested in the public health as well as possible social implications of our experiments. Mathematically, the entire set of principal components and the original data contained the same amount of information. However, realistically, the principal components were uninterpretable, while each independent variable in the original data was a search query. Hence, we were also interested in a method that could preserve the original data while performing well, especially against overfitting on the training data. These considerations led us to use the lasso method for training on and forecasting ILI data.

'Lasso' stands for *least absolute shrinkage and selection operator*, and differs from OLS in that it minimizes the square errors with an additional constraint that the sum of the absolute value of the betas, or the L-1 norm, cannot be greater than a constant value. Specifically, if there are M beta values $\boldsymbol{\beta}$ to be considered, N dependent variables $\mathbf{y}$ of ILI data in the training period, and some constant t, the lasso estimate is defined as (from [15]):

$$\boldsymbol{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{M} x_{ij}\beta_j \right)^2 \text{ subject to } \sum_{j=1}^{M} |\beta_j| \leq t. \tag{2.3}$$

An equivalent way to write this equation is in Lagrangian form:

$$\boldsymbol{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{M} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{M} |\beta_j| \right\} \tag{2.4}$$

In this form, it is evident that the Lagrangian serves as a complexity component to the minimization. Moreover, due to taking the L-1 norm, the penalization actually causes shrinkage in the sum of the absolute values by forcing certain beta coefficients to 0 while preserving others. Hence, this methodology, with an appropriate selection for t or λ (note that as t gets extremely large or as λ approaches 0, we obtain once

again the full OLS regression), can reduce the number of beta coefficients used in the regression, and therefore account for issues of overfitting. Moreover, empirically, one of the issues with multicollinearity is that beta coefficients can sometimes become extreme, and the penalization in lasso can also help take care of this issue.

In order to perform the lasso method, we employed the *lars* package in R. This package computed the complete lasso solution for all values of the shrinkage parameter until the full OLS regression values were reached. The package *lars* also reported Mallows' C_p statistic, which would be calculated for a particular set of independent variables of size P as

$$C_p = \frac{SSE_p}{S^2} - N + 2P \tag{2.5}$$

where SSE_p was the error sum of squares of the model with these P variables, S^2 was the residual mean square after regression on all independent variables, and N was the sample size for modeling [21]. This value as a minimum served as an adequate measure for model selection against overfitting, so we attempted always to select the model that minimized this value. Once models were selected, we dynamically updated lasso over the same time frame as OLS and PCA, and in a similar fashion. As we realized that lasso performed better than OLS and PCA during this time frame, and also that it made sense practically and theoretically, we continued lasso estimates beyond 2009 as well, eventually requiring the use of Google Trends to build a dataset for the most recent past, with which to produce lasso estimates for those weeks.

Chapter 3

Results

Table 3.1 showed the search queries Google Correlate returned to be most correlated to CDC ILI data from Sep. 28, 2003 through Mar. 11, 2007. Moreover, to supplement this main set of queries, the queries data used in forecasts by lasso after Oct. 14, 2012 were shown in table 3.2. As noted, only 70 of the 100 original Google Correlate queries could be included among the 697 total queries with weekly Trends volume data. However, 46 of these 70 remained in the top 100.

Table 3.3 revealed performance of several methods in various periods of time, notably the across-the-board strong performance of the lasso against original and updated GFT on the national level. This table and its comparison periods and statistics are adopted with modifications from the structure of reported values from Cook et. al., with 'Original GFT' values adopted directly from them [6], and updated GFT values downloaded from the Flu Trends website, but we chose to include a few other categories that we felt were helpful for better intuition or more valid comparisons. For example, we needed to modify the initial 'Pre-pH1N1' comparisons Cook et. al. drew because these began on Sep. 28, 2003. Since we did not have Google Correlate data from 2003, we created a modified 'Pre-pH1N1*' category that began on Jan. 4, 2004. Moreover, since GFT was updated on Sep. 13, 2009, we created a 'Common

	Correlate Queries				
1	influenza type a	35	is the flu contagious	68	fever in adults
2	bronchitis	36	flu in children	69	decongestant
3	influenza a	37	fever flu	70	normal body
4	symptoms of pneumonia	38	take action tour	71	low body temperature
5	flu incubation	39	flu remedies	72	a fever
6	influenza incubation	40	flu report	73	influenza a symptoms
7	flu contagious	41	nasal congestion	74	dangerous fever
8	influenza contagious	42	fever reducer	75	is flu contagious
9	flu incubation period	43	sinus infections	76	lauderdale florida
10	tussionex	44	rhode island wrestling	77	hotel fort lauderdale
11	benzonatate	45	symptoms of influenza	78	webmail shaw ca
12	influenza symptoms	46	castaway bay	79	high fever
13	a influenza	47	coral by the sea	80	robitussin ac
14	sinus	48	cold or flu	81	bronchitis contagious
15	pneumonia	49	respiratory infection	82	indoor driving
16	flu fever	50	take action	83	tussionex pennkinetic
17	flu duration	51	respiratory flu	84	wrestling report
18	taste of chaos	52	soweto gospel	85	walking pneumonia
19	bronchitis symptoms	53	soweto gospel choir	86	days inn miami
20	symptoms of bronchitis	54	illinois wrestling	87	body temperature
21	how long does the flu last	55	how long is the flu contagious	88	phlegm
22	symptoms of the flu	56	cold symptoms	89	flu relief
23	taste of chaos tour	57	the taste of chaos	90	mt sunapee
24	influenza incubation period	58	is bronchitis	91	harlem globe
25	sinus infection	59	upper respiratory	92	levaquin
26	flu recovery	60	afrin	93	strep throat
27	chaos tour	61	painful cough	94	coughing
28	type a influenza	62	laprepsoccer	95	whistler snow
29	flu symptoms	63	upper respiratory infection	96	fever temperature
30	tessalon	64	amoxicillin	97	sales tax credit
31	type a flu	65	ski harness	98	glitches
32	treat the flu	66	robitussin dm	99	pennkinetic
33	treating the flu	67	treating flu	100	histinex
34	how to treat the flu				

Table 3.1: The top 100 correlated search queries obtained through Google Correlate by uploading CDC data from Sep. 2003 to Mar. 2007. Queries are numbered from most to least correlated.

Forecast Period' that began on this date, as a possible cutoff where updated GFT values were completely in predictive mode (if GFT were not updated again afterwards), and could therefore be validly compared to lasso predictions for test error.

