Big Data and Disease: Using Twitter to Model the 2014 Outbreak of Chikungunya Fever in Puerto Rico

Big Data and Disease:
Using Twitter to Model the 2014 Outbreak of Chikungunya Fever in Puerto Rico

Wesley King Chen

A Senior Thesis Presented to the
Department of Applied Mathematics
in Partial Fulfillment for Honors in
Applied Mathematics in Computational Biology (AB)
and Computational Science and Engineering (SM)

Harvard University
Advised by: Mauricio Santillana

April 1, 2015

Table of Contents

Abstract

Big data has enabled an entirely new approach to solving and understanding problems. With the popularity of social media, data is created by individuals. We believe that embedded in the big data of social media, like Twitter, is the documentation of self-reporting illness. Through analysis of keywords in tweets geo-tagged to Puerto Rico, we seek to model the outbreak of Chikungunya fever, with initial correlations of around 0.86. Collected tweets were then divided into categories and treated as independent variables for Lasso regression. Although we train on imperfect suspected numbers for the outbreak from the Pan-American Health Organization (PAHO), we analyze the coefficients to understand the social implications behind both social media disease reporting and awareness in Puerto Rico. We see different phases of Twitter volumes pre-, during and post-initial outbreak. News and government tweets decrease during subsequent outbreaks when we see a corresponding relative increase of self-reporting tweets. Especially when applied to epidemiology, big data isn't about finding the perfect answer, but instead, about discovering the underlying story. This thesis is about the story of a Chikungunya outbreak in Puerto Rico from the eyes of Twitter.

All code is publicly available on Github at

`https://github.com/wesleykchen/Modeling-Disease-with-Twitter.git`.

The data, as a MongoDB database, must be separately requested due to file size.

If interested, contact Wesley at wesleychen@college.harvard.edu.

Chapter 1

Background

"There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days."

— Eric Schmidt as CEO of Google, 2010 [45]

1.1 Epidemiology and Modeling

Epidemiology is the study of modeling outbreaks of disease. The most traditional approach to do so is using the SIR (Susceptible-Infected-Recovered) model, where differential equations govern the relationship between the groups — for example, the rate of infection is dependent on the infected population size [50]. For each disease, the mechanics of the outbreaks, including any action, can complicate the model, often resulting in complex multi-state models including but not limited to the addition of an "Exposed" group [50] or a "Quarantined group" [41]. In these classical disease modeling methods, the data, which usually comes from official government reports, is used to fit the extended SIR model through parameter optimization techniques and new hybrid methods inspired by those

used to predict the weather such as Kalman filters [50]. The optimized models can have certain interpretable parameters such as R_0, the reproduction number, which represents the expected number of new cases one infection will generate. When $R_0 > 1$, this means that the epidemic will continue to spread and is often the most important number to compute [41].

Unfortunately these models have weaknesses, which big data seeks to address. First, they are highly dependent on the availability of data. Usually reliant on local governments, these numbers and the delayed acquisition, vary on the region of interest and the infrastructure set up to detect and track cases. The diseases which are available to model are also those that were chosen to have data collected — usually large epidemic outbreaks or annual diseases like the flu [50]. In order to access the full data, studies are almost always performed many years after the outbreak, with goals to understand the spread of disease and to analyze the effectiveness of the actions taken. An Ebola study in Africa, published in 2007, focused on models built from two outbreaks: one in 1995 and the other in 2000, which were twelve and seven years after the outbreak, respectively [41]. As a result of this delay, predictive modeling has not been performed as much as retroactive modeling or disease monitoring (more efficiently collecting data and tracking the disease) [50].

1.2 The Onset of Big Data

Ever since the consumerization of the Internet in the 90s, data generation and storage reached new peaks. The term "big data" was first mentioned in a NASA paper in 1999 [71]; that year, 1.5 exabytes of information were created [71]. The increased power of computation and decreased cost of memory has allowed data collection to explode. By 2012, digital data creation had surpassed all predictions, reaching 2937 exabytes in 2012, nearly 2000 times more data than 13 years ago [56].

The flood of information has led to increasing interest in using data to answer all questions [29]. In 2003, the Human Genome Project was completed, sequencing the entirety of the 3 billion base pair human DNA [51]. There was then a desire to use this data in various studies to better understand genes and in turn, to control and alter them. In 2006, one of the most famous big data contests began: the Netflix Prize [49]. This contest was a million dollar prize awarded to anyone who could beat Netflix's movie prediction engine based on a training set of movies that users had watched and liked. The prize was officially awarded three years later, in 2009 [49]. Just the year before, the prestigious biological sciences journal, *Nature*, released a special report on big data, foreshadowing the deluge of data-driven studies to come [56].

Big data has revolutionized our approach to solving problems [29]. It has also dissolved the boundaries around conducting scientific research. Anyone with an Internet connection can access mounds of publicly available data. The topics of study have been broadening as well. Even the most popular video game today, *League of Legends*, has released an API allowing access to a database of all past video game matches, including end game scores and player statistics [57]. But not all data is public. What would be one of the largest sources of online data is Google's search keywords, which has still been kept proprietary, though internal analysis has been performed sometimes leading to surfacing of Google's own projects.

Data has empowered researches to discover patterns and to find conclusions that are no longer spurred by higher level ideas, but by pure exploration of the data. Of course, data can still be used in the classical paradigm of supporting claims, but in the many applications of big data from biology to culture to epidemiology, data has evolved into the central analysis — exploring first and interpreting what is found afterwards.

1.3 Google Flu Trends and Lessons Learned

The access to the Internet has ushered in a new age, where communication is centered around the web and sites related to social media, therefore, generate mass data [7]. The goal of big data in epidemiology is to use real-time data from the population via the Internet to track and predict disease outbreaks. In 2009, a study estimated that 37-52% of Americans seek health related information on the Internet through searches [7]. Google realized that they had this data set and created one of the most famous applications of big data in epidemiology: Google Flu Trends [28]. Their approach was to select a subset of search keywords that correlated with the flu, train predictors on the past five years of outbreak data between 2003 and 2008, and to now-cast the current magnitude of the new flu season [22]. The tool created by Google is still available today, [28] and the same algorithm has also been published for Dengue Fever, also released by Google [27].

The largest critique of Google Flu Trends, other than non-duplicable studies due to proprietary data, was that the model severely over-predicted the severity of the flu outbreak for subsequent years [8]. A possible reason could have been the failure to account for people querying the keywords out of interest or concern and not self-reporting. With search data, there is no way to distinguish the reason behind the search and thus they had to assume that the proportion of unrelated searches would be proportionally constant [6], which it is not. Even an update to the Google Flu Trends algorithm, which improved the model, still overshot by 30% and was not not even as powerful as using CDC physician reports which has an acceptable two-week lag [43].

One result that was undoubtedly a breakthrough. however, was the speed at which predictions could be made. Regardless of the efficiency of the local government, Google data and predictions where independent of all else and could be made without any delay [22]. Fixing big data approaches in epidemiology required more data sources, and in

particular, those that could provide a textual context around each count. And then came Twitter [6].

1.4 Twitter as a Big Data Source

Twitter is a social media website that allows users to post 140 character messages (and photos) to the world [69]. Trending topics are often denoted using the hashtag, #. The nature of Twitter is to keep the messages length brief and focus on one trending topic at a time[67]. Most people use Twitter to either share personal information and thoughts or to raise awareness about a trending topic. What is popular on Twitter represents what is popular in the community [47, 46]. Users can also create profiles for their Twitter handles which allow others to re-announce their tweets (called retweeting), and can include some metadata about the tweeter personal information such as age, gender, name and location [69].

Since its incorporation in 2007, Twitter has gradually increase in popularity to its current state of 288 million active users per month, tweeting in 33 different languages around the world. Around 500 million tweets are sent per day, fluctuating with breaking news and seasonal trends [33]. From its conception, Twitter has indexed all public tweets ever sent which is over half a trillion as of late 2014 [72]. 80% of active Twitter users are on mobile where it is easy to opt in to geo-tagging, when the phone's GPS location (or triangulated) is saved and assigned to the tweet [67]. And it's not just individuals who have Twitter accounts. Business (around 63% [15]) have Twitter handles as well. Twitter continues to be popular around the world and has had its IPO in late 2013 [13].

Twitter provides a public API to access their database of tweets with [68], with an average query time of under 100ms [72]. This API allow access not just to the text of the

tweets, but also to all available metadata including user handle, location, date, language and others. The API is only limtied by a rate ceiling but companies like Datasift [14] and GNIP [23] have sought to monetize a full collection of Twitter's data.

A quick search into Google scholar with the search phrase "Twitter API" reveals the power of the Twitter data set, with 326,000 results [26]. Most of the published work using Twitter has been about network and graph problems. Only recently, with the growth of applying big data to social problems has Twitter's valuable collection of data from the individual been tapped into [40]. Twitter data holds an advantage over other big data due to this public APIl; most notably, Google's complete search data is proprietary [6].

The nature of Twitter data also makes it unique. Some Twitter data not only represents the thoughts of individuals but also at an estimated location, through geo-tagging, allowing each idea to be placed on the map [40, 44]. The location can be determined either from an opt-in feature of GPS tagging or from a user profile which is labeled with a qualitative location, usually a greater geographic area like a city. With trained classifiers, the geographic location supplemented by tweet content can predict the home location of the tweeter [44]. For epidemiology, location data is a special piece of information that is desired because of the power it gives to mapping local outbreaks. Another advantage of Twitter data is the power of context. Not only can keyword frequencies be counted, as the case with Google searches, but connotation can also be evaluated to filter for only the most relevant tweets. For these reasons, Twitter is seen as a big data set with much potential in the public health sector [6, 17, 7].

1.5 Past Epidemiological Work using Twitter

The Twitter dataset, with both location and connotation neatly packed into 140 character snippets, is perfect for epidemiology. When looking at health-related tweets, 17.6% were about personal experience which confirms the existentialistic of self-reporting by patients [39]. Different models have been created to understand the context of tweets have been studied, particularly in the health-related sense, such as the Ailment Topic Aspect Model which was able to find a strong set of keywords of descriptions, symptoms and treatments for broad categories of public health. Optimal subsets of keywords for categories including cancer, allergies and oral health have been published [54]. `www.healthtweets.org` is s site that uses Twitter to monitor public health. Although no predictive modeling is included, statistical classifiers are used to identify health-related tweets, where they are then annotated and geo-located using a tool developed by the creators called Carmen. The site has tracked trends over billions of tweets, at a processing rate of 10 million a day [17].

The relationship between Twitter and social behavior has been studied recently as well. Twitter activity related to disease outbreaks has been shown to go through different phases [36]. The volume of tweets is initially very high when news and public service awareness tweets are trying to get information out. When news of the outbreak gets older, this volume decreases and we observe a corresponding increase of self-reporting as the presence of the disease becomes common knowledge. In this study, we clearly see Twitter phases which we will further analyze in our results (see Section 3.4) The content of tweets has also been looked at. Social media often promotes false information, which is no different for medical knowledge. A study on the quality of information about Ebola on Twitter in Nigeria, revealed that almost 60% of information on the disease was inaccurate, and often retweeted uncorrected [52]. This misinformation can often lead to swellings of twitter volume as people react to a fear or concern.

The gold mine of Twitter data has resulted in many epidemiological studies most of which a shared similar overall approach. The first step is to use some type of classification system to find relevant tweets either from sentiment analysis, statistical classifiers or other natural language processing techniques. Then the analysis focuses around establishing a correlation from tweet frequency to the gold standard sometimes using clustering [24], naive Bayesian models [3] or support vector machine (SVM) techniques [63, 24, 3]. Occasionally, regression models are built for some predictive modeling but only for larger data sets, usually related to the flu [2].

But there are always nuances. Even for self-reporting, understanding content may require advanced linguistics to classify tweets [37]. The tenses of the verbs, as well as the subject/predicate relationships are easy for humans to distinguish but require more training and clever methods for automation. Methods usually train on labeled data from humans, with sample sizes on the order of 10,000 to classify a two or three trends over time [37].

For these reasons, although Twitter data is massive, there are still limitations surrounding the availability of data for certain geographic resolutions once multiple filters for keywords are applied. It is still easiest to model and track diseases that are the most common, like the flu, even on Twitter [54].

Flu predictions in the United States using Twitter data have shown positive results. Some of this can be attributed to the large amount of data available — as the flu is an annual epidemic that folllow similar seasonal trends. With a data set of 280,000 across the United States over past years, studies have achieved 80% F1-recall accuracy [3]. Another study used Twitter data as a correction factor to models built from other data sources to predict flu levels and reported that the error rate was reduced by 17-30% when Twitter data was incorporated. The most similar approach to the one presented in this paper, with respect to the treatment of processing of Twitter data, was a 2014 paper tracking the

2012-2013 flu outbreak in New York City. Geo-coded tweets were used with 3000 tweets collected on a keyword tally. These authors were able to reach a correlation of $R = 0.763$ which beat out the $R = 0.683$ of Google Search Query trends [48].

There have been a handful of studies using Twitter for non-flu outbreaks. As mentioned before, a Twitter study about Ebola aimed to find how much misinformation existed about the disease [52]. A geographic and time-based study on Dengue Fever in Brazil was also conducted and seen success; Brazil is a country with extremely high Twitter usage so this was a natural location to extend non-flu studies to [24]. Only recently has current work been done towards tracking and predicting lesser-known diseases in more localized regions, where Twitter usage is not exceptionally high, but still prevalent enough to warrant study.

1.6 Big Data and Epidemiology Today

Big data in epidemiology has continually grown as a field and studies are being conducted for various diseases around the world utilizing the spectrum of data available. The field has expanded in such a way that in 2014, the first IEEE conference was held on the exact topic of big data in computational epidemiology [32].

Other than social media, a variety of sites and mobile apps have sought to give access to cleaner data from the population by creating self-reporting portals. Rather than gleam self-reporting numbers from Twitter, sites like `www.sickweather.com` [62] and `www.flunearyou.com` [19] collect voluntary reports of various sicknesses. This high quality data comes at the cost of being low in volume, with SickWeather having only 8 cases of being "sick" over a 2 week period in all of Boston [62] at a randomly selected time. Even with cleaner data in terms of context, predictive modeling may not be the sole purpose of these sites; rather, they wish simply to be as local notifiers of relative intensity,

not real predictors of numbers computed from a model [19]. SickWeather, however, does have its own modeling methods, claiming to search through social media and even filter for context, but the method is not revealed, merely stating that it is "patent-pending".

