Deploying nEmesis: Preventing Foodborne Illness by Data Mining Social Media

Adam Sadilek, Henry Kautz, Lauren DiPrete, Brian Labus, Eric Portman, Jack Teitel, Vincent Silenzio

■ *Foodborne illness afflicts 48 million people annually in the US alone. More than 128,000 are hospitalized and 3000 die from the infection. While preventable with proper food safety practices, the traditional restaurant inspection process has limited impact given the predictability and low frequency of inspections, and the dynamic nature of the kitchen environment. Despite this reality, the inspection process has remained largely unchanged for decades. CDC has even identified food safety as one of seven "winnable battles"; however, progress to date has been limited. In this work, we demonstrate significant improvements in food safety by marrying AI and the standard inspection process. We apply machine learning to Twitter data, develop a system that automatically detects venues likely to pose a public health hazard, and demonstrate its efficacy in the Las Vegas metropolitan area in a double-blind experiment conducted over three months in collaboration with Nevada's health department. By contrast, previous research in this domain has been limited to indirect correlative validation using only aggregate statistics. We show that the adaptive inspection process is 64 percent more effective at identifying problematic venues than the current state of the art. If fully deployed, our approach could prevent more than 9000 cases of foodborne illness and 557 hospitalizations annually in Las Vegas alone. Additionally, adaptive inspections result in unexpected benefits, including the identification of venues lacking permits, contagious kitchen staff, and fewer customer complaints filed with the Las Vegas health department.*

Since its inception, social media have been routinely data mined for marketing consumer goods. Starting around 2010, researchers began to realize that the same techniques could be used for influenza surveillance (Culotta 2010). Since then, social media analytics for public health has been expanded to monitor a variety of conditions, including cholera (Chunara, Andrews, and Brownstein 2012), mental health (Golder and Macy 2011), and diet (Widener and Li 2014). This body of work has shown that social media can be a useful complement to traditional methods, such as surveys of medical providers or individuals, for gathering aggregate public health statistics. Our work extends the social media analytics approach to a new domain, foodborne illness. Our most important contribution, however, is that we go beyond simply monitoring population-level prevalence. Our system, nEmesis, provides specific, actionable information, which is used to support effective public health interventions.

The fight against foodborne illness is complicated by the fact that many cases are not diagnosed or traced back to specific sources of contaminated food. In a typical US city, if a food establishment passes its routine inspection, it may not see the health department again for up to a year. Food establishments can roughly predict the timing of their next inspection and prepare for it. Furthermore, the kitchen environment is dynamic, and ordinary inspections merely provide a snapshot view. For example, the day after an inspection, a contagious cook or server could come to work or a refrigerator could break, either of which can lead to food poisoning. Unless the outbreak is massive, the illness is unlikely to be traced back to the venue.

CDC has identified food safety as one of seven "winnable battles,"[1] along with vehicle accidents and HIV, but progress to date on eradicating the disease has been limited. Our work adds to the arsenal of tools we as humanity can use to fight disease.

We present a novel method for detecting problematic venues quickly — before many people fall ill. We use the term *adaptive inspections* for prioritizing venues for inspection based on evidence mined from social media. Our system (nEmesis) applies machine learning to real-time Twitter data — a popular microblogging service where people post message updates (tweets) that are at most 140 characters long. A tweet sent from a smartphone is usually tagged with the user's precise GPS location. We infer the food venues each user visited by "snapping" his or her tweets to nearby establishments (figure 1). We develop and apply an automated language model that identifies Twitter users who indicate they suffer from foodborne illness in the text of their public online communication. As a result, for each venue, we can estimate the number of patrons who fell ill shortly after eating there. In this paper, we build on our prior work, where we showed a correlation between the number of "sick tweets" attributable to a restaurant and its historic health inspection score (Sadilek et al. 2013). In this paper, we deploy an improved version of the model and validate its predictions in a controlled experiment.