Root mean square error (RMSE) was reported in Cook et. al., and we did so here as well, but we noted that the formula for RMSE for the N values of true ILI data $\mathbf{y}$

	Trends Queries				
1	tamiflu	35	sinus	68	tylenol cold
2	tamiflu side effects	36	influenza type a	69	low temperature
3	flu contagious	37	indoor driving	70	coricidin hbp
4	is the flu contagious	38	flu virus	71	upper respiratory infection
5	flu treatment	39	symptoms of pneumonia	72	stomach virus
6	type a flu	40	robitussin dm	73	common cold
7	flu in children	41	cough in children	74	wrestling report
8	influenza a	42	cold or flu	75	cold remedies
9	symptoms of the flu	43	cough and cold	76	the stomach flu
10	flu symptoms	44	stomach flu	77	normal body
11	flu incubation period	45	walking pneumonia	78	symptoms of strep
12	flu incubation	46	cold vs flu	79	nasal congestion
13	symptoms of flu	47	tussionex	80	cold and flu
14	flu remedies	48	cough medicine	81	nyquil
15	flu medicine	49	robitussin ac	82	benzonatate
16	influenza b	50	afrin	83	baby cough
17	influenza symptoms	51	how long does the flu last	84	is bronchitis
18	robitussin	52	cold symptoms	85	fever in children
19	bronchitis	53	bronchitis contagious	86	the flue
20	harlem globe	54	sinus infection	87	cold medicine
21	cough remedies	55	tessalon	88	coughing
22	type a influenza	56	infant cough	89	bad cough
23	cough remedy	57	child fever	90	pneumonia contagious
24	the flu	58	mt sunapee	91	decongestant
25	influenza	59	illinois wrestling	92	stomach flu contagious
26	remedies for cough	60	strep throat	93	stop coughing
27	pneumonia symptoms	61	flu symtoms	94	robitussin cough
28	symptoms of bronchitis	62	strep	95	pneumonia treatment
29	bronchitis symptoms	63	low body temperature	96	barking cough
30	delsym	64	respiratory infection	97	chest cough
31	coricidin	65	upper respiratory	98	strep throat symptoms
32	the flu virus	66	dry cough	99	coughing up
33	cough suppressant	67	toddler cough	100	sinus pain
34	pneumonia				

Table 3.2: The top 100 correlated queries obtained through Google Trends, chosen by correlating each of 697 queries against ILI data from Dec. 27, 2009 through Oct. 7, 2012. 46 of the 70 original Correlate queries that were included in the Trends queries appeared in this list.

and predictions of ILI $\hat{\mathbf{y}}$ was (adapted from [22]):

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{y}_n)^2} \tag{3.1}$$

Unfortunately, while RMSE values provided a way to compare models with similarly scaled outputs, it did not offer an intuitive interpretation for the amount models erred, because the square difference between true and predicted ILI data was not

	Pre-pH1N1	Pre-pH1N1*	pH1N1 Overall	pH1N1 Wave 1	pH1N1 Wave 2	post-pH1N1	Common Forecast Period
Dates begin	Sep '03	Jan '04	Mar '09	Mar '09	Aug '09	Dec '09	Sep '09
end	Mar '09	Mar '09	Dec '09	Aug '09	Dec '09	Feb '13	Feb '13
Correlation							
Lasso	n/a	0.991	0.976	0.828	0.970	0.975	0.982
Original GFT	0.906	n/a	0.912	0.290	0.916	n/a	n/a
Updated GFT	0.957	0.959	0.990	0.940	0.987	0.866	0.854
OLS	n/a	0.990	0.962	0.660	0.953	n/a	n/a
PCA	n/a	0.989	0.859	0.619	0.867	n/a	n/a
Relative Error							
Lasso	n/a	10.2%	13.8%	11.9%	15.0%	12.2%	12.1%
Original GFT	n/a	n/a	n/a	n/a	n/a	n/a	
Updated GFT	20.3%	20.0%	11.8%	6.38%	14.7%	31.7%	30.7%
OLS	n/a	10.9%	18.2%	17.8%	18.2%	n/a	n/a
PCA	n/a	10.8%	38.3%	36.5%	38.8%	n/a	n/a
RMSE							
Lasso	n/a	0.001	0.004	0.002	0.006	0.002	0.003
Original GFT	0.006	n/a	0.018	0.008	0.023	n/a	n/a
Updated GFT	0.004	0.003	0.005	0.001	0.006	0.010	0.010
OLS	n/a	0.002	0.006	0.003	0.007	n/a	n/a
PCA	n/a	0.002	0.012	0.005	0.016	n/a	n/a

Table 3.3: Performance summary, with correlation, relative error, and root mean square error (RMSE) reported. For RMSE, the unit for weekly outputs is the fraction of U.S. population with ILI, so RMSE gives some measure of the absolute error, on average, of trained/predicted vs. actual fractions each week. However, because RMSE does not take into account the absolute level of ILI each week, relative error provides the more intuitive and correct interpretation of how well a method is performing relative to the true data. Formulas can be found in the report for more details on this discrepancy. In terms of the table rows and columns, 'Pre-H1N1*' is a modified pre-H1N1 period that begins in Jan. 2004 rather than Sep. 2003 because Google Correlate does not output pre-2004 results. The 'Common Forecast Period' marks off Sep. 13, 2009 onward, because this date is the last reported date of when GFT queries were updated, so we presume GFT is entirely in predictive mode afterwards and can therefore be validly compared with lasso for predictive errors. Original GFT' values are taken from [6], while 'Updated GFT' values are downloaded from the GFT website.

taken relative to the true ILI rate each week. That is, a difference of 1% in true versus predicted ILI data for a week if the true rate were 10% and a difference of 1% if the true rate were 2% for a given week do not carry the same amount of practical significance. Therefore, we also wrote down a formula that made sense in this context:

$$\text{Relative Error } (\%) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{y_n - \hat{y}_n}{y_n} * 100\%\right)^2} \tag{3.2}$$

where we multiplied by 100 to let the value easily be read and interpreted as a percentage. This formula could approximately be interpreted as the average percent error a model's output deviated from the truth. RMSE did not allow this easy interpretation because the average that RMSE reported, if divided by the average ILI level over a certain period, would still not account for variations in the ILI, such as the hypothetical example of 2% versus 10% true rates, and would therefore typically understate the severity of the error.