The greatest drawback of these sites is that the data is collected either too generic (with keywords like "sick" or "fever") which could be any disease or too specific, built only for the flu, which has quality data already. The combination of different data sources, however, is presented by another team running a webiste used to track all diseases — `www.healthmap.org` [**?**]. The different diseases can have varying prediction methods, with the FluCast [31] being the most developed, but the site contains automated real-time predictions and visualizations for all. FluCast trains separate classifiers on the different sources of flu data (including Google Flu Trends, Google Search Queries, AthenaHealth and FluNearYou) and combines them to form one predictive model in an ensemble mixture-of-experts approach. It is able to outperform any of the data signals alone but is still only as powerful as the availability of data, currently tied to the flu [31].

However, for diseases not as popular as influenza and in more specific regions, the effect of big data has not sped up modeling yet. Predictive modeling the outbreak of Chikungunya fever in Puerto Rico or even the greater region, at the time of this thesis, has not been published despite the first large outbreak occurring half a year ago in July 2014. The CDC released a report in December about monitoring the disease and reporting the outbreak levels from May to August [61], but that there were no predictive elements. Another paper in Spanish analyzed lessons learned from the handling of the outbreak in the Dominican Republic, which although forward-looking, still did not given any measure of predictive capability [55]. We believe there is still ground to be broken, not just for Chikungunya in Puerto Rico, but also in developing a pipeline that can be generalized to any disease or region (see Section 4.1.3).

1.7 Puerto Rico, the Internet and Twitter

One of the requirements of any social media study for a region is Internet penetration, which is how prevalent access to the Internet is. Increased penetration leads to greater usage of social media, which is the source of epidemiological big data studies [32, 7]. Puerto Rico has a relatively high Internet penetration compared to its Caribbean neighbors like the Dominican Republic [25]. Google information citing the World bank claims a penetration of 73.9% in 2013 [25]. Other sources report a percentage of 57 % for those 12 and older, which represents close to 2 million Internet denizens [35, 34]. With regard to social media in particular, even in 2013 when a study was done, 89% of Puerto Rican Internet surfers connected to social media sites [9, 34], with Facebook being the number one site at almost 79% and Twitter being the 4th most popular (behind Facebook, Instagram and Google+) at 23.1%. Even at this percentage, Puerto Rico would have 400,000 Twitter users. Since then, Puerto Rico Internet usage has been projected to increase [9] and grow more mobile [9, 35], which will increase geo-tagging percentage. For 2014, we collected over 10 million geo-coded tweets with location coordinates in Puerto Rico, see Section 2.2.

1.8 Chikungunya Fever

Chikungunya fever is caused from a virus that is spread to humans from the bite of an infected female *Aedes aegypti* or *Aedes albopictus* mosquito [59, 10, 70, 58]. The first reported case was during 1952 in southern Tanzania [70]. Due to the same mosquito vector and similar symptoms /citeWHOchik, Chikungunya can be confused with dengue fever — with hepatomegaly, the swelling of the liver, being a key differentiating factor between Chikungunya and classical dengue fever [58].

A unique factor that is important to bear in mind when trying to control outbreaks

is that a human with Chikungunya fever can "back-infect" other mosquitoes who bite the human host [10]. Diagnosis of the disease involves blood tests for specific antibodies [70]. The incubation period between the infected bite and expressing symptoms can be anywhere from 2 to 12 days [70].

The primary symptoms are an abrupt fever coupled with debilitating arthralgia, or joint pain. The severe joint pain experienced has led to the naming of the disease; "Chikungunya" is a Kimakonde word meaning "to become contorted" as a result of the body pains [70]. Other symptoms that are reported include muscle aches, nausea, rashes, fatigue and headache [10, 58, 70, 59] Fortunately, the disease is not lethal, barring co-morbid complications, but can lead to chronic arthralgia [70].

Currently, there are no vaccines to prevent Chikungunya fever [10, 58, 70, 59] and no antiviral drugs [70, 59]. Any medicine prescribed is for alleviating symptoms and ensuring plenty of fluids and rest is the only medical suggestion to all patients once infected [10, 58]. Prevention is the same for preventing mosquito bites and other mosquito vector disease: bug-repellent, long clothing and being wary during active mosquito hours [70]. Once infected with Chikungunya, however, the person is likely to develop an immunity for the future [10].

Chikungunya has been reported in the Americas, Asia, Africa and Europe in over 60 countries [70, 59, 58]. In the Americas, it was first reported in the Caribbean region in December of 2013 but the largest outbreaks occurred in July, with subsequent outbreaks in October and late November [10]. As of February 27th, the Pan-American Health Organization (PAHO) has tallied 1.2 million reported cases [10]. On the Center for Disease Control (CDC) website, there are "nowcasts" that are summaries of the reported cases per moth across Central and South America though on the scale of 20 dots per month for a visualization encompassing the entire region. The CDC does its best to track outbreaks of Chikungunya as immediately as possible but most studies of outbreaks are retroactive

by at least a few months. A report on the first outbreak in July was released four months after the relative end of the outbreak, after another outbreak had already occured [61]. Speeding up this process to be in real time is one of the primary goals of our exploration using Twitter data.

1.9 Lasso Selection and Least Angle Regression

The Lasso, or Least Absolute Shrinkage and Selection Operator, is a selection method for linear regressions [30, 65], first proposed by Robert Tibshirani of Stanford in 1996. [66]. Lasso fits the same linear model (Eq. 1.1) as more well-known methods like ordinary least squares (OLS) [65], where once the coefficients, b_is are obtained, they can be used to fit and predict predict out of sample data with a new set independent variables.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n \tag{1.1}$$

The development of the Lasso method was to improve on OLS estimates in two ways: prediction accuracy (minimizing variance on estimates) and interpretation (determine smaller subsets of important variables), both of which are important in epidemiological applications [66]. Previously, there have been two common improvements for OLS estimation: subset selection and ridge regression. Subset selection allows coefficients to be set to 0 which effectively removes them from the model allowing for easier interpretation when there is a large set of possibly explanatory variables. However, small changes in the data can result in vastly different subsets [66]. Ridge regression continuously shrinks coefficients to reduce their importance in the regression model, which is more numerically stable than subsetting, but the coefficients can never reach 0 to be eliminated from the model. Lasso seeks to combine these two approaches (hence the name) through changing

the objective function which is minimized to using the L1-norm as the penalizing function with a threshold tuning parameter, see Algorithm **??**.

Algorithm 1: The Lasso Problem Formation [66]

Input: (x^i, y_i) where $i = 1, 2, \ldots, N$ and $x^i = (x_{i1}, \ldots, x_{ip})^T$ standardized to $\mu = 0, \sigma^2 = 0$

Input: tuning parameter $t \geq 0$

Let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)^T$

Let Lasso estimate be, $(\hat{\alpha}, \hat{\beta})$, where:

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin} \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \text{ under the constraint } \sum_j |\beta_j| \leq t$$

The tuning parameter, t, controls the amount of shrinkage. The amount of shrinkage can be approximately quantified. If $t = t_0/2$, where $t_0 = \Sigma|\hat{\beta}_j^\circ|$, the result will be similar to finding the optimal subset of size $p/2$ [66]. In general, when $t < t_0$, more coefficients will be shrunk towards 0 and possibly being set to exactly 0 [66].

Solving for Lasso becomes a quadratic programming problem with linear constraints [30, 66], so despite a final linear model, solving the lasso occurs through non-linear paths [20]. Often used is the "shooting algorithm" [21] which is a type of coordinate descent optimization, proposed two years after Lasso which solved the Lasso with multiple parameters. The regularization path for Lasso is computed via a slight modification on LARS (Least Angle Regression and Shrinkage) [18], which is a model selection algorithm that is more efficient than traditional forward selection algorithms [18]. We summarize the forward selection algorithm, in Algorithm 2 and compare it to the LARS algorithm, Algorithm 3.

Algorithm 2: Forward stepwise regression [65]

Input: coefficients $b_i = 0, \forall i$

Start with empty model

foreach predictor not in model **do**

    Add predictor x_i most correlated to y to model

    Compute residual:

        $r = y - \hat{y}$

Algorithm 3: Least angle regression [65]

Input: coefficients $b_i = 0, \forall i$

Start with predictor x_i most correlated to y and empty model
Compute residual:
  $r = y - \hat{y}$
foreach predictor not in model **do**
  while $\text{corr}(x_i, r) < \text{corr}(x_j, r)$ where $x_i \in$ model and x_j is any predictor $\notin$ model **do**
    Increase coefficients of $b_i \forall i \in$ model in the direction of joint least squares
  Add equallly-correlated predictor to model
  Update residual:
    $r = y - \hat{y}$

The extension of the least angle regression algorithm to compute the entire path of lasso solutions can be achieved by removing all zero-valued coefficients from the set of predictors and then recomputing the joint least squares direction [65], as $t = [0, \text{inf}]$ Since its development, Lasso variants have been published for adaptive quantile estimators and to model error when moments of the error are unknown. Work has also been done to optimize the process of solving the quadratic programming problem [12].

Chapter 2

Methodology

"Hiding within those mounds of data is knowledge that could change the life of a patient, or change the world."

— Atul Butte, Stanford School of Medicine [60]

2.1 Twitter Database

The research group under John Brownstein has been routinely pulling tweets using the public Twitter API [68] to maintain a database of historic geo-tagged tweets. The tweets are stored into a MongoDB database and hosted online. Due to the large amount of metadata available per tweet, our working database was built selecting only certain fields of interest and then including a unique ID to another database where the entire metadata directly from the Twitter query is stored. This processing step selects for the essentials like user id, tweet content, tweet date, and accessdate, but also saved fields that were not necessarily in all tweets, such as our geo-coding profile vs. GPS numbers, language of tweet, etc. The database is being maintained by Clark Freifeld of the Brownstein lab, so that we can rerun

the analysis on up-to-date data by rerunning our pipeline. Their are different clones of the database which are optimized for different operations — and for our read-only needs, there was one such database set up to increase query performance.

The REST API limit allow access of around 1% of the total daily tweets — amounting to 5 million a day [67]. There are other sources such as GNIP[23] and DataSift[14] that allow for almost full volume (at a price) but we have not found this necessary — the extent of Twitter data already available for us is enough for large enough sample sizes of our refined queries. Even with our API query limit, we are able to saturate our daily quota with geo-tagged tweets (so we believe that the true percentage of geo-coded tweets may be somewhere from 2-5% of all the tweets 1.4. The selection process of only geo-tagged tweets comes from querying for the existence of the longitude and latitude fields (between [-180,180] and [-90, 90]) to capture all tagged tweets. The geographic information can exist in two forms: either as a profile location (where we translate the given geographical name like a city into a bounding box and then save the midpoints) or even more simply, a GPS reading at the location the tweet was made. We stored both types of geo-information in our database into separate fields to allow for differentiation. The profile location will thus be a looser bound but since they are saved differently, during analysis, this data can be processed. As will be discussed later in Section 3.1, 70-80% of the Puerto Rico tweets had both profile and tweet locations and were consistent.

2.2 Removing the "Big" from Big Data

Filtering the complete Twitter database was the first step. We wanted to pair down our data in two separate processes — the first would be a coarse filter, where we would run an expensive operation on the entire database once in our analysis and the second would be filtering the working database, which could be further queried as analysis decisions

were made. Our coarse filter used `mongodump` to export the database to one that could be restored locally, with the below constraints:

1. Tweet Creation Date (field `cr`) $> 1/1/2014$
2. Profile Geo-Tag (fields `plt` and `pln`) or Tweet Geo-Tag (fields `tlt` and `tln`) are within the bounding box including Puerto Rico (Latitude $[17.5°N, 18.7°N]$, Longitude: $[65°W, 68°W]$

After the coarse filter, we kept 14430635 tweets (via `db.collection.find().count()`) — representing the total number of geocoded tweets from January 1st 2014 to January 15th, 2015 (the date of the most recent query).

The fine-grained filters applied varied slightly depending on research question, but the primary operations revolved around using PyMongo and other Python libraries to output datafiles of regex keyword queries fed to the `db.find()`. Our Regex search for the `t`, tweet content, field of our database was of the format "`/\b.*KEYWORD.*\b/i`" where keyword is from the below Figure 2.1 of keywords of interest, using the Python `re` library. This query, for those unfamiliar with regular expressions, will match any text between spaces that contains our keyword, case-insensitive. This means a keyword "Chikungunya" will be able to match even more complex hashtags sometimes found in tweets like "#hateCHIKUNGUNYA!!"

Category	**Keyword**
Chikungunya Related	"Chikungunya"
	"Chikv"
Symptoms Related	"rash"
	"high fever"
	"joint pain"
	"nausea"
	"vomit"
	"photophobia"
	"arthralgia"
Other Control	"Dengue"
	"flu"
	"sick"
	"cough"

Figure 2.1: List of Keywords queried against Puerto Rico geo-coded tweets by Category

The queries were grouped approximately by week (with deviations to match the inconsistencies in our gold standard, see Section 2.3) and counts (using the PyMongo function `db.collection.count()`) were displayed in a comma-separated value datafile.

We considered allowing for misspellings (through flexibility in our regular expression) but not only would wildcards drastically increase query time (evening adding the wild card repeats right before the keyword resulted in a 10-20x slow down, but we observed that this consideration was not necessary. After reviewing our results on the keyword searches, including the fact that CHIKV, the abbreviation of the Chikgunya virus was never mentioned and that adding the wild card repeat only altered the result by around 1% we believe that less than 1% of all mentions of Chikungunya were spelled incorrectly

on Twitter and decided to ignore misspellings in favor of computation time.