The Southern Nevada Health District (SNHD) conducted a three-month controlled experiment with nEmesis beginning January 2, 2015. Venues with the highest predicted risk on any given day were flagged and subsequently verified through a thorough inspection by an environmental health specialist. For each adaptive inspection, we perform a paired control inspection independent of the online data to ensure full annual coverage required by law and to compensate for the geographic bias of Twitter data. During the first three months, the environmental health specialists inspected 142 venues, half using nEmesis and half following the standard protocol. The latter set of inspections constitutes our control group. The inspectors were not told whether the venue comes from nEmesis or control.

nEmesis downloads and analyzes all tweets that originate from Las Vegas in real time. To estimate visits to restaurants, each tweet that is within 50 meters of a food venue is automatically "snapped" to the nearest one as determined by the Google Places API. We used Google Places to determine the locations of establishments because it includes latitude/longitude data that is more precise than the street address of licensed food venues. As we will see, this decision allowed nEmesis to find problems at unlicensed venues.

For this snapping process, we only consider tweets that include GPS coordinates. Cell phones determine their location through a combination of satellite GPS, WiFi access point fingerprinting, and cell-tower triangularization (Lane et al. 2010). Location accuracy typically ranges from 9 meters to 50 meters and is highest in areas with many cell towers and Wi-Fi access points. In such cases, even indoor localization (for example, within a mall) is accurate.

Once nEmesis snaps a user to a restaurant, it collects all of his or her tweets for the next five days, including tweets with no geo-tag and tweets sent from outside of Las Vegas. This is important because most restaurant patrons in Las Vegas are tourists, who may not show symptoms of illness until after they leave the city. nEmesis then analyzes the text of these tweets to estimate the probability that the user is suffering from foodborne illness.

Determining if a tweet indicates foodborne illness of the user is more complex than simply scanning for a short list of key words. By its nature, Twitter data is noisy. Even a seemingly explicit message, such as "I just threw up," is incomplete evidence that the author of the tweet has a foodborne illness. By using a language model rather than relying on individual key words, our method is able to better model the meaning behind the tweet and is therefore able to capture even subtle messages, such as "have to skip work tomorrow" or "I need to go to a pharmacy." Figure 1 lists the 20 most significant positive and negative language features that contribute to the score.

nEmesis then associates the individual sickness scores to the food venues from which the users originally tweeted. Each snapped twitter user is a proxy for an unknown number of patrons that visited but did not tweet. Since contracting foodborne illness and tweeting at the right times and places is a relatively rare occurrence, even a single ill individual can be a strong evidence of a problem. The web interface (figure 2) is used by the managing health specialist to sort venues by the number of sick users and to dispatch inspectors.

Figure 3 illustrates the full nEmesis process. On a typical day we collect approximately 15,900 geotagged tweets from 3600 users in the Las Vegas area. Approximately 1000 of these tweets, written by 600

Positive Feature		Negative Features	
Feature	**Weight**	**Feature**	**Weight**
stomach	1.7633	think i'm sick	– 0.8411
stomachache	1.2447	i feel soooo	– 0.7156
nausea	1.0935	f--k i'm	– 0.6393
tummy	1.0718	@ID sick to	– 0.6212
#upsetstomach	0.9423	sick of being	– 0.6022
nauseated	0.8702	ughhh cramps	– 0.5909
upset	0.8213	cramp	– 0.5867
naucious	0.7024	so sick omg	– 0.5749
ache	0.7006	tired of	– 0.5410
being sick man	0.6859	cold	– 0.5122
diarrhea	0.6789	burn sucks	– 0.5085
vomit	0.6719	course i'm sick	– 0.5014
@ID i'm getting	0.6424	ifi'm	– 0.4988
#tummyache	0.6422	is sick	– 0.4934
#stomachache	0.6408	so sick and	– 0.4904
i've never been	0.6353	omg i am	– 0.4862
threw up	0.6291	@LINK	– 0.4744
i'm sick great	0.6204	@ID sick	– 0.4704
poisoning	0.5879	if	– 0.4695
feel better tomorrow	0.5643	i feel better	– 0.4670