This table, 3.3, made comparisons during the interesting periods for GFT, namely during pH1N1 and during the recent 2012-13 flu season, but we further wished to highlight these periods in graphical form. First, figure 3.1 showed two graphs, ranged from Jan. 4, 2004 through Dec. 27, 2009, displaying updated GFT, lasso, OLS, and PCA. To avoid overfitting when using OLS as well as to address the unstable outputs in PCA seen after pH1N1 began, lasso was used to make predictions for the rest of the time after 2009 and was compared to updated GFT.

Figure 3.2 zoomed in on the pH1N1 period, with 'Original GFT' values extracted from Cook et. al. using the program GraphClick v3.0.2. Lasso performed far better than the original GFT during this period, even in an estimated fashion of the latter's output, and it performed comparably to the updated GFT. It was important to note that the update for GFT occurred during pH1N1. Figure 3.3, on the other hand, showed that despite GFT's update, lasso outperformed updated GFT significantly in the most recent 2012-13 flu season. Updated GFT estimated a peak in the season of over 10%, while the peak actually tipped at around 6% in ILI data, and lasso successfully estimated this turnaround.

Finally, it was instructive to plot and examine coefficient values of betas that were 'non-zero' in the various models. Because OLS and PCA were not expected to produce any coefficients of 0 value, 'zero' here was defined as any absolute value of an output for a particular model, from Mar. 18, 2007 through Dec. 27, 2009 (when

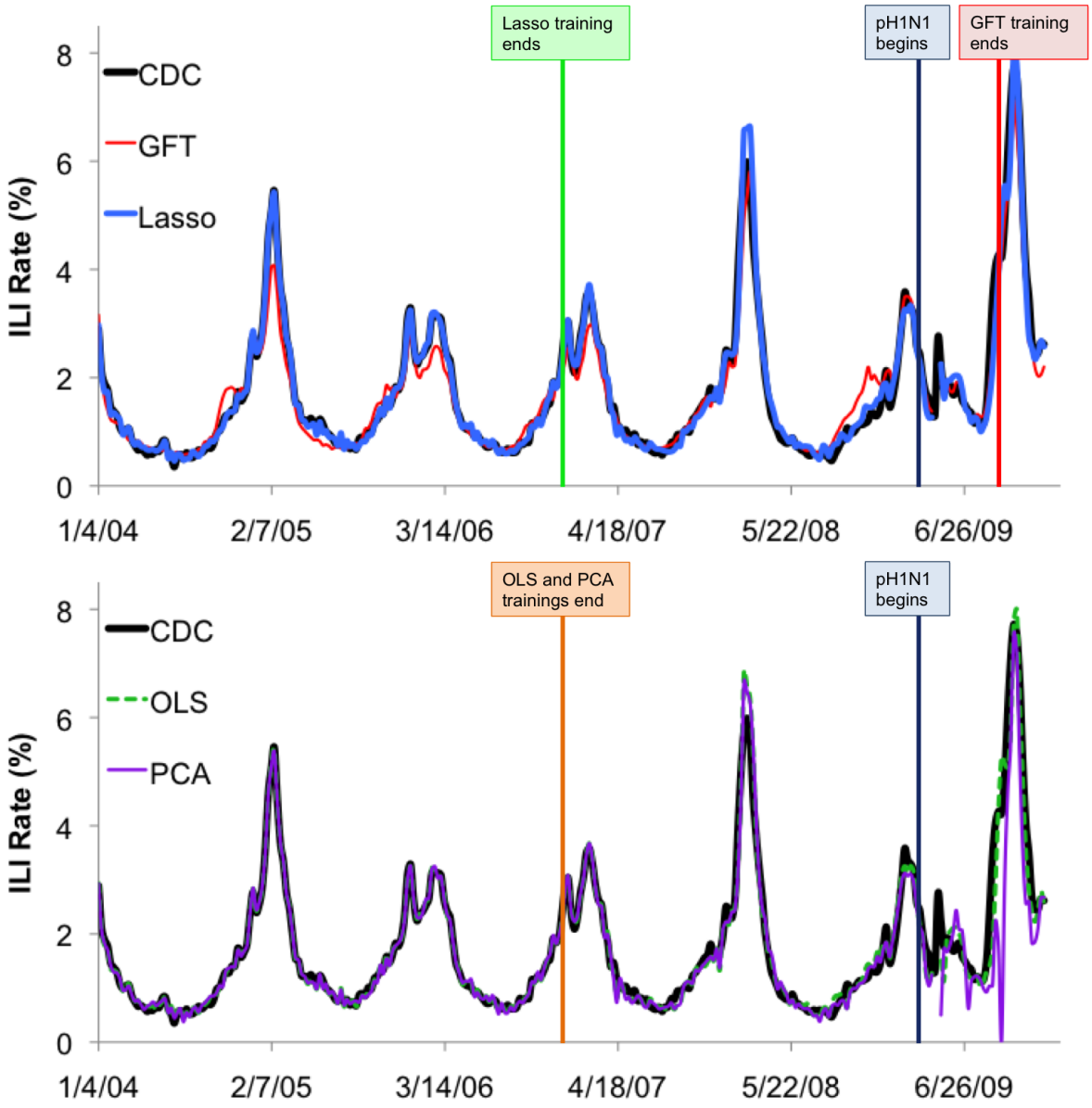

Figure 3.1: Graph of all the methods over time, until end of 2009. Because of unstable outputs in the OLS and PCA methods in these initial training and predictive periods, the lasso method was chosen as the method of comparison against Google Flu Trends for the rest of the time post-2009. The training period for updated GFT ends Sep. 13, 2009, after the end of the training period for the other methods, which end Mar. 11, 2007, also marking the end of the training period of the original GFT model.

lasso, OLS, and PCA were all three in predictive mode), that was less than or equal to $0.01 * \max\{\text{abs. value of outputs}\}$, or 100 times less than the maximum absolute value of outputs.