2.3 Our "Golden Standard"

When evaluating any predictive or informative modeling, it is necessary to have a gold standard. In our case, we needed to find a published report of the number of Chikungunya cases in Puerto Rico. The World Health Organization (WHO) and the Pan-American Health Organization (PAHO) publish nearly weekly PDFs that summarize the cases of Chikungunya in various countries in the Americas — of which Puerto Rico is independently tallied. The PAHO data starts as early as the first full week of 2014 [53]. For the purposes of comparisons, our Twitter data models match the same time periods as the releases of the PAHO Chikungunya in the Americans reports. Initially, when the disease is first starting to be diagnosed in the country, PAHO will only report the number of confirmed cases. As the disease continues to spread, however, PAHO will model a "suspected" case number. It is this number that we believe will be a better gold standard for our study since Twitter, by utilizing self-reporting, should be able to track the actual number of cases and not just those that go to hospitals. For Chikungunya in Puerto Rico, the first case was confirmed the week of May 25 but it wasn't until several weeks after until other cases were confirmed before PAHO started to publish a suspected cases number on their weekly reports (June 22 was the first week of reported suspected cases). There is also an imported case number that is tracked for cases that come to the island via travel but this number was one order of magnitude lower and very irregular, so we chose to ignore it.

It is important to note that all of the reported numbers are cumulative. To compute the weekly numbers, pairwise differences were taken from week to week. Corrections are also sometimes made into the data and unfortunately, do not get highlighted and are instead observed as artifacts in the data, sometimes resulting in negative new cases in a week.

2.4 Exploring the Dataset

Exploring the data was an iterative process as we did not know what to expect. The process was initiated with mockup graphing done in the Excel workspace. Excel made it very easy to setup different sheets for slight variations of data, graph results of multiple methods and compare them, all in one workbook. The majority of the initial work was to consolidate our counts into the same time intervals as in our gold standard, and then correlate the counts for each keyword to the standard. Graphs were also created to help see relative movement and predictions of peaks, which is often a feature that epidemiologists are interested in [41].

We also experimented with normalization at this phase. When graphing, the relative movements are important and different normalization techniques were used including scaling from 0 to 1, computing z-scores for each time series and scaling by weekly tweet volume. Each normalization method was then rationalized. Scaling from 0 to 1 by dividing by the maximum would be model the shape of the disease data — the outbreak is always compared to a peak value for the year. The z-scores would help us detect numbers that deviated and were very surprising, possibly helping identify periods of outbreak. Normalizing by weekly tweet volume would help remove bias of weeks simply having more tweets and season increased tweeting trends, but may not always be appropriate as the changes in volume could in fact be due to self-reporting which we are trying to track. All of these normalized numbers had different interpretations and could all be used as different sets of independent variables for our regression and at this phase, were all computed and kept.

Regarding the keywords of symptoms, not specific to Chikungunya (like cough and fever), the plan was to eliminate a mean background which would be computed from the counts before the outbreaks. But after observing that the symptoms were so rarely

tweeted, we chose to ignore the keywords and only work with the principle keyword of "Chikungunya".

To answer the question of geographic location, we D3.js with a TopoJSON representation (from GDAL [1]) of Puerto Rico to visualize where the tweets were occurring. The setup for the map was inspired by a D3 work [5] and adapted to allow for interactivity with tweet visualizations. The direct longitude and latitude coordinates were plotted on top of the map of Puerto Rico along with variable slider bars done in jQuery to control the dates. Upon user input, the changed date would repopulate the points on the map but only those that fit the time window. This visualization helped us confirm where the tweets were coming from and was a useful tool to visualize the scale of the outbreak around Puerto Rico.

2.5 Tweet Curation

After selecting for keywords, we wanted to understand the connotation behind the usage of each keyword. We wanted to "curate", or categorize, each tweet into different categories for context. For this project, we dumped the content field of each tweet that we queried into a datafile that was sent to curators (see Acknowledgments 4.2). For continued analysis of new data as we continue to monitor the outbreak from the eyes of Twitter, we now utilize a previously designed web interface , designed by Jared Hawkins for future curations (see Acknowledgements 4.2). The curators were bilingual as the Tweets were 58%/42% for English/Spanish by accessing the `lang` field. If additional languages or work hours were required, Amazon's Mechanical Turk has also previously been used by the Brownstein group [48]. Initially, some basic categorizations were offered to the curators along with some flexibility to create new groups. The two curators' groupings were then evaluated and cross-validated. After understanding the content of the tweets, larger bins

were grouped to be used by Lasso as the independent variables. We also kept the more specific breakdown to help understand the breakdown of the otherwise larger "Other" Categories. The discussion of some observed categories is found in Section 3.7, but our final groupings are seen in Figure 2.2 below.

Category	**Description**
Reporting for Self	Being sick or showing symptoms
Reporting for Others	Others being sick or showing symptoms
Government/Educational	Any news from official sites (often hyperlinked)
Other Awareness	Including the prevention, the song or general public awareness
Other Irrelevant/Unsure	Including jokes/threats or other non-awareness mentions

Figure 2.2: Finalized groupings for curation with descriptions for curators, created after the initial curation to understand what content was in the tweets

This initial process was not automated as the data set turned out to be on the order of just over 1400 tweets. The automation process would require large enough data sets for each category to allow for distinctions to be made using automated language classification called Natural Language Processing (NLP). The training sets would need to be built for each language as well — requiring a lot of data. For this exploratory project, we did not have enough data to run a full natural language processing training and decided that the level of accuracy and understanding granted from human curation, at least at first, would be beneficial. The NLP automation, however, is a definite extension, see Section 4.1.1.

2.6 Lasso Regression

The Lasso regression (introduced in Section 1.9) was run using scikitlearn's `sklearn.linear_model.lassocv()` with automatically computed α along each step.

Given no prior information for our Lasso model, we did not chose to set a list of αs. The cross validation was computed over $n = 3$, which is around 10-20% of our full data set, for smaller runs and up to ten-fold, $n = 10$, if allowed by the data (must have at least that many points of data to cross-validate) using `sklearn.cross_validation.KFold()`. The trained classifier was then use to fit data given another set of independent variables and that prediction was stored and computed to the outfile. The decisions of how Lasso was used to train and predict will be discussed in 3.8.2. In general, to allow for the automatic computation of α, we did not train on any null data, meaning rows where we had no signal in our independent variables (our tweet counts) and this allowed for numeric stability of Lasso.

Chapter 3

Results and Discussion

"You can have data without information, but you cannot have information without data"

— Daniel Keyes Moran [45][60][16]

3.1 Evaluating Twitter as a Dataset for Puerto Rico

The process of exploring our data set first required analysis on how well our data set corresponded to our assumptions that Twitter is pervasive enough in Puerto Rico to be used as a signal. We looked at weekly Twitter volume for geo-coded tweets either by profile or by tweet location tags. From Figure 3.1, we see the annual trend of volume for 2014. We highlight a couple particularly interesting features. First, we see that there appears a relatively stable volume of around 100,000 tweets per week until the first week of April. There then seems to be a sudden spike in volume to over 250,000 for 15 or so weeks afterward. We hypothesize that this may have been Twitter API related, possibly an API rate limit increase or a change in the geo-tagging indications, discussed shortly. We do expect winter months to have less volume due to greater activity, travel and tweets in

the warmer months but the sharp corner and sudden spike was not expected. We see a gradual peak during July and August which coincides with the warmest months of travel as well as the World Cup of 2014, a very frequently tweeted topic [69]. After the summer, the volumes decreased to around 13 million, with the exception of the week of 11/16/2014 which had an abnormally low 43,000 tweets. This error is probably due to tweet retrieval.

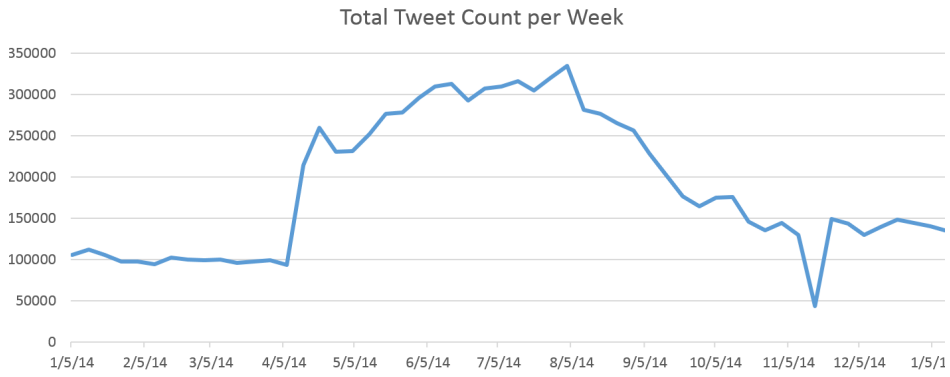


Figure 3.1: Weekly totals number of geo-coded tweets in Puerto Rico

A possible discrepancy when working with our Twitter data is a change in the the way the API stored different methods of geo-tagging. Twitter can keep track of two types: profile-specific, which is tied to the user, and tweet-specific locations, the location recorded at the time of the tweet. There seemed to have been a change in the way these methods were being stored on Twitter's end. Our filtering accepted both profiles and tweet locations as location information was what we wanted. In most cases, tweets had both tweet and profile locations but there were some tweets with only one type of geo-coding. We wanted to see if this would affect any analysis by trying to see how well the total number of tweets with each geo-tagging method correlated. In Figure 3.2, we see that our regression slope is 0.93 indicating that indeed most tweets which have geo-tagging have both. We see that with the exception of a cluster around the $x = 100000$ where several weeks have more tweet-location tweets than profile location tweets, we notice that profile-based locations were slightly more common which is as expected since profiles are usually established by

users and the tweet location is an opt-in feature requiring GPS or triangulation of cellular signal.

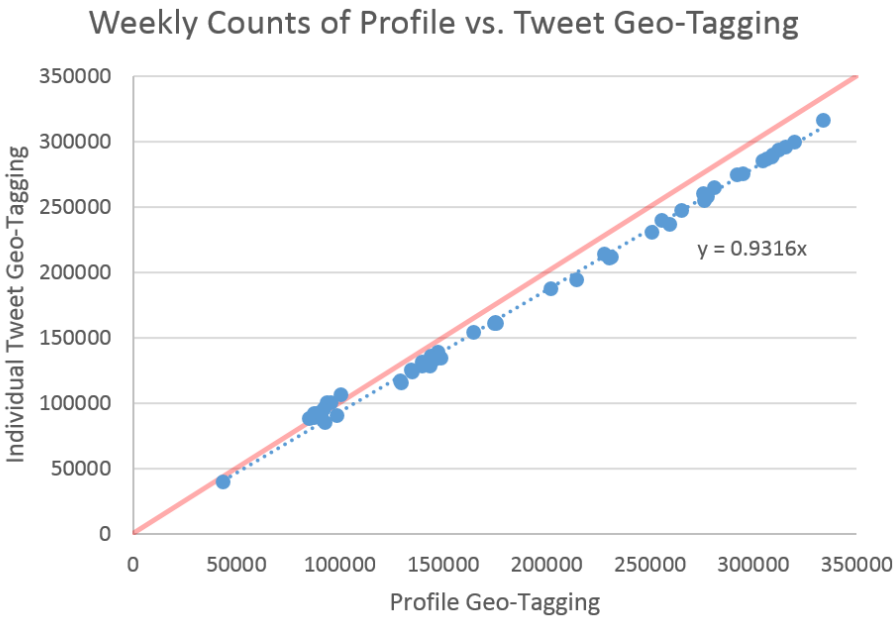


Figure 3.2: Scatter plot of tweet volume with profile geo-tagging (X-axis) vs. tweet location geo-tagging (Y-axis), the majority of geo-tagged tweets have both

But to further investigate, we wanted to see if the unique tweet-location only tweets were during the same time interval so we ran a time series of the volume of tweets that are only marked by one type of geo-tagging, as seen in Figure 3.3. We notice that there is a sudden change from being tweet-location favored to profile-location favored right at April 4, which also corresponds to the sharp increase of the weekly volumes of geo-coded tweets we saw in Figure 3.1, further confirming a possible change in our Twitter data as extracted from the public source.

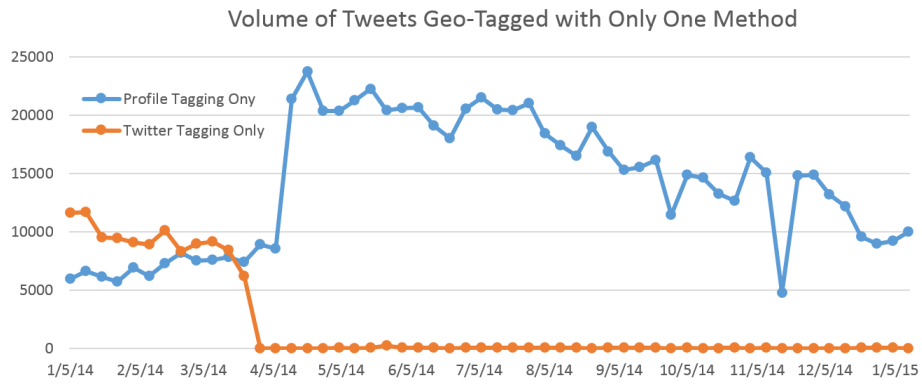


Figure 3.3: Weekly totals number of geo-coded tweets in Puerto Rico

The bottom line is that our Twitter data is still good for our analysis. With respect to volume, the observed low signal occurs at the beginning of 2014, which is outside the range of the Chikungunya outbreak (starting in late May to June). The only volume error inside our time period of interest is the aforementioned week of 11/16/2014. In the past year, there seems to have been changes in the way geo-tagging is stored in the Twitter API which is unfortunately out of our control, though we have seen that the two types of geo-tagging are generally consistent. However, we do not believe that this will affect our results much. We acknolwedge, one possible issue, or lying, where profile based locations are not as accurate and can be made up. Yet the analysis in this thesis focuses on the entire region of Puerto Rico which profiles will generally be accurate. The nature of our study, reliant on self-reporting, inherently trusts the average user. It is possible that we will miss some tweets where the user has made up an imaginary location (flast negative), but fewer people are expected to make up a profile claiming to be in Puerto Rico when not (false positive). Our curation step can also aid us in filtering away the noisy tweets.