Figure 1. The Top 20 Most Significant Negatively and Positively Weighted Features in Our Language Model.

unique users, snap to a food venue. nEmesis then tracks these 600 users and downloads all their subsequent tweets for the following five days. These subsequent tracked tweets are then scored by the language model. Finally, venues are ranked based on the number of tweets with sickness score exceeding the threshold of 1.0 determined on a withheld validation set. During the experiment, nEmesis identified on average 12 new tweets per day that were strongly indicative of foodborne illness. Figure 4 shows a distribution over health scores inferred by nEmesis.

Significance of Results

To the best of our knowledge, this is the first study that directly tests the hypothesis that social media provide a signal for identifying specific sources of any disease through a controlled, double-blind experiment during a real-world deployment. By contrast, prior work has been anecdotal, limited to finding correlations, and/or didn't include a control group.

Related Work

Since the famous cholera study by John Snow (1855), much work has been done in capturing the mechanisms of epidemics. There is ample previous work in computational epidemiology on building relatively coarse-grained models of disease spread through differential equations and graph theory (Anderson and May 1979, Newman 2002), by harnessing simulated

Figure 2. nEmesis Web Interface.

The top window shows a portion of the list of food venues ranked by the number of tweeted illness self-reports by patrons. The bottom window provides a map of the selected venue, and allows the user to view the specific tweets that were classified as illness self-reports.

populations (Eubank et al. 2004), and by analysis of official statistics (Grenfell, Bjornstad, and Kappey 2001). Such models are typically developed for the purposes of assessing the impact a particular combination of an outbreak and a containment strategy would have on humanity or ecology (Chen, David, and Kempe 2010).

However, the above works focus on aggregate or simulated populations. By contrast, we address the problem of predicting the health of real-world populations composed of individuals embedded in a social structure and geo-located on a map.

Most prior work on using data about users' online behavior has estimated aggregate disease trends in a large geographical area, typically at the level of a state or large city. Researchers have examined influenza tracking (Culotta 2010; Achrekar et al. 2012; Sadilek and Kautz 2013; Broniatowski and Dredze 2013; Brennan, Sadilek, and Kautz 2013), mental health and depression (Golder and Macy 2011; De Choudhury et al. 2013), as well as general public health across a broad range of diseases (Brownstein, Freifeld, and Madoff 2009; Paul and Dredze 2011b).

Some researchers have begun modeling health and contagion of specific individuals by leveraging fine-grained online social and web search data (Ugander et al. 2012; White and Horvitz 2008; De Choudhury et al. 2013). For example, in Sadilek, Kautz, and Silenzio (2012) we showed that Twitter users exhibiting symptoms of influenza can be accurately detected using a model of language of Twitter posts. A detailed epidemiological model can be subsequently built by following the interactions between sick and healthy individuals in a population, where physical encounters are estimated by spatiotemporal colocated tweets.

Our earlier work on nEmesis (Sadilek et al. 2013) scored restaurants in New York City by their number of sick tweets using an initial version of the language model described here. We showed a weak but significant correlation between the scores and published NYC Department of Health inspection scores. Although the data came from the same year, many months typically separated the inspections and the tweets.

Other researchers have recently tried to use Yelp restaurant reviews to identify restaurants that should be inspected (Harrison et al. 2014). Key words were used to filter 294,000 Yelp reviews for New York City to 893 possible reports of illness. These were manually screened and resulted in the identification of 3 problematic restaurants.

Background: Foodborne Illness

Foodborne illness, known colloquially as food poisoning, is any illness that results from the consumption of contaminated food, pathogenic bacteria, viruses, or parasites that contaminate food, as well as the consumption of chemical or natural toxins such as poisonous mushrooms. The US Centers for Disease Control and Prevention (CDC) estimates that 47.8 million Americans (roughly 1 in 6 people) are sickened each year by foodborne disease. Of that total, nearly 128,000 people are hospitalized, while just over 3000 die of foodborne diseases (CDC 2013).