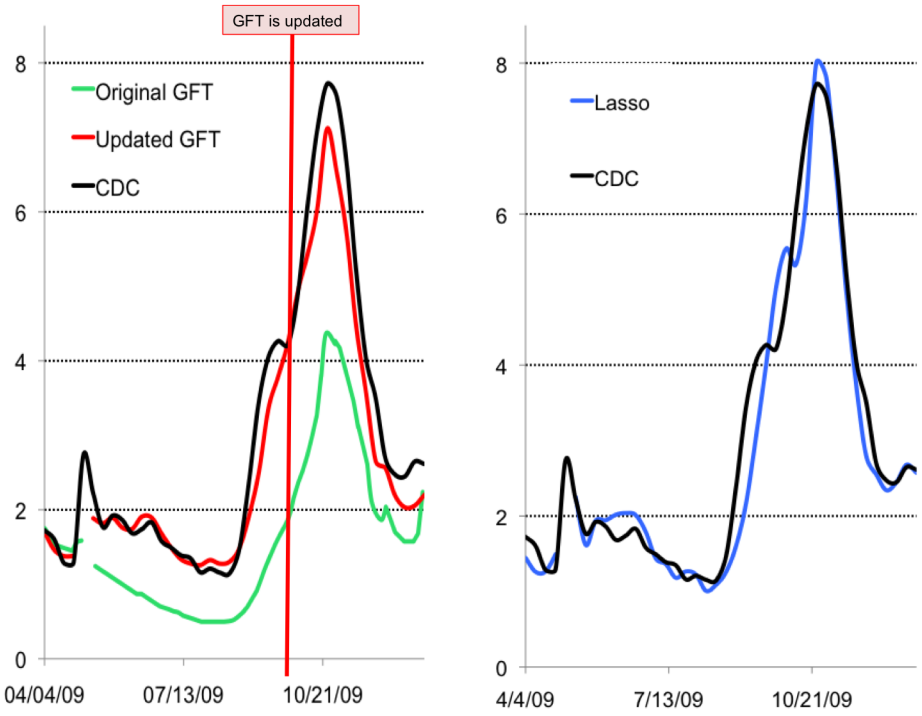

Figure 3.2: On the left, an estimated graph of the original Google Flu Trends (GFT) model, its update, compared to a graph on the right of the lasso method during the entire pH1N1 season from Mar. 29, 2009, through Dec. 27, 2009. GFT's update is also shown.

Under this heuristic, figure 3.4 showed the number of these non-zero coefficients present in the models over time, as well as the average over these periods. We also plotted heat maps of lasso, OLS, and PCA coefficients for each query over time in figures 3.5-3.6. For figure 3.5, we plotted the query coefficients over the entire predictive timespan of lasso from Correlate queries, until Oct. 14, 2012. We noted that even at our heuristic, lasso produced far more 'zero' coefficients than OLS did, which supported our view that lasso would help us prevent overfitting through the regularization component. Moreover, figure 3.6 showed, as expected, that while OLS coefficients varied in strength across queries over time, PCA displayed a pattern of high coefficients to low coefficients as we moved left to right in the bottom graph, therefore confirming that the methodology was working and that PCA managed to find principal components such that most of the variance of the original data could be accounted for in the first few principal components.

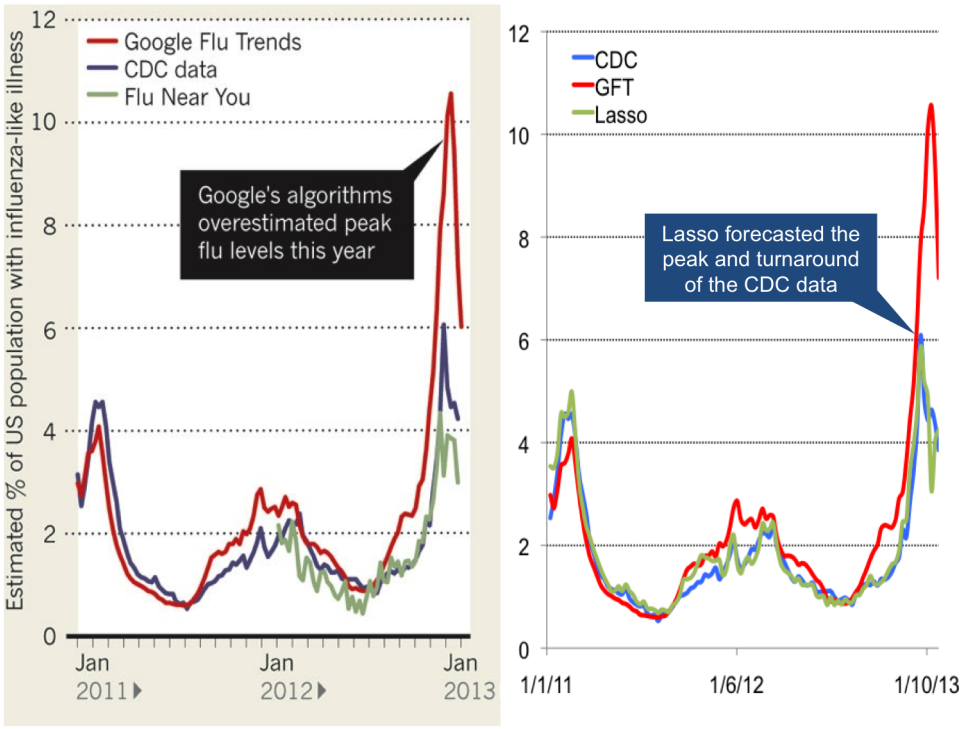

Figure 3.3: On the left, a recent graphic in Nature [2] examining Google Flu Trends' (GFT) failure to accurately forecast the recent flu season, drastically overshooting the peak of the season in its forecast. On the right, a graph during the same period of GFT and CDC data along with forecasts using lasso, clearly showing lasso's more accurate results.