3.2 The Not-So Golden Standard

When translating the weekly outbreak PDFs from the PAHO website [**?**] into a workable data format, we noticed many inconsistencies with the reported cases. Throughout the year of outbreak data, the format of outbreaks changed leading us to believe that the method of documenting cases has differed. Not only was the format different, but with the changes in format sometimes included lapses in data, when one report would suddenly miss data despite being in the middle of an outbreak. Because all of the reported numbers were cumulative, most of these lapses were very easy to detect (where two consecutive weeks had the same numbers). Sometimes, the confirmed number of cases decreased, which is a result of PAHO making retroactive corrections to past data but since this correction is not highlighted and previous reports not correct, when calculated weekly new cases, there are sometimes negative numbers.

Most notable was the failure of publications during certain weeks towards the end of the year. The missing or inconsistent weeks of data were the weeks beginning on: November 16, November 23, December 14 and December 28. To correct for this, since the reported numbers were cumulative, the difference between reports was still computed and assigned to a larger interval (up to 2-3 weeks combined). Since we are unsure about the distribution between weeks, we made no assumptions in trying to interpolate weekly numbers from this data. Therefore, not all of time intervals are even from points after November but for the purposes of this thesis, when we work with the data, we will still call it weekly despite the aggregations used.

Finally, the method of he suspected numbers that we are focusing on seem to be directly related to the confirmed case number. However, the method of computing the suspected number is not revealed. Due to the changes and sudden changes week to week, we do not suspect an SIR model to be used. PAHO's calculation of a suspected number adds one

more layer of complexity and error.

Unfortunately our golden standard clearly has faults, as often is the case with epidemiological data, but this data may be the best available for the time. We will continue to use it as a reference and to train our models when a target is needed but we do keep in mind the inaccuracies in the PAHO data and also seek to make qualitative analyses about the social behaviors of Puerto Ricans responding to the outbreak as well.

3.3 Insights from the Dataset

After the usage of Twitter was validated, we proceeded down the pipeline to filter for keywords using our regular expressions. We recovered weekly counts of the filtered tweets and plotted this against the background volume to see if at least our most likely keywords were significantly different from varying due to Twitter volume. From Figure 3.4, we can clearly see how there is no signal for much of the year until the out break happens, upon which there are clear spikes indicating significant activity in response to the outbreak going on.

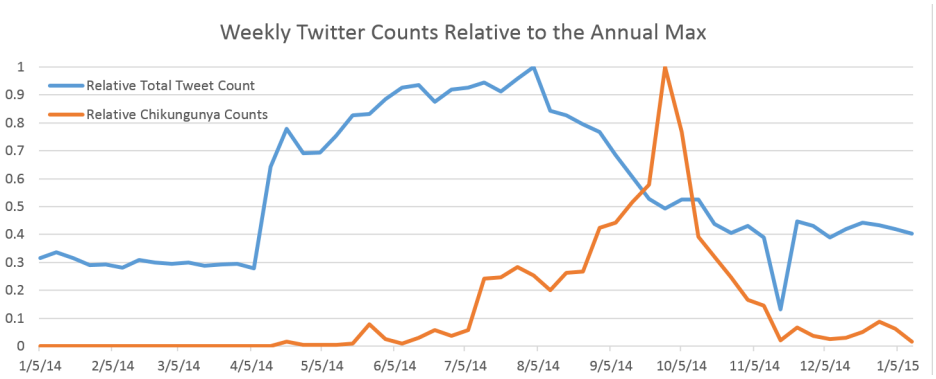


Figure 3.4: Relative weekly proportions, relative to maximum of the year, of the Chikungunya twitter signal and the total Twitter volume

From looking at our Twitter frequencies, we observed that all keywords tested in Figure 2.1, other than "Chikungunya" itself, did not have significant signals. In fact, the other keywords had surprisingly low signals, even our controls searching for other common diseases highlighting how alarming the Chikungunya outbreak was. We summarize the total and maximum weekly counts for our keywords below in Figure 3.5. There were nearly as many mentions of "Chikungunya" as "sick". A generic keyword like "sick", as expected, showed a very even signal throughout the year, unlike the outbreaking Chikungunya. Other diseases like Dengue and symptoms like rash, were an order of magnitude lower. Interestingly, we point out that the peak of "Dengue" hits was the same as the peak week of "Chikungunya" which shows how Chikungunya fever is often confused with the more well-known Dengue fever. The people of Twitter have shown this confusion as well.

Category	**Keyword**	**Total**	**Most in a Week**
Chikungunya Related	"Chikungunya"	1449	194
	"Chikv"	1	1
Symptoms Related	"rash"	317	15
	"high fever"	1	1
	"joint pain"	1	1
	"nausea"	51	4
	"vomit"	131	8
	"photophobia"	0	0
	"arthralgia"	0	0
Other Control	"Dengue"	240	28
	"flu"	86	7
	"sick"	1983	77
	"cough"	302	16

Figure 3.5: Total tweets and max hits per week of keywords, same ones as Figure 2.1

We then computed the correlation of our "Chikungunya" keyword twitter signal with the suspected number of cases in Figure 3.6. The correlation is moderate at 0.77.

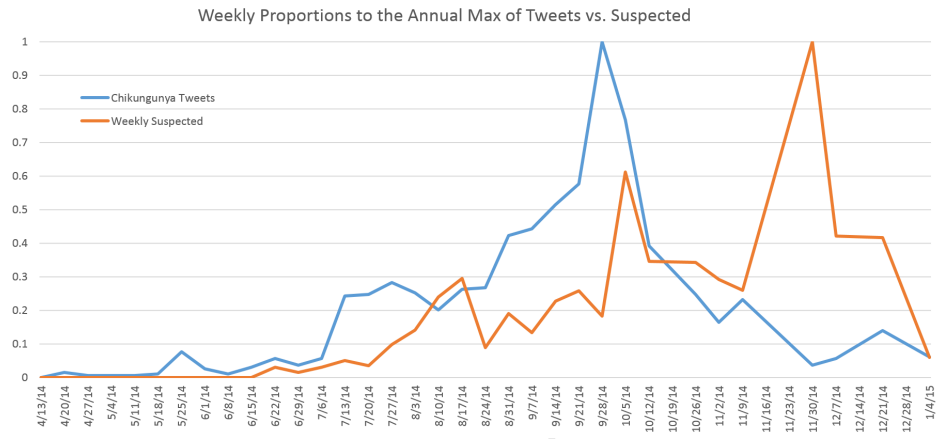


Figure 3.6: Relative weekly proportions, relative to the annual maximum, of the Chikungunya Twitter signal and the the suspected numbers of cases; notice the offset of the peaks in September and November

We believe that the social media scene should more accurately model the number of people who are actually sick, which PAHO can only guess with its suspected number. This is in contrast to the confirmed cases number which is from a count of the patients in hospitals, which PAHO can know with greater certainty. We show that these two numbers do not actually correlate well in Figure 3.7. The confirmed case number fluctuates greatly and appears to have an upper limit, which would be the number of patients seen able to be seen by doctors during the week.

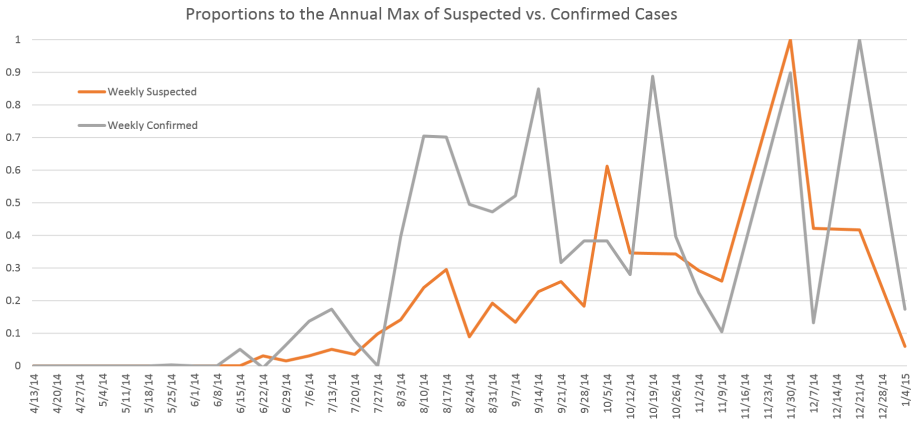


Figure 3.7: Relative weekly proportions, relative to the annual maximum, of the suspected vs. confirmed cases in the PAHO data

We observed that two very obvious peaks in both graphs, one in late September and the other in late November but offset. Upon closer inspection, the November offset which seemed larger was actually just one time interva's worth of data, just a larger time interval which was an artifact of the aggregation process to deal account for the missing weekly report, see Section 3.2. A reasonable explanation for this one week lag could be because it took some time for people to post about illness before actually going to the hospitals to get counted (and in turn increase suspected cases). This lag, set at one week, would seem to help our correlation, as seen in Figure 3.8.

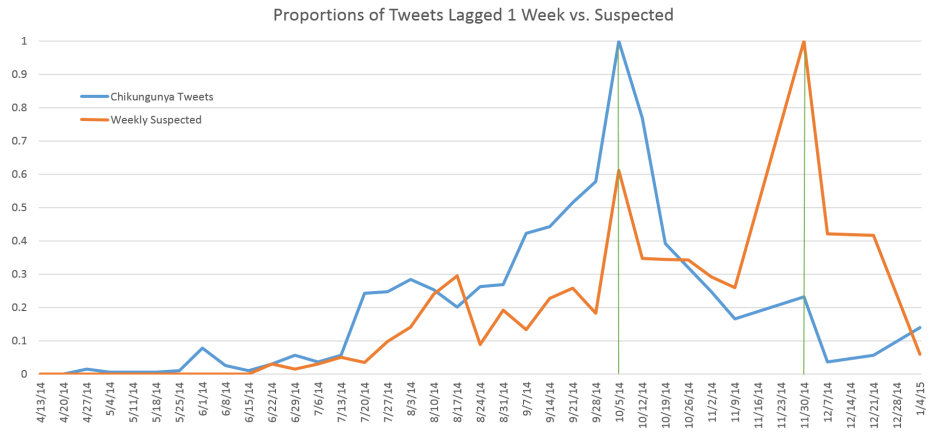


Figure 3.8: same as Figure 3.6, but with the Chikungunya Twitter signal lagged by 1 week to simulate actual time delay; the green lines show alignment of the peaks

After applying the one week lag on the Twitter data, we see a much better consistency with the peaks between Twitter and the suspected case outbreaks, with a computed correlation of 0.86, which is higher than the correlation between the suspected and confirmed cases as reported by PAHO.

In addition, we wanted to ensure that tweets were coming from independent sources, not the same few people tweeting repeatedly throughout the week. We counted the number of unique users and plotted the scaled proportions to the maximum per week against the number of tweets mentioning "Chikungunya" in Figure 3.9. We see that the shape is extremely similar with a correlation of 0.985 so we do not believe the number of unique users will bias our tallies.

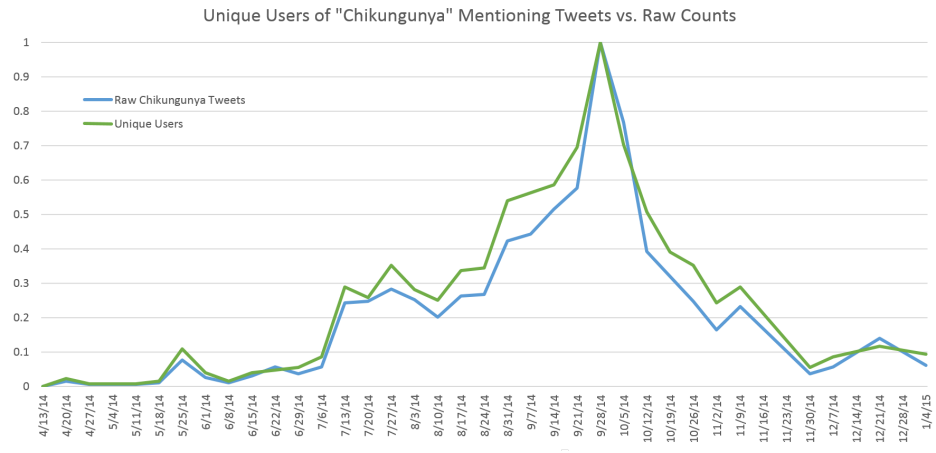


Figure 3.9: Weekly proportions of unique users mentioning "Chikungunya" vs raw number of tweets

3.4 The Phases of Twitter

In analyzing the new lagged time series (Figure **??**, we have detected features of our graph that may be explainable by social behaviors, highlighted in Figure 3.10. The first phase is called the "Onset Phase", where the medical community has reported enough cases from the local population that the government and/or PAHO are keeping watch on the disease in case it becomes an epidemic. The first relative maximum occurs during this phase. The volume of tweets will then gradually increase as the disease becomes more and more known, and we see moderate levels of disease — enough to warrant precaution from the health agencies. The second phase is the "Crisis Phase". Here we see an increasing amount of suspected cases with a corresponding swell of related Twitter activity in what is actually a second outbreak and larger outbreak. In the case of Chikungunya, a disease not as well-known as some of its counterparts like Dengue, the fear of an outbreak and the announcement by various news sources will amplify the signal to abnormal highs. The third phase is the "Post-Crisis Phase", where although the disease will outbreak to new heights, the activity on Twitter is much diminished since the initial fear and speculation

around the disease is not as large. There are fewer tweets with respect to news, education and reporting since the population has already been educated previously. What is left, we hope, is self-reporting. This last phase is the one that will be stable over time, where the flu currently stands, which makes it easier for social media to predict.