CDC classifies cases of foodborne illness according to whether they are caused by one of 31 known foodborne illness pathogens or by unspecified agents. These 31 known pathogens account for 9.4 million (20 percent of the total) cases of food poisoning each year, while the remaining 38.4 million cases (80 percent of the total) are caused by unspecified agents.

Food poisoning episodes associated with these 31 known pathogens account for an estimated 44 percent of all hospitalizations resulting from foodborne illness, as well as 44 percent of the deaths. Of these 31 known pathogens, the top five (*Norovirus*, *Salmonella*, *Clostridium perfringens*, *Campylobacter* species, and *Staphylococcus aureus*) account for 91 percent of the cases of foodborne illness, 88 percent of the cases that require hospitalization, and 88 percent of the cases that result in death. The economic burden of health losses resulting from foodborne illness are staggering. One recent study estimated the aggregated costs in the United States alone to be $77.7 billion annually (Scharff 2012).

Despite the variability in the underlying etiology of foodborne illness, the signs and symptoms of disease overlap considerably. The most common symptoms include vomiting, diarrhea (occasionally bloody), abdominal pain, fever, and chills. These symptoms can be mild to serious, and may last from hours to several days. Some pathogens can also cause symptoms of the nervous system, including headache, numbness or tingling, blurry vision, weakness, dizziness, and even paralysis. The gastrointestinal fluid losses can commonly result in dehydration, leading to secondary symptoms such as excessive thirst, infrequent urination, dark-colored urine, lethargy, and lightheadedness. Typically, symptoms appear within hours, but may also occur days to even weeks after exposure to the pathogen (Morris and Potter 2013). According to the US Food and Drug Administration (FDA), the vast majority of these symptoms will occur within three days (FDA 2012).

Public health authorities use an array of surveillance systems to monitor foodborne illness. In the United States, the CDC relies heavily on data from state and local health agencies, as well as more recent systems such as sentinel surveillance systems and national laboratory networks, which help improve the quality and timeliness of data (CDC 2013). An example of the many systems in use by CDC would include the Foodborne Diseases Active Surveillance Network, referred to as FoodNet. FoodNet is a sentinel surveillance system using information provided from sites in 10 states, covering about 15 percent of the US population, to monitor illnesses caused by seven bacteria or two parasites commonly transmitted through food. Other systems include the National Antimicrobial Resistance Monitoring System (NARMS), the National Electronic Norovirus Outbreak Network (CaliciNet), and the National Molecular Subtyping Network for Foodborne Disease Surveillance (PulseNet), among many others.

A major challenge in monitoring foodborne illness is in capturing actionable data in real time. Like all disease surveillance programs, each of the systems currently in use by CDC to monitor foodborne illness can entail significant time lags between when cases are identified and the data is analyzed and reported.