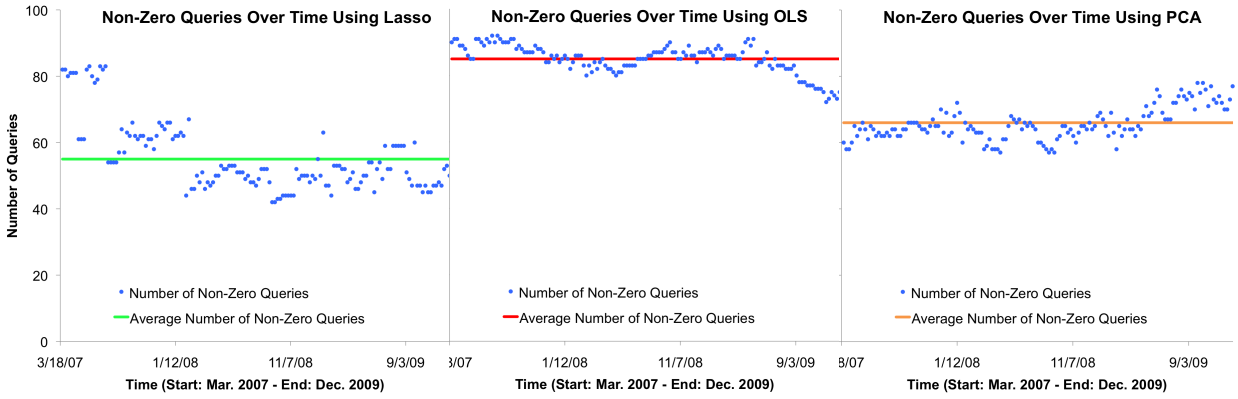

Figure 3.4: Graphs showing the number of non-zero coefficients of terms over time, from left to right showing the methods lasso, OLS, and PCA, respectively, starting in Mar. 2007 through Dec. 2009. 'Zero' for each model is defined as any output for that model having an absolute value less than or equal to 0.01**(max abs value of the outputs)* during this period.

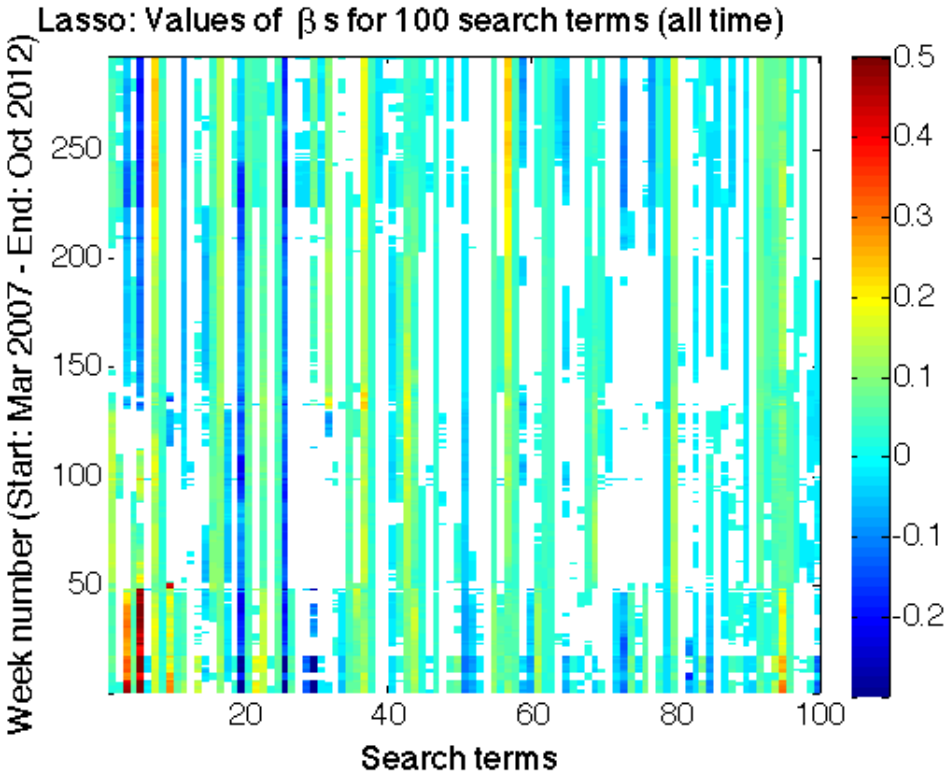

Figure 3.5: Displays which coefficients of search queries used in the lasso method are non-zero over time, from the weeks of Mar. 18, 2007 through Oct. 14, 2012. The x-axis shows the search queries obtained from Google Correlate and ordered by correlation to uploaded CDC data from Sep. 28, 2003 to Mar. 11, 2007 (see Table 3.1 for more details). For all the methods, to create these graphs, a query with a 'zero' coefficient is defined as a coefficient value of absolute value less than or equal to 0.01*(max abs value)*, in other words, 100 times less than or equal to the maximum coefficient value over time for any query using the lasso method. These 'zero' values are shown in white in the graph.