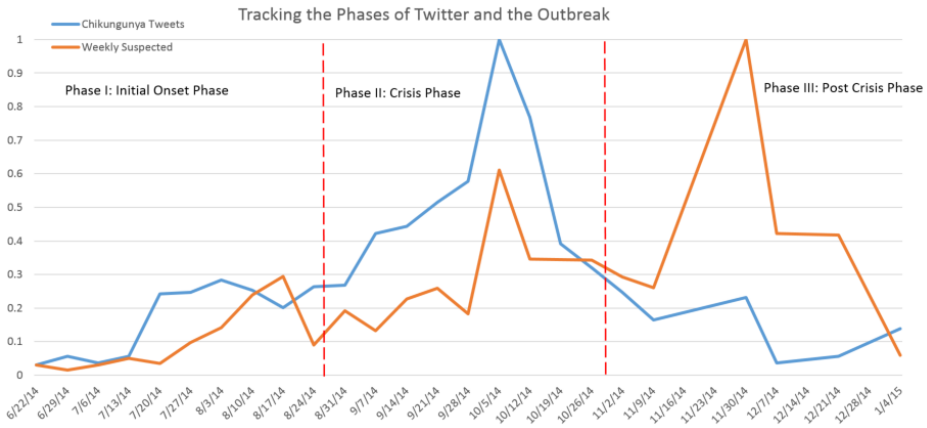


Figure 3.10: The Twitter phases plotted on top of the suspected cases and the one-week lagged Twitter signal, as in Figure 3.8; Phase I, "the Onset Phase", leads up to the first outbreak from the first suspicions of a possible new epidemic; Phase II, "the Crisis Phase", is during the first large outbreak; Phase III, "the Post-Crisis Phase", includes subsequent outbreaks to present

3.5 D3 as a Visualization Tool

The goal of the visualization was to ensure that the tweets were not all concentrated in one city or province. Although we expect more people to report from cities due to higher population and more bodies to spread the disease from, we wanted to observe the geographic distribution as well. In Figure 3.11, we see a screenshot of the interactive tool developed in D3.js which places a point for every tweet mentioning "Chikungunya" in a time range that you can select and drag on the top. We also highlight the computed number of counts in this time window as well.

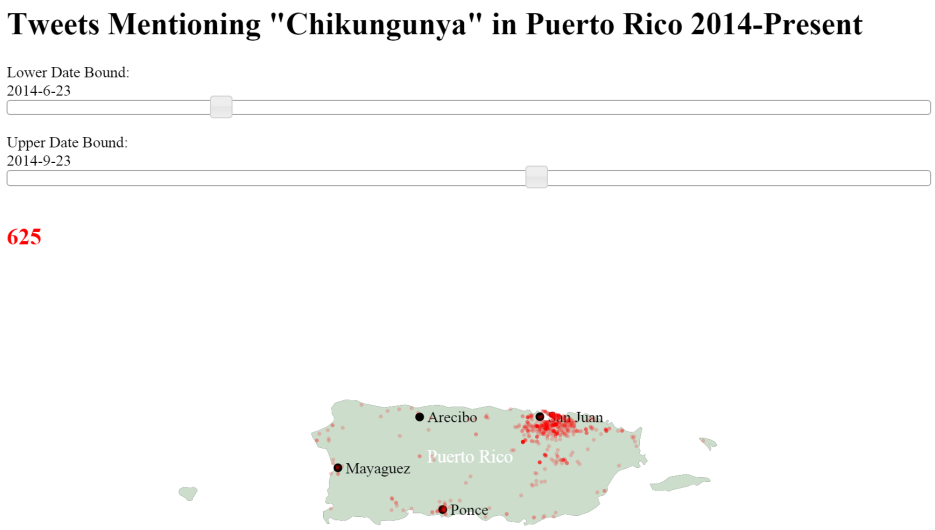


Figure 3.11: The interactive tool in D3.js which allows for interactive adjustable time frames, currently viewing from June 23 to September 23, spanning the first outbreak in Puerto Rico, centered around San Juan

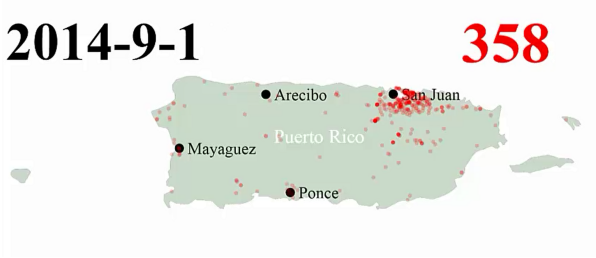
From the visualization, we also created an automated video to highlight the outbreak of the disease over time. This allowed us to see the progression over time. If we presume that our Twitter frequencies act as proxies for the actual cases, we can tell a story of the outbreak, which we will accompany with Figure 3.12. We notice that the first cases were near San Juan, which is as expected since it is the capital and largest city as seen in Figures 3.12a and 3.12b. For the first few weeks (Figure 3.12c, these cases are intermittent but during the months of July and August, we observe the first geographic outbreak in San Juan (Figures 3.12d and 3.12e in our Twitter space which is confirmed by the previously graphed PAHO data **??**. The second and large outbreak happens (refer again back to Figure **??**) where we see Chikungunya spread to Ponce (Figure 3.12fand then to the non-urban areas of Puerto Rico (Figure 3.12g). The outbreak subsides slightly in November but the newer cases that we do observe are more in the non-urban regions (Figure 3.12h. We conclude our geographic story of the Puerto Rico outbreak the cumulative graph of all the Twitter locations which we believe does model the outbreak, Figure 3.13.

(a) through May 1

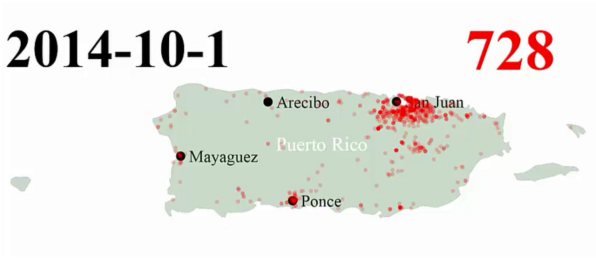
(b) through June 1

(c) through July 1

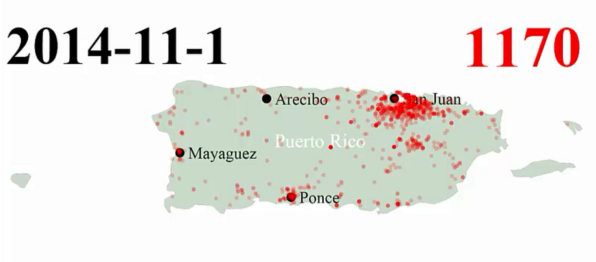
(d) through August 1



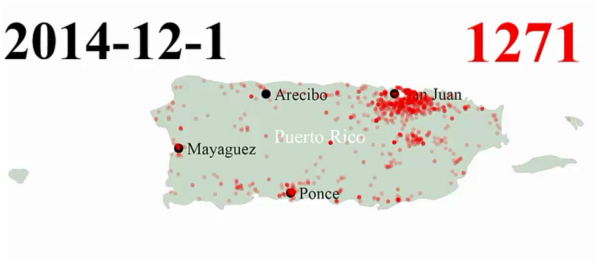
(e) through September 1



(f) through October 1



(g) through November 1



(h) through December 1

Figure 3.12: Cumulative tweets mentioning "Chikungunya" during for each month of the outbreak, each semi-transparent red dot is one hit

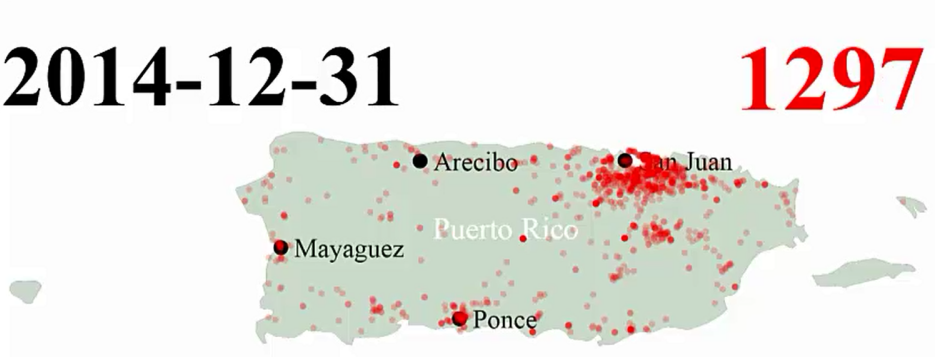


Figure 3.13: All of the locations of the classified tweets geographically plotted on a map, A screenshot from the video form of the deliverable

3.6 Variations on Using the Data

When working with the time series data, rather than using the weekly Twitter counts, we considered normalizing for that week's Twitter volume. Normalization allows the signal to be unbiased by weekly volume. In our analysis seen previously, from Figure 3.1 and Section refsec:weekly, we see that even with some seasonal fluctuation of volume, the range is around double, whereas our signal is very clear and is around 4-5 fold different between outbreaks. Still, we did try normalization over weekly volume and found that overall the shape is very similar, but normalization does lose the signal tracking the last outbreak during the post-scare phase discussed in Section 3.4, (Figure 3.14. This may be because being sick with Chikungunya during the epidemic may actually lead to an increased volume of tweets from those who are sick and normalization would actually bury the strength of the signal.

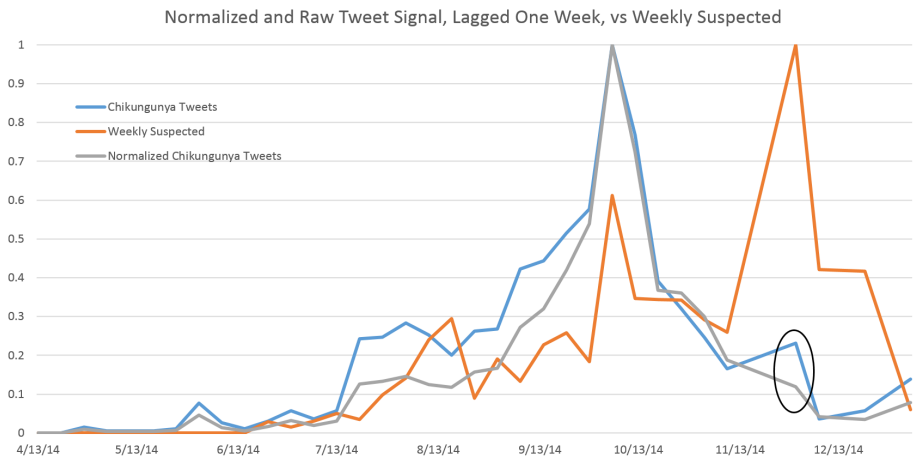


Figure 3.14: Normalized and raw Twitter signals over time, each scaled as a proportion to their maximum week, one week lag for Twitter signal is included; notice the black circle highlighting the loss of a relative peak which corresponds to a post-scare phase outbreak when accounting for normalization

We also considered normalizing by the mean of each signal, computing z-scores but with an expected background of zero with skewed left data (most weeks will have no signal on keywords when there is no outbreak), the mean and z-scores may not well represent sharp outbreak peaks.

3.7 Lessons from and Results of Curation

Due to our small data size, the 1500 tweets were classified by humans to ensure greatest accuracy. The categories would serve as independent variables to our regression. From our initial duration of the tweets, we found some surprising categories, which did not fall into the standard categories we were expecting but seemed relevant. Two such surprises were the Chikungunya Song (an actual pop song by artist Wayne J) and the Chikungunya Bracelet (a mosquito-repelling bracelet). Originally, they were marked as "Irrelevant", but later, were transferred to the "Awareness and Other Related" grouping. Although

not directly about being sick, these tweets were expected to correlate with the number of people sick, as increased awareness is generally due to people seeing the effects of the epidemic.

We sample some of the tweets below to show example tweets of each category in Figure 3.15.

Category	**Sample Tweet Text**
Reporting for Self	*"Me han diagnosticado Chikungunya, sinceramente esto es HORRIBLE!"* "Just been diagnosed with Chikungunya, honestly this is HORRIBLE!"
Reporting for Others	*"My uncle got sick with the chikungunya flu. #rip #2sick4life #getwell"*
Government/Educational	*"Reporte semanal de de Salud refleja 133 casos confirmados del chikungunya, lo que eleva a 386 los casos confirmados en el ao."* "Weekly Health Report confirms 133 cases of Chikungunya, brining the total to 386 cases this year"
Other Awareness	*"No encuentro mi pulsera anti chikungunya #nohayflow"* "I can't find my anti-Chikungunya bracelet #nohayflow"
Other Irrelevant/Unsure	*"Est la Chikungunya pero nos gusta ms la Ganya-chiku"* "It's 'Chikungunya', but we prefer 'Ganya-chiku'."

Figure 3.15: Sample tweets from each curation groups, tweets in Spanish translated in non-italics

We tried to get a sense of how each new category affected the correlation to see if there was a principal category that contributed the most. The below Figure 3.16 shows how certain subsets of the curation categories correlated to the CDC suspected case data prior to

the regression modeling. We see that most subsets do not perform as well as the total, which confirmed that during different phases of the Twitter scene, the most important curated categories would change in time. We do notice that news and awareness has the lowest correlation as the volume of news tweets steeply fall off in subsequent outbreaks. Slightly surprising was the strong correlation of the "Joke/Unknown" category. We hypothesize that this could be due to the nature of Twitter being very brief and sometimes cryptic with the 140 character cap. Twitter content can be private and meant for friends, and jokes may encoded for self-reporting, but unknown to our readers, as a third party.

Curation Category Subset	**Correlation to Suspected Cases**
Reporting for Self	0.711
Reporting for Others	0.745
News Related	0.459
Joke/Unknown	0.854
Other Relevant	0.869
Combined Reporting	0.742
Non-News	0.859
All	0.865

Figure 3.16: Correlation values to the suspected PAHO outbreak numbers, lagged one week, taking only tweet counts sorted into the categories named in each row

3.8 Predictive Modeling Using Lasso

We chose to run all regressions with Lasso (see Section **??** and **??** due to the easy to explain linear coefficients but with non-linear optimization paths. All training was done on data that was clipped to start on the week of June 22. This was the first week when

PAHO reported a suspected case number (confirmed numbers occurred a couple weeks earlier). During these weeks, the occurrence of "Chikungunya" on Twitter was already present, but still low enough to warrant claiming that we starting our predictive model training at the start of the "Onset Phase" of Twitter and the Chikungunya outbreak.