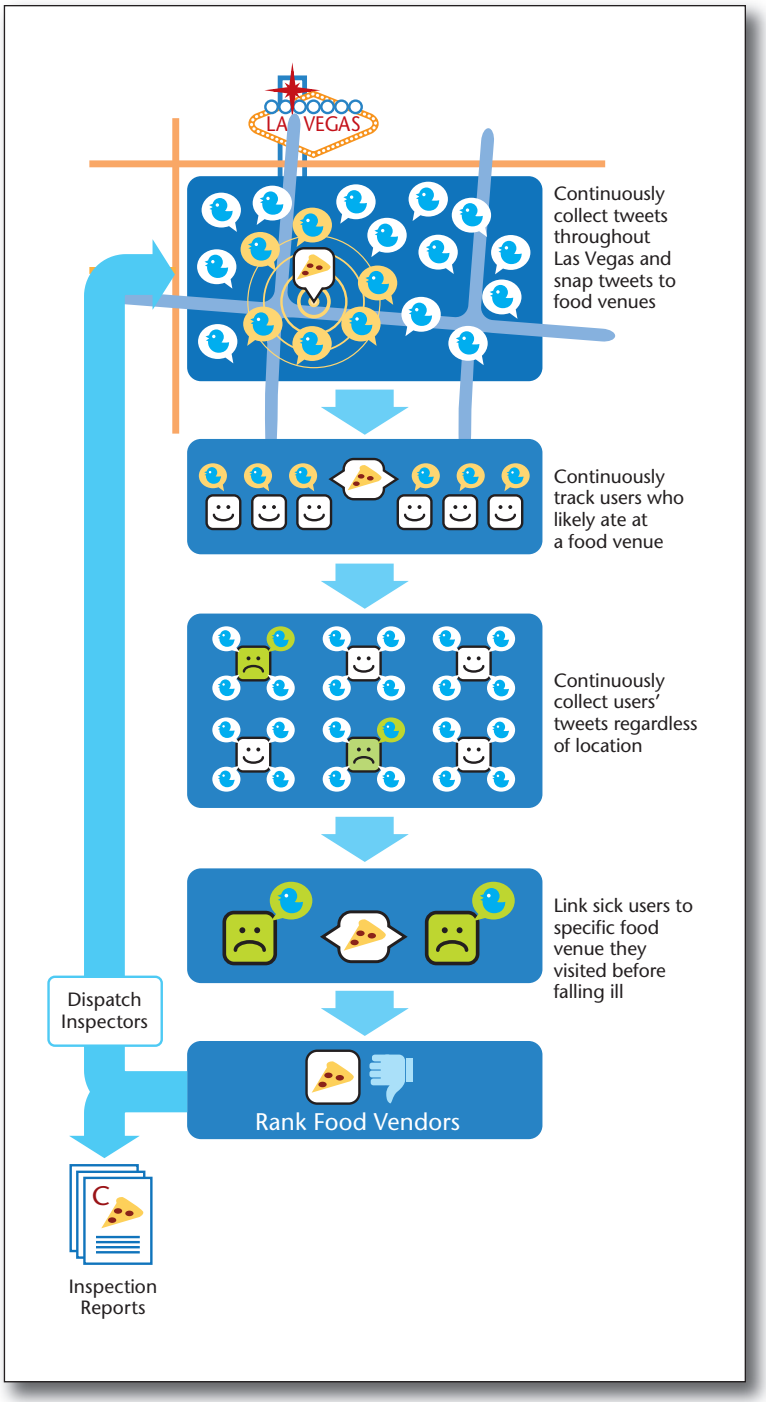

Figure 3. Adaptive Inspection Process.

Starting from the top: All tweets geo-tagged in the Las Vegas area are collected. Tweets geo-tagged within 50 meters of a food venue are snapped to that venue, and the Twitter IDs of the users are added to a database of users to be tracked. All tweets of tracked users are collected for the next five days, whether or not the users remain in Las Vegas. These tweets are evaluated by the language model to determine which are self-reports of symptoms of foodborne illness. Venues are ranked according to the number of patrons who later reported symptoms. Health department officials use the nEmesis web interface to select restaurants for inspection. Inspectors are dispatched to the chosen restaurants, and findings reported.