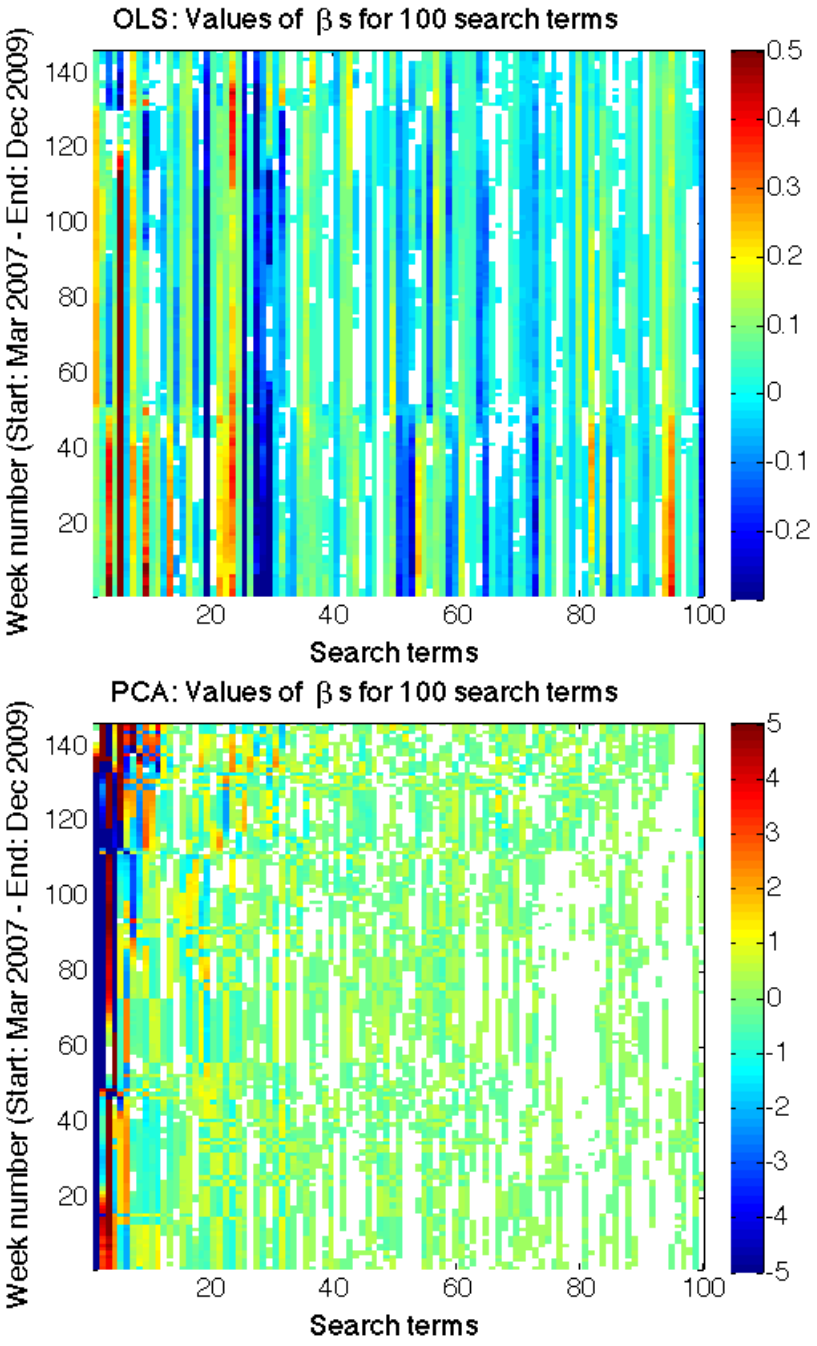

Figure 3.6: Graphs showing the number of non-zero coefficients of terms over time, the top figure of OLS and the bottom figure of PCA, from Mar. 18, 2007 through Dec. 27, 2009. The y-axis shows number of weeks since Mar. 11, 2007, and the x-axis shows the search queries from Google Correlate for OLS, ordered by correlation to uploaded CDC data from Sep. 28, 2003 through Mar. 11, 2007 (see Table 3.1 for more details), and the principal components for PCA. For all the methods, to create these graphs, a query with a 'zero' coefficient is defined as a coefficient value of absolute value less than or equal to 0.01*(max abs value)*, in other words, 100 times less than or equal to the maximum coefficient value over time for any query using the lasso method. These 'zero' values are shown in white in the graph. Note that while significant coefficients for OLS are scattered throughout the queries, the significant coefficients for PCA are concentrated in the first few principal components.

Chapter 4

Discussion

4.1 Impact of my Work

In summary, I studied the Google Flu Trends model and noticed two inherent weaknesses: first, that it used unreleased private data, and second, that it trained a simple logit regression on its search query data. Moreover, there was the unproven question of whether GFT could actually forecast ILI, or whether it simply forecasted spurious behavior similar to ILI during seasonal influenza's annual crests and troughs. When pH1N1 occurred, the original GFT model performed badly, requiring an update and retraining in its set of search queries from which the model was built. What truly compromised this approach, however, was that the update took place during the week of Sep. 13, 2009, several months after pH1N1 broke out in the United States. Therefore, GFT essentially admitted defeat in the face of the task of forecasting this off-season flu epidemic, for it retrained partially with pH1N1 data, perhaps in order to take into account changing online search behaviors among users in its body of search queries. However, in machine learning parlance, this decision made what should have been a holy grail of test data, namely off-season pH1N1 data and therefore untouchable by the model except for testing, partially available as the training data, altogether

compromising GFT from showing it could make any crucial accurate predictions of ILI rates in the United States over time.

I took these questions and issues into consideration as I built my own model of ILI surveillance. I used Google Correlate and Google Flu Trends to build open-source datasets of search queries that were highly correlated to ILI data. I tried several multivariate approaches to forecasting ILI, ultimately settling on lasso both for its robustness against overfitting and for its ability to zero out certain search queries while preserving others. I made sure always to keep clear my training from my predictions. While I displayed in table 3.3 a pre-H1N1* period that included both training and test data, this column was only meant to replicate the category used in Cook et. al., and I always focused instead on comparing predictions with predictions. On this focus, I achieved excellent performance in the two periods since training when forecasts were most important. First, during pH1N1 (see figure 3.2), my model achieved Pearson correlation of 0.828 during pH1N1 Wave 1, the most problematic period for original GFT, which achieved 0.290. Second, during the most recent influenza season (see figure 3.3), a lot of buzz occurred a few months prior to this writing that a severe influenza season was in session: it certainly did not turn out to be a light season, but definitely lighter than the 10%+ that updated GFT forecasted, versus an actual ILI rate of approximately 6%. In my reported statistics, this discrepancy and my better performance were reflected in the post-2009 statistics, where I achieved Pearson correlation of 0.975 and an average relative error of 12.2% in my forecasts. In contrast, updated GFT achieved correlation of 0.866 and an average relative error of 31.7%, almost triple my error percentage.

Therefore, I felt my model improvements were substantial. I managed to produce a model using lesser-quality, open-source data (which I would be more than happy to release) that over the past six years when it could have hypothetically operated, would have consistently provided accurate estimates of ILI rates, particularly when

such alerts were most crucial. Finally, my results showed that improvements to the underlying regression techniques were sufficient in gaining substantially improved forecasts, and that the updated queries process Cook et. al. undertook was probably unnecessary. Public health officials would like accurate, immediate surveillance as the first step in defending against possible outbreaks of epidemics, whether natural or man-made, and Google Flu Trends provided an exciting initial model of surveillance using search queries that could signal the current condition of ILI in the United States faster than the weekly reports by the CDC, generally delayed by two weeks. With the two updates I made, while relatively simple and perhaps even rudimentary compared to Google's wealth of data and computing capacity, I managed to update the GFT model to a more mathematically correct and practically relevant tool for ILI surveillance.

4.2 Remaining Considerations

In this section, I would like to provide more details about issues that may have occurred to readers while following the methods and results.

4.2.1 Training vs. Forecasting

It may have occurred to readers that while we highlighted Cook et. al.'s update of GFT as having compromised test data, we also provided an update to our list of search queries in Oct. 2012, albeit without choice after Google Correlate provided no more data. However, we would contend that we proceeded carefully throughout our methodology to provide valid forecasts that were not tainted in some fashion by the training of our model. Therefore, it would be important at this point to speak more carefully about the two-tiered training process of an ILI surveillance model using search queries. Training the ILI surveillance model required two steps. First, data had

to be collected, and whether GFT or my model, this process always required ILI data from the CDC be correlated in some way to a database of search queries, taking the queries with highest correlation as model-worthy search queries. The period during which this correlation occurred was considered a "training" period, but throughout our discussions, we tried not to refer to it as training and always either described it or referred to it as a correlation period. Once data was collected, a model had to be trained on this search query, and here GFT and our model differed, where GFT used a simple logit regression, while we used the lasso method.