3.8.1 Discovering Meaningful Coefficients

With a less-confident source of actual weekly suspected cases, there is something to be learned from the coefficients of our linear model — which is one of the advantages of Lasso over traditional least-squares techniques. We trained our model on each of the three phases (discussed in Section 3.4) and then perform in-sample fitting. We observe high correlations (Figure **??** of our in-sample fitting indicating that the outbreak can be well represented by the linear model. We see the correlation of the lowest phase is the post-crisis phase, which is the one with the nosiest data in the PAHO reports, see Section 3.2. The graph of the time series shows the high quality fit except for the last phase, Figure 3.18.

Phase	**Correlation to Suspected Cases**
Onset Phase	0.943
Crisis Phase	0.929
Post-Crisis Phase	0.797
Total	0.897

Figure 3.17: Correlation values to the suspected PAHO outbreak numbers, lagged one week, taking only tweet counts sorted into the categories named in each row

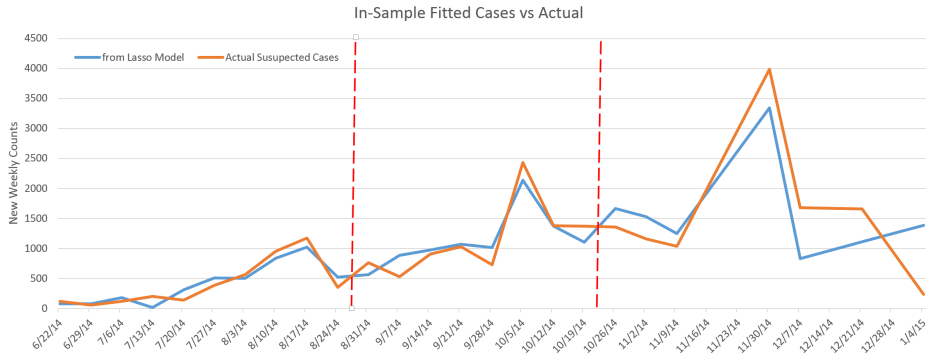


Figure 3.18: The fitted data (combined into one series) for all three phases show with the actual number of suspected cases

Most importantly, we look at Figure 3.19 to see the coefficients change given training for the three phases. Overall, news is almost always set to 0 which informs us that the volume of news tweets does not correlate with the outbreak, at any point. For the first phase, the two kept coefficients are for both grouping of "other", which can be because with fewer tweets, our model is more susceptible to noise. We believe that we are not missing any categories and that the onset phase is simply hard to predict using social media since volumes are low. For the crisis phase, we see the rise of importance of self-reporting, as related "other" tweets which comprised of a lot of peer education about the disease. During the post-crisis phase, unfortunately this reaches into the weak data, so we observe that again the "other irrelevant" tweets has the sole weight. With increased time points, and excising poor data from the gold standard, we believe that we will see more self-reporting only, as the education and news phase will have diminished.

Phase	Category	Coefficient b_i
Phase I: Onset	Reporting for Self	0
	Reporting for Others	0
	News Related	0
	Other Relevant	58.3
	Other Irrelevant	43.11
Phase II: Crisis	Reporting for Self	21.75
	Reporting for Others	0
	News Related	0
	Other Relevant	34.03
	Other Irrelevant	21.06
Phase III: Post-Crisis	Reporting for Self	0
	Reporting for Others	0
	News Related	0
	Other Relevant	0
	Other Irrelevant	139.32

Figure 3.19: Coefficients for each curation category for each of the three phases, as computed by our Lasso model

3.8.2 Out-Sample Prediction Models with Lasso

Static Train and Extend

After understanding the coefficients, we wished to demonstrate some out-sample tests to see correlation. We started with training on the same phases of data we distinguished before, so that the coefficients would be the same. We extended the fit to model all points

in the time series rather than just the time points in sample. We tried both training on phase I only, Figure 3.20, and phase II only, Figure **??** as these two phases showed the best correlation and had cleaner outbreak data.

We saw what we had expected. Since the early phase training is prone to noise, our model built from "other relevant" and "other irrelevant" categories did not expect to perform well. The measured correlation was 0.283. This model tracks the crisis phase fairly but in the post-crisis phase, fails to even pick up a signal with the third outbreak in November. When we trained on phase II, where "self-reporting" and "other relevant" tweets were incorporated, we saw a better model, with a correlation of 0.436, though still not the best. Our phase II model does detect the peak in late November (though only relatively as it misses the magnitude) but where the models fails is during the onset phase, as the first outbreak in August is missed. From these two static prediction models, we have shown that our predictions are very sensitive to how we categorize and where we train. However, some general guidelines, if needing to build a static training model, would be to pick an outbreak with significant volume and train on that data, just like our phase II in this case.

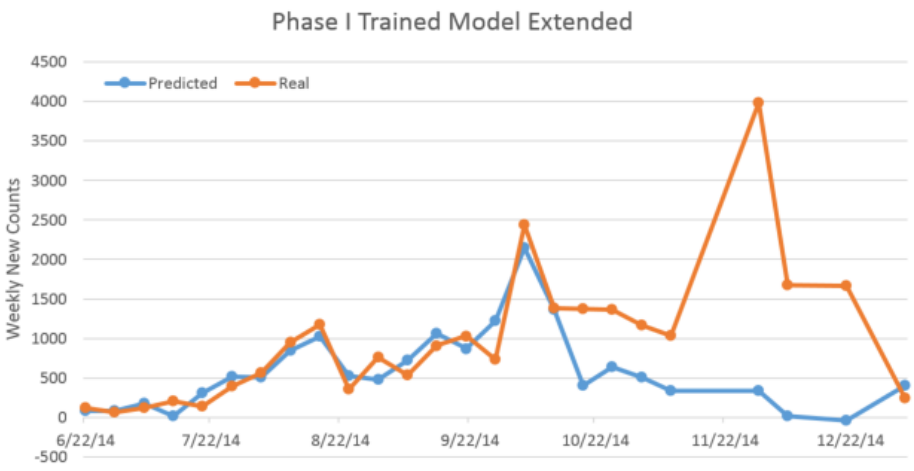


Figure 3.20: Predictions with onset phase coefficients of Figure **??** training extended to entire period

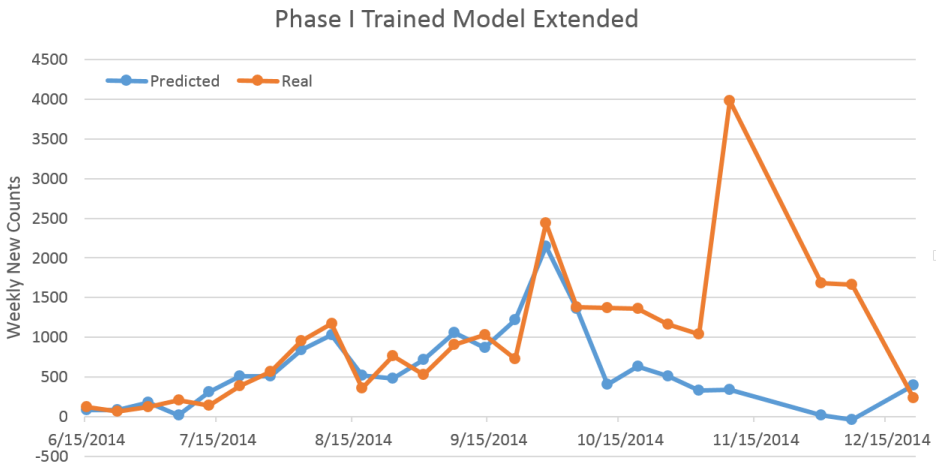


Figure 3.21: Predictions with peak phase coefficients of Figure **??** training extended to entire period

Full Dynamic Training

Since our static predictive modeling approach missed some peaks, we sought to improve our approach by allowing dynamic training. In dynamic training, we retrain our Lasso model at every week, to account for the most recent available data. With this paradigm, predictions made at the end of the 20 or so weeks train on all previous weeks. We see in Figure 3.22 the predictions of our dynamic model. The reported correlation for this model is only 0.179 and this confirms the visual analysis of not being a very good fit. We believe that Twitter trends will greatly vary from phase to phase and the cumulative trends are actually not strong. Using all previous data becomes detrimental. In addition, we have almost set ourselves up for low correlations because the cross-validation size is 3 requiring only 3 weeks of training data to predict. Note that in Figure 3.22, the first value we predict is the week of July 13 instead of June 22. This portion of our Twitter signal is quite variable as we've discussed and when we have only a few, noisy data points to train on, we can do no better than to make a wild prediction — which is why the early part of Figure 3.22 is much different from the actual values. We also visualize the coefficients

(normalized to be between 0 and 1 for each category) of the model are tracked over time in the form of a heatmap shown in Figure 3.23. We see that In the beginning, "irrelevant" tweets have larger weights but as time goes on, more and more weight is being put towards self-reporting, the single most consistently correlated category.

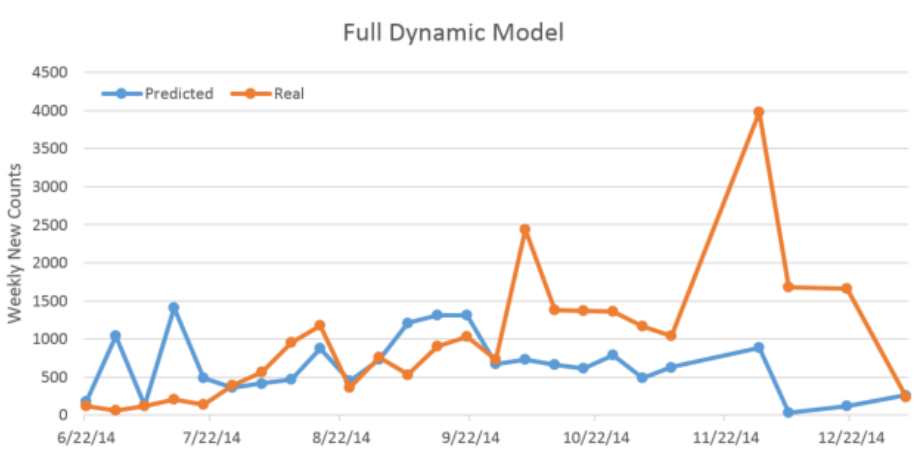


Figure 3.22: Predictions of the dynamic training model retrained weekly using all past n weeks of previous data to predict the $n+1$ value

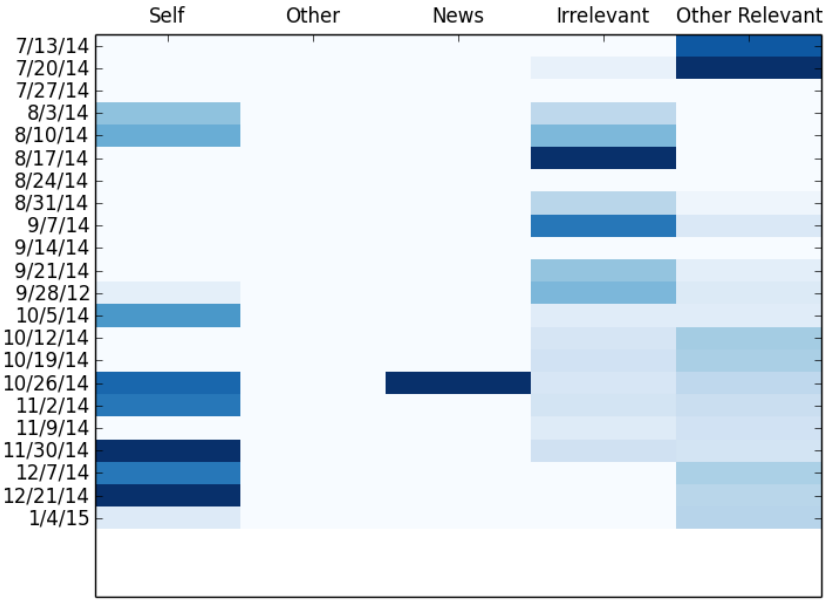


Figure 3.23: Coefficients of the dynamic training model retrained weekly using all past n weeks of previous data to predict the $n+1$ value; dark blue is maximum, white is minimum

Rolling Window Dynamic Training

From our full dynamic and static models, we have observed that phase to phase changes required active modeling to adjust for but that cumulative training did not work well. We then decided to try a rolling-window approach, where we would train on n past weeks to predict the value for the $(n+1)^{st}$ week. After each week, the new observed point would be added, the rolling window shifted and the Lasso model recomputed. We tried rolling window sizes of 3 and 5 and compared the predicted values to the actual suspected case data. With this approach, we hoped to be able to capture things that were locally relevant while using the most recent set of data possible to account for social behavior changes on Twitter.

The graphs of the predicted values for $n = 3$ is shown in Figure 3.24 and for $n = 5$, in Figure 3.26 Despite having correlations of 0.243 and 0.348 for $n = 3$ and $n = 5$ rolling window models, we note the improvement from the cumulative dynamic model in the previous section. As before, a large part of the low correlation comes from the noisy phase III data, which has too few reliable data points to fit to our linear model. Recall that even the in-sample fitting produced correlations of less than 0.80.

Our dynamic rolling approach is sensitive to the window size, n, and for this example, were chosen arbitrarily to be approximately half the size of the aforementioned Twitter phases. But with enough points in the future, the optimal size of the window can be computed. We notice that an $n = 3$ model is good at getting general levels right from week to week, but misses out on large spike weeks, namely the spike in early October and late November. The $n = 5$ model is better at predicting levels for the post-crisis phase but does not seem sensitive enough to pick up early outbreaks when self-reporting is not as prevalent — like for the August outbreak.

We again show the coefficients normalized within category for the time series as a heatmap for both (see Figures 3.25 and 3.27) rolling window models. In both models, we see the importance of the "other relevant" category which indicates that general awareness corresponds to outbreak levels. For the $n = 3$ model, we see that news is relevant during the first outbreak in the crisis phase in August, and then becomes irrelevant. Although in the $n = 5$ (Figure 3.27 model, we see news appear to have some weighting, this may be the fault of normalization, where things are scaled relative to their maximum magnitude. If the coefficient was very low, it will still normalize to dark blue since news is not important anywhere else it the time series. Encouragingly, we see in the $n = 5$ rolling window model that as time goes on moving away from the onset and crisis phases, we see the increased importance of self-reporting.