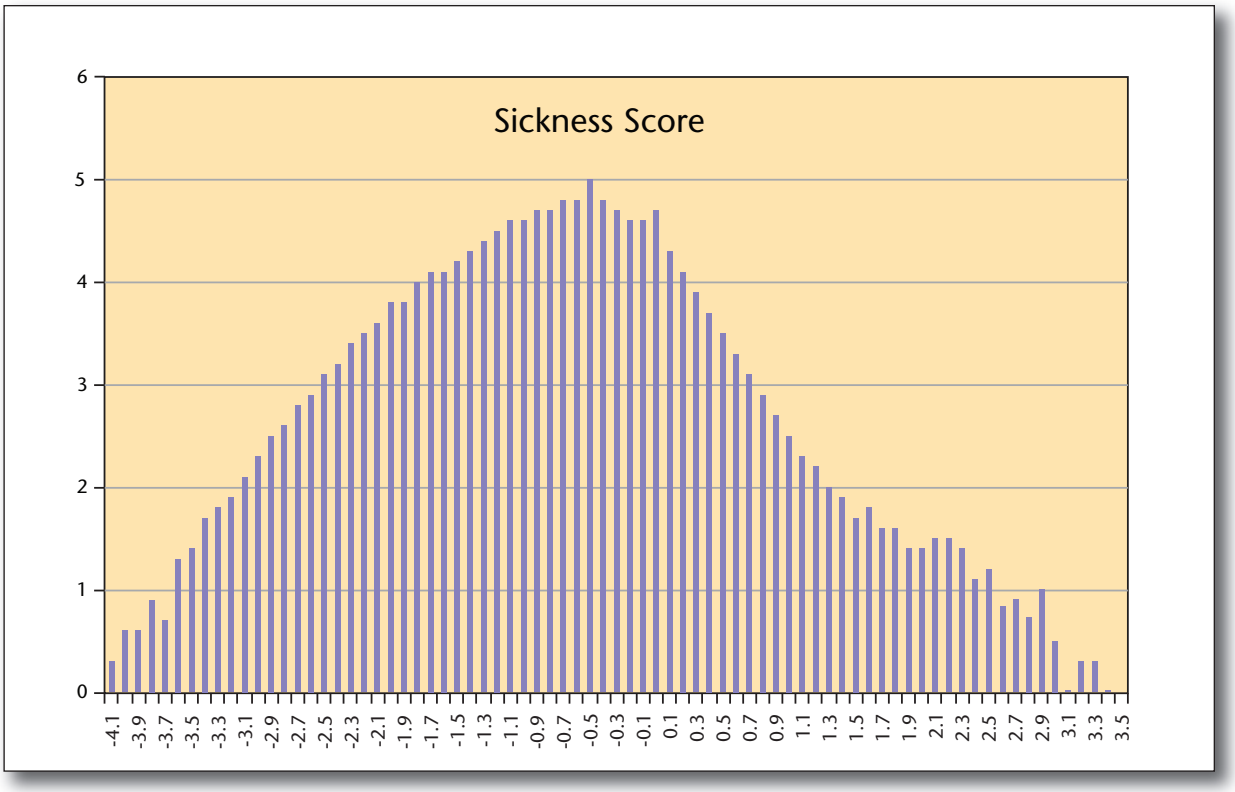

Figure 4. Distribution of Inferred Health Scores (Horizontal Axis) for One Week's Worth of Tweets.

The vertical axis shows the common logarithm of the number of messages with a particular health score. Higher scores indicate increased probability of being sick. Note that a tiny proportion of tweets (scores larger than 1.0) confidently show a foodborne illness.

Whereas this is not as important a limitation in terms of epidemiological surveillance, using surveillance data to actively intervene in outbreaks of foodborne illnesses can be challenging when surveillance data may not infrequently identify cases after the window of opportunity needed to prevent additional cases (Heymann 2004).

Methods

There are three general types of restaurant inspections conducted by health departments. First, restaurants are inspected prior to receiving a permit to ensure that the facility is designed and constructed in a way that allows food to be handled, prepared, and served in a safe manner. For example, inspections would ensure that food contact surfaces were durable and able to be easily cleaned, backflow prevention devices were installed in the plumbing system, and that commercial-grade appliances were installed. Once this type of inspection is completed for a facility, it would not be conducted again unless the facility was renovated.

The second, and most common, type of inspections are routine inspections. Routine inspections are not driven by the occurrence of problems, but are conducted periodically to prevent foodborne illness by ensuring that the facility is operating in accordance with good food-handling practices. Nevada law requires that these types of inspections happen at least annually. A routine inspection is a risk-based process addressing a food establishment's control over the five areas of risk for foodborne illness: personal hygiene, approved food source, proper cooking temperatures, proper holding times and temperatures, and sources of contamination.