The key, however, was that **both types of training had to occur before any forecasts could be made**. Cook et. al. admitted that the updated model provided "retrospective estimates from July 2003 through September 2009" [6]. It was still unclear to us the exact training period used after queries were updated, but at worst, training of updated GFT's simple logit model occurred during the same period as the extended update of search queries, ended Sep. 13, 2009, and at best, training occurred during the original GFT's training period, ended Mar. 11, 2007, with "forecasts" thereafter. Nevertheless, in either case, GFT already updated its set of search queries with correlated queries from a period after pH1N1 began, which constituted integrating a portion of the test period into the training period. The reasoning against this practice was simple: hypothetically, if we could return to Mar. 2009, when pH1N1 began, at best we could have updated our search queries in its correlation to ILI data up to this time, and not through Sep. 2009. Therefore, the scheme Cook et. al. proposed for updating their model was impractical in terms of forecasting. Moreover, our results in the end demonstrated this scheme to be unnecessary in improving GFT.

In contrast, we sought to maximize the quality of our forecasts while never violating this hypothetical boundary between past and future. Even when we updated our search queries after Google Correlate data ended, we were careful to find correla-

tions between Trends data and ILI data two weeks prior to the period when forecasts were required, ended Oct. 7, 2012 (with appended forecasts beginning Oct. 21, 2012). Finally, while we constantly retrained lasso with new beta coefficients in a dynamic process as new data would have hypothetically become available, we always ended training two weeks prior to our desired week of forecasting, which simulated the unavailability of ILI data for a given week until two weeks after that time passed. Therefore, while we maximized the amount of training we could validly perform, we always kept a careful and solid boundary between training and forecasting.

4.2.2 Future Work: Coefficients, the Ridge Regression, and Transformations

As I described, we chose the lasso method because of its regularization component to prevent overfitting, as well as for its ability to preserve some queries while zeroing out the coefficients of others, because it constrained the L-1 norm. Whether by doing so, the model found any significant patterns in search behavior over time was left undetermined. Certainly, as shown in figure 3.4, lasso did not use all of the search queries over time, and as seen in figure 3.5, some of the queries seemed to remain zero throughout the trainings and forecasts, while non-zero queries gradually faded in and out of use. Hence, perhaps there could be some important social cues to be gathered in analyzing particular search queries over time, but it remained beyond the current scope of this paper to do so in a rigorous framework, and we avoided merely speculating on social implications. The results, however, would remain available for future explorations in this area.

Instead, the results showed empirically that while lasso performed the best, a full least squares regression (OLS) also performed somewhat well. Therefore, making sure our model could adapt over time in a smart fashion was crucial. Empirically, lasso performed this adaptation wonderfully, but from a mathematical perspective in this

context, we had no reason to prefer the L-1 norm to, say, the L-2 norm. Specifically, adapting equation 2.3 and 2.4, we could alter the constraint such that the equations became, respectively (from [15]):

$$\boldsymbol{\beta}^{lasso} = \arg\min_{\beta} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{M} x_{ij}\beta_j \right)^2 \text{ subject to } \sum_{j=1}^{M} \beta_j^2 \leq t. \tag{4.1}$$

and in Lagrangian form:

$$\boldsymbol{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{M} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{M} \beta_j^2 \right\} \tag{4.2}$$

Known as the ridge regression, because of the constraint on the L-2 norm, this equation's minimization would not generally force beta coefficients to zero. However, mathematically, since we were not interested in solving the equations analytically or serving other mathematical purposes where the lasso and ridge differed, the regularization present in the ridge should have forecasted comparably to the lasso. It would be interesting in future work to see if this would be the case.

Finally, possibly because forecasts were so close to actual ILI values each week, the forecast values never dipped below 0 nor did they ever exceed 100, the accepted range of outputs since forecasts were percentages, though nothing in my model actually prevented these possibilities from occurring. On the one hand, mathematically, this issue was not particularly grave, because models would always have shortcomings and could only approximate the true model, if there were such a thing. In fact, this issue seemed analogous to certain statistical modeling scenarios, where a Gaussian model with a tight variance would be fitted to a particular set of data and found to forecast well, when in fact the data could never be negative. Indeed, empirically our results were fine, so in a sense, this issue with our model could be accepted. On the other hand, similar to the Gaussian model and how a log-transformation could

restrict the domain appropriately, in future improvements to our model, we would explore a transformation similar to the logit regressions performed in GFT in order to restrict the possible range of our model's outputs appropriately.

4.3 Concluding Thoughts

While I hope my work was presented clearly, the past few months of exploration were far from straightforward, and I felt extremely fortunate that careful analysis and testings yielded substantial improvements to the existing GFT model. Usually cynical about finding significant results, I nevertheless had to conclude for myself that the results here were, in fact, significant, having only tried three different methods with educated motivations for moving onto each, and then having an overwhelming body of test data post-2009 with which to test my final model. In other words, I was guarded against philosophical issues concerning data snooping, where even when no significant results should be found in the data, if a slight probability exists *a priori* that a test would find good results anyway, then the total probability that one of my tests would yield good results anyway increases as more tests are performed.

I felt fortunate as well that I managed to tackle a project from start to finish according to my original academic wishes, that I could take a real-world issue and use data mining and machine learning techniques correctly to make substantial contributions. For example, even now, I am adapting my research into a separate paper that hopefully could be published among the public health community. In addition, as this influenza season played out, some big news sources picked up on GFT's overestimate of ILI, and I felt a sense of humble disbelief that my work mattered in the midst of these stories. Of course, in perspective, my work probably seemed like it mattered more because I focused on it for these past few months, so any mention in a common news outlet with minor relevance to it would seem important to me. Still, I would

say that if my work did not move real mountains in the world, it at least could not be buried under rugs of irrelevancy. Within the public health community, indeed my model represented substantial improvements to the original and updated GFT models, and hopefully over time the issues I addressed and the solutions I proposed could be further studied, discussed, and implemented. More importantly to me, however, this process showed me the relevance of my schoolwork to the real world, that I could continue from this point to adapt these methods for other issues, perhaps to issues that could move real mountains, to build applications that could inspire still others with stories as I had been inspired.