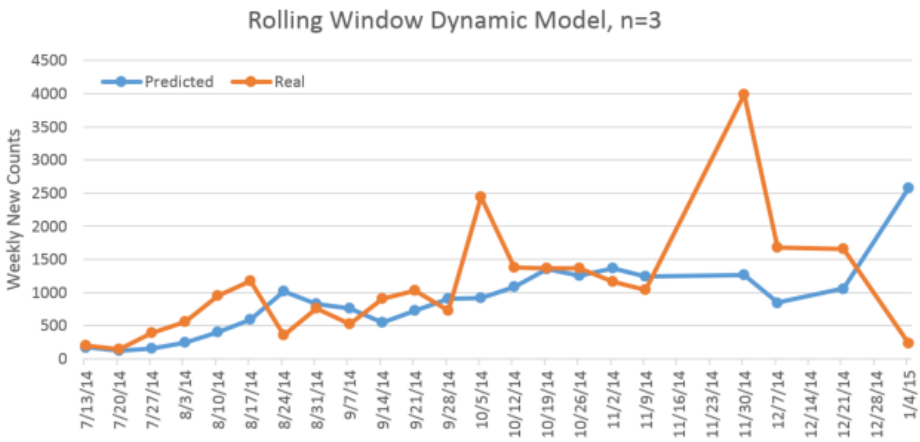


Figure 3.24: Predictions of the Lasso model retrained weekly accounting for the past 3 weeks of data

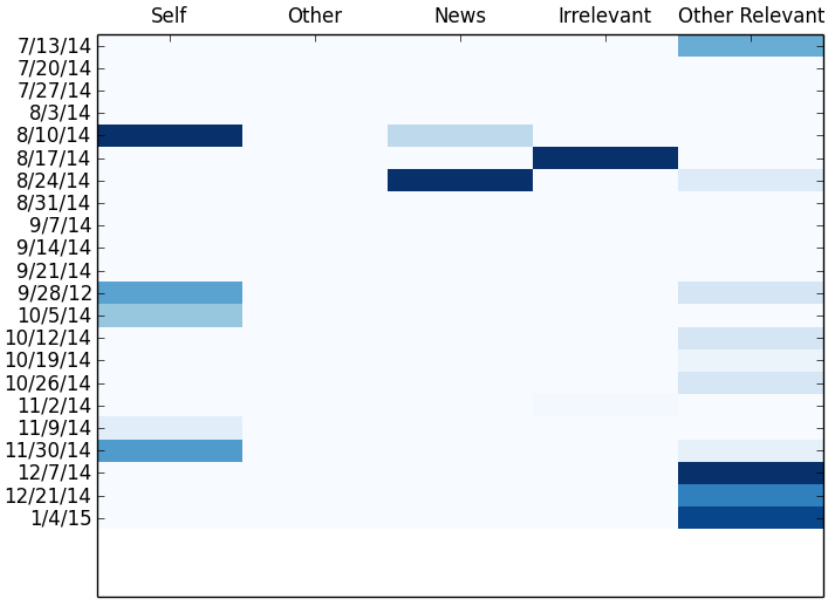


Figure 3.25: Coefficients of the Lasso model retrained weekly accounting for the past 3 weeks of data; dark blue is maximum, white is minimum

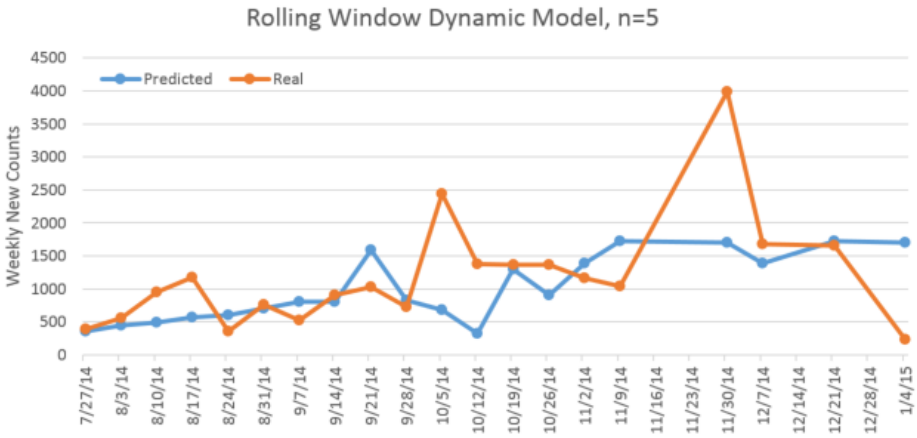


Figure 3.26: Predictions of the Lasso model retrained weekly accounting for the past 5 weeks of data

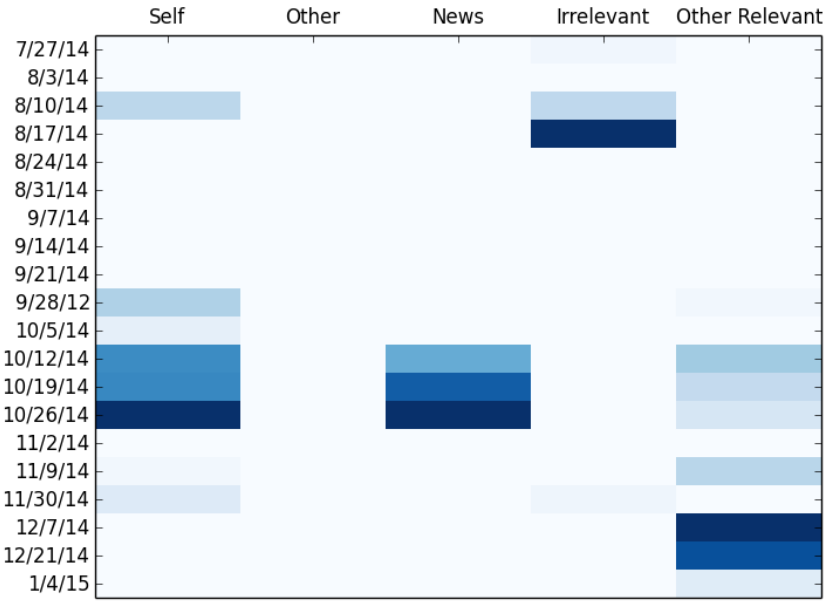


Figure 3.27: Coefficients of the Lasso model retrained weekly accounting for the past 5 weeks of data; dark blue is maximum, white is minimum

Chapter 4

Conclusion

"The goal is to turn data into information, and information into insight."

— Carly Fiorina, former CEO of Hewlett-Packard [16]

4.1 Future Work

4.1.1 Natural Language Processing

One of the most critical steps in our predictive pipeline is dividing up our flagged tweets by keyword into various categories. In our study, tweet curation was done by human readers, see Section 2.5 but with larger data sets in the future, or when applied to larger countries or more popular diseases, the process will need to be automated. The natural extension of that is to use NLP (Natural Language Processing) models, where the computer can automatically classify text by connotation, from using frequencies, structures and other linguistic techniques to build classifier [11, 4, 42].

One common model used is the bag-of-words model, where word frequencies are tallied in a grammar-free way and given the word counts used, a Bayesian method or other clustering methods run on the bag of words frequencies are used to determine how to classify each text [38, 11]. This model has shown to work well in many cases such as email spam filtering [38]. Extensions in this way are definitely worth testing but may run into issues due to our small dataset with texts that are limited in word counts. Twitter has the 140 character cap and this will hinder the recall of the bag-of-words model.

More complicated natural language processing models also exist, some of which even consider the grammar of the language under study [64]. We recognize that the nature of tweets can ignore grammar and be nonsensical, which will probably result in lower than expected F1 recall scores for most algorithms but different NLP models can be tried to help us get the most accurate automated curation [64, 4, 42].

One final consideration is that Puerto Rico (and other countries) may have tweets in two languages, requiring training sets for all categories in both languages, doubling the requirement of data [42]. Over time, as more tweets are collected, the training data will readily increase but if looking to categorize the tweets at an early phase of an outbreak when there are few tweets, manual curation may be a necessity.

4.1.2 Refinements to Continued Dynamic Modeling

We may also consider trying different models from which to predict the number of cases (the dependent variable) given our categorical counts (the independent variable). Since we believe there may be more complicated phases involved, models could be developed with additional parameters around time, which can in one model, build in an excitement decay, τ. Other modeling options can even extend to non-linear functions. Although our modeling resulted in linear coefficients for dependent variables, we were able to suggest

social implications explained by these linear coefficients. If our goal is to best predict the outbreak level in real time, non-linear models can be chosen that may have coefficients harder to map to social behavior, but may fit the problem better.

Finally, it may be useful to develop a method to compute error bars around the gold standard and in turn our predictions, to get a sense of variance. If another source of gold standard data were released, perhaps by either private studies or another government branch (even local government), numbers could be cross-validated and errors could be computed. Errors can also be considered by measuring week to week fluctuations if we assume some epidemiological model and fit certain parameters.

4.1.3 Generalizations

With a more detailed data set, we see that modeling the same disease for a more localized geographic region, say provinces of Puerto Rico, becomes a recursively similar task. We can easily apply the same workflow just by changing the bounding polygon from which the filters were first applied. Then we could repeat for each province. The only thing that will limit the extent of this "zooming in" will be the availability of gold standard data sets to train our regression model with. A contact at the CDC has informed of us possible provincial level data that the CDC maintains which may be further used to evaluate the power of Twitter data. We would hope to see that the trends and fitting carry over even once looking at a province or clustered by large cities. This would provide not only outbreak level but also outbreak location, a harder but even more important number to know when managing the outbreak in real time.

The beauty of the method which we present in this paper is lies in generalization. Our study is in name, applied to Chikungunya fever in Puerto Rico, but the pipeline is established to analyze any subject that would be expected to included in tweets over a

given period of time, within a certain location. For other diseases and other areas, the extension is trivial — apply different coordinates and search for the name of a different disease. However, to trace other subjects, such as revolution or awareness of an event, slight changes are envisioned. When wishing to track breaking news or trending topics like, it may be better to break down the groupings not from curation but from different keywords as most tweets in the onset and crisis phases are of general awareness. Different keywords can be tracked and serve as the independent variables for Lasso (such as two different mottoes of a revolutionary movement). If connotation is important, the curation process may need to be extended to group multiple keyword-filtered tweets into the same relevant buckets. As mentioned above in Section 2.5, this approach may limit the training size which will hinder natural language processing and other automation in the curation phase. An extra consideration when using Lasso to study coefficients will be to determine the "gold standard" for the data, which can be even more uncertain than in this epidemiological case. For cases that the gold standard itself is more uncertain, we favor the in-sample studies of coefficients to understand the story whereas cases which have verifiable correct answers (which is rare), we would favor the out-sample predictive modeling approaches.

In summary, we highlight that the following requirements to allow feasibility with our Twitter analytics method:

- Internet connectivity and Twitter usage in region of interest
- Topic that can be narrowed into a short list of keywords
- Sample sizes that are small enough to be hand curated (order of 10000) [37] or data sets as large as possible (and evenly distributed per dependent variable) for automated curation using NLP (see Section 4.1.1)

4.2 Final Insights

Chikungunya in Puerto Rico was a good model disease and country for our pipeline of using categorized Twitter counts to predict volume. There was high enough Twitter usage and geo-coded tweet data to be able to track the evolution of keywords over time. The golden standard of comparison was cleaned data scraped from near-weekly outbreak reports by PAHO. Initial correlations showed hope of matching Twitter frequency counts to one week lagged data of PAHO suspected cases, with a correlation of 0.86. This one week lag was explained by nature of Twitter posting being prior to medical visits. Although only one keyword was ultimately used to the next phase of Lasso regression due to the low signals of other tested keywords (like abbreviations or primary symptoms), multiple dependent variables were constructed by categorizing the connotations of tweets. With a handful of independent variables, Lasso regressions were run to understand importance of variables over time. We observed three distinct behavioral phases — an onset phase, followed by a crisis phase and then slowly transitioning to the stable, post-crisis phase. Each phase was characterized by coefficients computed from in-sample Lasso regression where we saw a large number of news tweets initially which transitioned to self-reporting and individual awareness tweets in the post-scare phase.

Finally, we tested predictive models with out-sample data to simulate being able to predict outbreaks of Chikungunya in real time given just the Twitter counts. Despite the correlations of our dynamic models to be mediocre at best, we attribute much of our error to the diminished quality of our data set towards the end of our time frame. We saw the weakest fits at the end of 2014, we believe that this is due to our small sample size and the uneven time intervals (loss of resolution for our data). We have seen from the coefficient analysis that we are studying useful categories that seem to explain social response to disease outbreak on Twitter. We believe that the rolling window Lasso regression method should provide the best results, given a few more outbreak months of data to evaluate. We

will continue to run our analysis as additional data becomes available through the spring to further evaluate out of sample predictions. If another outbreak occurs in the Spring, we will be able to fully test out-of-sample predictions.

When trained on the proper data after an initial Twitter excitement phase, we believe that Twitter data can be valuable to predict disease. From this exploration, we have begun to understand how initial outbreaks of disease diffuse through the Twitter space and how geo-tagging can help us locate not just a summary of the disease, but trace the location as well. Once the Twitter dust settles, the self-reporting nature of tweets, along with other awareness categories, can help estimate the extent of the current outbreak before any other case numbers are reported. This early response, done almost in real time, can help the medical world best prepare for and deal with diseases in the localized area to cure the most number of people.

Acknowledgments

First of all, a big thank you to my advisor, Mauricio Santillana, who first alerted me to the questions asked by epidemiology with Google Flu Trends applications of Applied Math 50. His ability to inspire new directions for the project was what allowed the thesis to come to its final form. He has shown me that epidemiology is truly a blend of math, computer science and storytelling. Under his guidance, it's clear that asking the right questions is the most critical.

I'd also like to thank Jared Hawkins whose expertise in working with the Twitter database has been a great idea wall to bounce ideas off of when otherwise just feeling for the dark in the mass of big data.

By maintaining the Twitter Database, Clark Freifeld has also helped this project come together..

Another shoutout to Tobi Skotnes and Courtney Baur who aided in the tweet curation process as well as to Andre Nguyen, who has also worked with Mauricio in sibling projects and helped start me off with the Lasso algorithm.

Finally, to John Brownstein and his lab who works on very similar projects all across the world of epidemiology, a nod of appreciation for allowing me me to present at group meetings as the data came in and asking questions that helped inspire some of the analysis.

References

[1] Geospatial Data Abstraction Library. http://www.gdal.org/, 2015. [Open source software; online; accessed 1-Feb-2015].

[2] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, pages 702–707. IEEE, 2011. [Online; accessed 23-Mar-2015].

[3] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Twitter Improves Seasonal Influenza Prediction. Technical report, University of Massachusetts Lowell, 2012. [Online; accessed 23-Mar-2015].

[4] James F Allen. Natural language processing. 2003. [Online; accessed 26-Mar-2015].

[5] Mike Bostock. Geospatial Data Abstraction Library. http://bost.ocks.org/mike/map/, 2012. [Online; accessed 31-Jan-2015].

[6] David A. Broniatowski, Michael J. Paul, and Mark Dredze. Twitter: Big data opportunities. *Science Letters*, 345(6193):148, 2014. [Online; accessed 23-Mar-2015].

[7] John S. Brownstein, Clark C. Freifeld, and Lawrence C. Madoff. Digital Disease Detection - Harnessing the Web for Public Health Surveillance. *New England Journal of Medicine*, 2009. [Online; accessed 25-Mar-2015].

[8] Declan Butler. When Google got flu wrong. *Nature News*, 494(7436), 2013. [Online; accessed 23-Mar-2015].

[9] Caribbean Business PR. Study: internet use still on rise in PR. http://www.caribbeanbusinesspr.com/news/study-internet-use-still-on-rise-in-pr-96953.html, May 2014. [Online; accessed 24-Mar-2015].

[10] CDC. About the Chikungunya Virus. http://www.cdc.gov/chikungunya/, 2015. [Online; accessed 19-Mar-2015].

[11] Gobinda G Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003. [Online; accessed 26-Mar-2015].

[12] Gabriela Ciuperca. Adaptive Lasso model selection in a multiphase quantile regression. *arXiv*, 1309(1262), 2014. [Online; accessed 4-Mar-2015].

[13] CNN Money. WOW Twitter soars 73 percent in IPO. http://money.cnn.com/2013/11/07/technology/social/twitter-ipo-stock/, 2013. [Online; accessed 25-Mar-2015].

[14] DataSift. Twitter. http://datasift.com/platform/datasources/twitter/, 2015. [Online; accessed 19-Mar-2015].

[15] Dell Inc. Dell Survey: Midmarket Companies Aggressively Embrace Big Data Projects. http://www.dell.com/learn/us/en/uscorp1/press-releases/2014-04-28-dell-software-big-data-midmarket-survey, 2014. [Online; accessed 17-Mar-2015].

[16] Amarendra B. Dhiraj. 25 Insightful And Thought-Provoking Quotes about Big Data. https://www.linkedin.com/pulse/

20140502105616-8781298-25-insightful-and-thought-provoking-quotes-about
May 2014. [Online; accessed 16-Mar-2015].

[17] Mark Dredze, Renyuan Cheng, Michael J. Paul, and David Broniatowski. HealthTweets.org: A Platform for Public Health Surveillance using Twitter. *Association for the Advancement of Artificial Intelligence*, 2014. [Online; accessed 25-Mar-2015].

[18] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least Angle Regression. 32(2):407–499, 2004.

[19] FluNearYou. Flu Activity in the United States. https://flunearyou.org/, 2015. [Online; accessed 21-Mar-2015].

[20] Emily Fox. Lasso Regression. Technical report, University of Washington, 2013. [Online; accessed 16-Feb-2015].

[21] J. Fu, Wenjiang. Penalized Regressions: The Bridge Versus the Lasso. 7(3):397–416, 1998.

[22] Jeremy Ginsberg, Matthew H. Mohebbi, Lynnette Patel, Rajan S. Brammer, Mark S. Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457, 2009. [Online; accessed 8-Dec-2014].

[23] GNIP. Twitter Data. https://gnip.com/sources/twitter/, 2015. [Online; accessed 19-Mar-2015].

[24] Janaina Gomide, Adriano Veloso, Wagner Meira Jr., Virgilio Almeida, Fabricio Benevenuto, Fernanda Ferraz, and Mauro Teixeira. Dengue surveillance based on a computational model of spatio-temporal locality of Twitter. *WebSci*, 2011. [accessed 3-Mar-2015].

[25] Google Public Data. Internet users as percentage of population for Puerto Rico. https://www.google.com/publicdata/explore?ds=d5bncppjof8f9_

&met_y=it_net_user_p2&idim=country:PRI, 2015. [Online; accessed 24 -Mar-2015].

[26] Google Scholar. Scholar Search for Twitter API. https://scholar.google.com/scholar?q=Twitter+API, 2015. [Online; accessed 18-Mar-2015].

[27] Google.org. Google Dengue Trends. https://www.google.org/denguetrends/, 2014. [Online; accessed 1-Dec-2014].

[28] Google.org. Google Flu Trends. https://www.google.org/flutrends/us/#US, 2014. [Online; accessed 1-Dec-2014].

[29] Elaine Grant. The promise of big data. http://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/, 2012. [Online; accessed 24-Mar-2015].

[30] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition, 2009.

[31] HealthMap.org Team. US Flu activity: Estimate of percentage of people with flu like symptoms. http://www.healthmap.org/flucast/, 2015. [Online; accessed 25-Mar-2015].

[32] IEEE. BDCE 2014 First IEEE International Workshop on Big Data in Computational Epidemiology. http://www.vbi.vt.edu/ndssl/upcoming-events/upcoming-events-view/bdce-2014-first-ieee-international-workshop-on-big-data-in-computationa 2014. [Online; accessed 24-Mar-2015].

[33] Internet Live Stats. Twitter Usage Statistics. http://www.internetlivestats.com/twitter-statistics/, 2015. [Online; accessed 23-Mar-2015].

[34] Michelle Kantrow. 88.1 percent of local social media users connect daily, most gravitate toward Facebook. http://newsismybusiness.com/88-1-of-local-social-media-users-connect-daily-most-gravitate-toward-fa 2012. [Online; accessed 24-Mar-2015].

[35] Michelle Kantrow. Study: Puerto Rico Internet users growing , more mobile. http://newsismybusiness.com/study-puerto-rico-internet-users-growing-more-mobile/, May 2013. [Online; accessed 24-Mar-2015].

[36] Alex Lamb, Michael J. Paul, and Mark Dredze. Investigating Twitter as a Source for Studying Behavioral Response to Epidemics. *Association for the Advancement of Artificial Intelligence*, 2012. [Online; accessed 25-Mar-2015].

[37] Alex Lamb, Michael J. Paul, and Mark Dredze. Separating Fact from Fear: Tracking Flu Infections on Twitter. *Association for Computational Linguistics*, 2013. [Online; accessed 25-Mar-2015].

[38] S. Lazebnik, A. Torralba, L. Fei-Fei, D. Lowe, and C. Szurka. Bag-of-Words models. Technical report, New York University, 2012. [Online; accessed 26-Mar-2015].

[39] Joy L. Lee, Matthew DeCamp, Mark Dredze, Margaret S. Chisolm, and Zackary D. Berger. What Are Health-Related Users Tweeting? A Qualitative Content Analysis of Health-Related Users and Their Messages on Twitter. *Journal of Medical Internet Research*, 16(10), 2014. [Online; accessed 25-Mar-2015].

[40] Kalev H. Leetaur, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. Mapping the global twitter heartbeat: The geography of twitter. *FirstMonday*, 18(5–6), May 2013. [Online; accessed 18-Mar-2015].

[41] J. Legrand, R. F. Grais, P. Y. Boelle, A. J. Valleron, and A. Flahault. Understanding the dynamics of Ebola epidemics. *Epidemiology and Infection*, 135:610–621, 2007. [accessed 4-Mar-2015].

[42] Elizabeth D Liddy. Natural language processing. 2001. [Online; accessed 26-Mar-2015].

[43] Steve Lohr. Google Flu Trends: The Limits of Big Data. `http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/`, 2014. [Online; accessed 1-Dec-2014].

[44] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Home Location Identification of Twitter Users. *arXiv*, 1403(2345), 2012. [Online; accessed 19-Mar-2015].

[45] Bernard Marr. The Top 10 Big Data Quotes of All Time. `http://smartdatacollective.com/bernardmarr/232941/top-10-big-data-quotes-all-time`, 2014. [Online; accessed 16-Mar-2015].

[46] Michael Mathioudakis and Nick Koudas. Twittermonitor: trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 1155–1158. ACM, 2010. [Online; accessed 23-Mar-2015].

[47] Mor Naaman, Hila Becker, and Luis Gravano. Hip and trendy: Characterizing emerging trends on Twitter. *Journal of the American Society for Information Science and Technology*, 62(5):902–918, 2011. [Online; accessed 23-Mar-2015].

[48] Ruchit Nagar, Qingyu Yuan, Clark C. Freifeld, Mauricio Santillana, Aaron Nojima, Rumi Chunara, and John S. Brownstein. A Case Study of the New York City 2012-2013 Influenza Season with Daily Geocoded Twitter Data From Temporal and Spatiotemporal Perspectives. *Journal of Medical Internet Research*, 16(10), 2014. [Online; accessed 23-Mar-2015].

[49] Netflix. Netflix Prize. http://www.netflixprize.com/, 2009. [Online; accessed 24-Mar-2015].

[50] Elaine Nsoesie. Digital Disease Detection: Using Social Media To Predict Trends. http://www.healthmap.org/site/diseasedaily/article/digital-disease-detection-using-social-media-predict-flu-trends-31114, 2014. [Online; accessed 25-Mar-2015].

[51] Natioanl Institude of Health. The Human Genome Project Completion: Frequently Asked Questions. https://www.genome.gov/11006943, 2010. [Online; accessed 27-Mar-2015].

[52] Sunday O. Oyeyemi, Elia Gabarron, and Rolf Wynn. Ebola, Twitter, and misinformation: a dangerous combination? *BMJ Letters*, 349(6178), 2014. [Online; accessed 25-Mar-2015].

[53] PAHO. PAHO Chikungunya Statistic Data. http://www.paho.org/hq/index.php?option=com_topics&view=readall&cid=5927, 2015. [Online; accessed 10-Jan-2015].

[54] Michael J. Paul and Michael Dredze. You Are What You Tweet: Analyzing Twitter for Public Health. *Association for the Advancement of Artificial Intelligence*, 2011. [accessed 1-Mar-2015].

[55] Raquel Pimentel, Ronald Skewes-Ramm, and Jose Moya. Chikungunya en la Republica Dominicana: lecciones aprendidas en los primeros seis meses [Chikungunya in the Dominican Republic: lessons learned in the first six months]. *Rev Panam Salud Publica*, 36(5):336–341, 2014. [accessed 3-Mar-2015].

[56] Gil Press. A Very Short History of Big Data. http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/, May 2013. [Online; accessed 24-Mar-2015].

[57] Riot Games, Inc. League of Legends Full API Reference. http://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub, 2015. [Online; accessed 26-Mar-2015].

[58] Gurvinder Rull. Chikungunya Fever. http://www.patient.co.uk/doctor/chikungunya-fever, 2011. [Online; accessed 19-Mar-2015].

[59] SA Dept. of Health. About Chikungunya. http://www.sahealth.sa.gov.au/wps/wcm/connect/public+content/sa+health+internet/health+topics/health+conditions+prevention+and+treatment/infectious+diseases/chikungunya+virus/chikungunya+virus+-+symptoms+treatment+and+prevention, 2014. [Online; accessed 19-Mar-2015].

[60] Parker Schultz. 10 Great Quotes About Big Data. http://www.businessproweekly.com/business-intelligence/10-great-quotes-about-big-data/?mode=featured, 2014. [Online; accessed 16-Mar-2015].

[61] Tyler M. Sharp, Nicole M. Roth, Jomil Torres, Kyle R. Ryff, Nicole M. Perez Rodriguez, Chanis Mercado, Mario del Pilar Diaz Padro, Maria Ramos, Raina Phillips, Matthew Lozier, Carmen S. Arriola, and Michael Johansson. Chikungunya Cases Identified Through Passive Surveillance and Household Investigations — Puerto Rico, May 5 to August 12, 2014. *Centers for Disease control and Prevention Morbidity and Mortality Weekly Report*, 63(48), 2014. [accessed 15-Jan-2015].

[62] SickWeather. SickWeather Live Map. http://www.sickweather.com/live-map.php, 2015. [Online; accessed 21-Mar-2015].

[63] Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic. *PLOS one*, May 2011. [Online; accessed 23-Mar-2015].

[64] P Spyns. Natural language processing. *Methods of information in medicine*, 35(4):285–301, 1996. [Online; accessed 26-Mar-2015].

[65] Robert Tibshirani. A simple explanation of Lasso and Least Angle Regression. Technical report, Stanford University. [Online; accessed 17-Feb-2015].

[66] Robert Tibshirani. Regression Regression and Selection via the Lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996. [Online; accessed 12-Feb-2015].

[67] Twitter Inc. About Twitter, Inc. `https://about.twitter.com/company`, 2015. [Online; accessed 18-Mar-2015].

[68] Twitter Inc. Twitter REST APIs. `https://dev.twitter.com/rest/public`, 2015. [Online; accessed 18-Mar-2015].

[69] Twitter Inc. What Is Twitter? `https://about.twitter.com/what-is-twitter`, 2015. [Online; accessed 25-Mar-2015].

[70] WHO. About Chikungunya. `http://www.who.int/mediacentre/factsheets/fs327/en/`, 2015. [Online; accessed 19-Mar-2015].

[71] WinShuttle. Big Data and the History of INformation Storage. `http://www.winshuttle.com/big-data-timeline/`, 2014. [Online; accessed 24-Mar-2015].

[72] Yi Zhuang. Building a complete Tweet Index. `https://blog.twitter.com/2014/building-a-complete-tweet-index`, 2014. [Online; accessed 23-Mar-2015].