Appendix A

PCA Derivation

Principal component analysis (PCA) uses a combination of statistical and linear algebra methods to break a set of data with possible correlation into uncorrelated components, each capturing as much of the variance in the original data as possible. PCA can be performed equivalently either by 1) eigenvalue decomposition of a covariance matrix of data, or 2) singular value decomposition of the data itself. Explanations to both, as well as a brief description of the relationship between them, will be provided.

First, using the covariance matrix, a data matrix $\mathbf{A}_0$, the matrix can be presented as

$$\mathbf{A}_0 = \begin{pmatrix} & & & & \\ \vec{x}_1 & \vec{x}_2 & \vec{x}_3 & ... & \vec{x}_N \\ & & & & \end{pmatrix}, \text{where } \vec{x}_n, \forall n \in N \text{ is an } M \text{ x 1 column vector.}$$

An important step at this point in principal component analysis is to subtract each entry by the mean of the column. This step ensures the first principal component

yields the direction of maximum variance. **Let A represent this mean-centered matrix.** Of these vectors, an N x N sample covariance matrix $\mathbf{\Sigma}$ between the column vectors can be derived by:

$$\mathbf{\Sigma} = \frac{1}{N-1}\mathbf{A}^T\mathbf{A} \tag{A.1}$$

The orthogonal eigenvectors of this matrix are the principal components if ordered by decreasing eigenvalues. For large amounts of data, efficient extraction of eigenvalues and eigvenvectors can be performed using the QR algorithm.

For PCA by singular value decomposition and its relationship to the sample covariance method, any MxN matrix $\mathbf{A}$ can be decomposed into

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}_0\mathbf{V}^T \tag{A.2}$$

where $\mathbf{U}$ is an MxM matrix containing the eigenvectors of $\mathbf{A}\mathbf{A}^T$, $\mathbf{V}$ is an NxN matrix containing the eigenvectors of $\mathbf{A}^T\mathbf{A}$, and $\mathbf{\Sigma}_0$ is an MxN matrix containing what are known as the singular values of $\mathbf{A}$, which lie as diagonal entries in $\mathbf{\Sigma}_0$ and are comprised of the square roots of the non-zero eigenvalues of $\mathbf{A}\mathbf{A}^T$ or $\mathbf{A}^T\mathbf{A}$. Therefore, it becomes evident that decomposing the sample covariance matrix to obtain the eigenvectors is equivalent, up to the constant factor, to finding $\mathbf{V}$. Hence, singular value decomposition also provides a method to find the principal components.

Bibliography

[1] Ayres, I. *Super Crunchers: Why Thinking-By-Numbers is the New Way To Be Smart*. Bantam: New York, 2007.

[2] Butler, D. When Google Got Flu Wrong. *Nature*, 494. doi:10.1038/494155a, 2013.

[3] Centers for Disease Control and Prevention. FluView. `http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html`.

[4] Centers for Disease Control and Prevention. Reconstruction of the 1918 Influenza Pandemic Virus. `http://www.cdc.gov/flu/about/qa/1918flupandemic.htm`, last updated Feb. 8, 2011.

[5] Centers for Disease Control and Prevention. Overview of Influenza Surveillance in the United States. `http://www.cdc.gov/flu/weekly/overview.htm`, last updated Oct. 5, 2012.

[6] Cook S., Conrad C., Fowlkes A.L., and Mohebbi M.H. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. *PLoS ONE*, 6.8: e23610. doi:10.1371/journal.pone.0023610, 2011.

[7] Davies G.R., Finch R.G. Sales of Over-the-Counter Remedies as an Early Warning System for Winter Bed Crises. *Clinical Microbiology and Infection*, 9:858–63, 2003.

[8] Duhigg, C. How Companies Learn Your Secrets. *New York Times Magazine Online*, `http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all`, Feb. 16, 2012.

[9] Espino, J., Hogan, W. and Wagner, M. A Timely Data Source for Surveillance of Influenza-Like Diseases. *AMIA: Annual Symposium Proceedings*, pages 215–19, 2003.

[10] Eysenbach, G. Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. *AMIA: Annual Symposium Proceedings*, pages 244–8, 2006.

[11] Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature*, 457:1012–4, 2009.

[12] Google Correlate. `http://www.google.com/trends/correlate/`.

[13] Google Flu Trends. `http://www.google.com/flutrend/`.

[14] Google Flu Trends. `www.google.com/trends/`.

[15] Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer: New York, 2009.

[16] HealthMap. Flu Near You: Do You Have it in You? `https://flunearyou.org/`.

[17] Hogan W.R., Tsui F.C., Ivanov O., et. al.; Indiana-Pennsylvania-Utah Collaboration. Detection of Pediatric Respiratory and Diarrheal Outbreaks from Sales of Over-the-Counter Electrolyte Products. *JAMIA*, 10:555–62, 2003.

[18] Hulth, A., Rydevik, G., and Linde, A. Web Queries as a Source for Syndromic Surveillance. *PLoS ONE*, 4.2: e4378. doi:10.1371/journal.pone.0004378, 2009.

[19] Johnson, H., et. al. Analysis of Web Access Logs for Surveillance of Influenza. *MEDINFO*, pages 1202–6, 2004.

[20] Magruder S. Evaluation of Over-the-Counter Pharmaceutical Sales as a Possible Early Warning Indicator of Human Disease, 2003.

[21] Mallows, C.L. Some Comments on C_p. *Technometrics*, 15(4):661–75, 1973.

[22] Murphy, K.P. *Machine Learning: A Probabilistic Perspective*. MIT: Cambridge, 2012.

[23] Polgreen, P. M., Chen, Y., Pennock, D. M. and Forrest, N. D. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, 47:1443–8, 2008.

[24] Welliver R.C., Cherry J.D, Boyer K.M., et. al. Sales of Nonprescription Cold Remedies: A Unique Method of Influenza Surveillance. *Pediatric Research*, 13:1015–7, 1979.

[25] World Health Organization. Influenza (Seasonal). `http://www.who.int/mediacentre/factsheets/fs211/en/index.html`, last updated Apr. 2009.