A third type of inspection is a complaint-driven inspection initiated by either consumer complaints or the identification of a foodborne illness occurrence that may be associated with the facility. These inspections have a narrow focus but look in depth at a problem. For example, an inspection based on a complaint of improper handwashing at a restaurant would result in the inspector evaluating the handwashing facilities (that is, the availability of hand sinks, hot water, soap, and paper towels) and observing employees as they wash their hands, but would not result in a complete inspection of the facilities. If the inspection were related to foodborne illness, the

inspection would focus on the preparation of the particular foods consumed and the risk factors for the contamination, proliferation or amplification, and survival of the causative organism. This type of inspection is reactive in nature, and while it may prevent additional disease, problems in the facility have already occurred. The ultimate goal of all of these types of inspections is to prevent foodborne illness. Historically, there has been no way to easily identify restaurants having a decline in food handling practices and easily prevent illness, as inspections are based largely on the elapsed time from a previous inspection. As a result, these types of inspections represent the bulk of inspection activities but tend to be rather inefficient in identifying problem facilities. Complaint-driven inspections, while important, identify the problems after they have occurred, which is too late to prevent disease. More importantly, foodborne illnesses are frequently underdiagnosed and underreported (Scallan et al. 2011), preventing public health officials from identifying the source of illness for most foodborne infections.

Clark County, Nevada, is home to more than 2 million people and hosts over 41 million annual visitors to the Las Vegas metropolitan area. The Southern Nevada Health District (SNHD) is the governmental agency responsible for all public health matters within the county and is among the largest local health departments in the United States by population served. In 2014, SNHD conducted 35,855 food inspections (of all types) in nearly 16,000 permitted facilities. In Southern Nevada, inspection violations are weighted based on their likelihood to directly cause a foodborne illness and are divided into critical violations at 5 demerits each (for example, food handlers not washing hands between handling raw food and ready to eat food), to major violations at 3 demerits each (hand sink not stocked with soap), to good food management practices with no demerit value (leak at the hand sink). Demerits are converted to letter grades, where 0–10 is an A, 11–20 is a B, 21–39 is a C, and 40+ is an F (immediate closure). A repeated violation of a critical or major item causes the letter grade to drop to the next lower rank. A grade of C or F represents a serious health hazard.

Controlled Experiment: Adaptive Inspections

During the experiment, when a food establishment was flagged by nEmesis in an inspector's area, he was instructed to conduct a standard, routine inspection on both the flagged facility (adaptive inspection) and also a provided control facility (routine inspection). Control facilities were selected according to their location, size, cuisine, and their permit type to pair the facilities as closely as possible. The inspector was blind as to which facility was which, and each facility received the same risk-based inspection as the other.

Labeling Data at Scale

To scale the laborious process of labeling training data for our language model, we turn to Amazon's Mechanical Turk.[2] Mechanical Turk allows requesters to harness the power of the crowd in order to complete a set of human intelligence tasks (HITs). These HITs are then completed online by hired workers (Mason and Suri 2012).

We formulated the task as a series of short surveys, each 25 tweets in length. For each tweet, we ask "Do you think the author of this tweet has an upset stomach today?" There are three possible responses ("Yes," "No," "Can't tell"), out of which a worker has to choose exactly one (figure 5). We paid the workers 1 cent for every tweet evaluated, making each survey 25 cents in total. Each worker was allowed to label a given tweet only once. The order of tweets was randomized. Each survey was completed by exactly five workers independently. This redundancy was added to reduce the effect of workers who might give erroneous or outright malicious responses. Inter-annotator agreement measured by Cohen's κ is 0.6, considered a moderate to substantial agreement in the literature (Landis and Koch 1977). Responses from workers who exhibit consistently low annotator agreement with the majority were eliminated.

Workers were paid for their efforts only after we were reasonably sure their responses were sincere based on inter-annotator agreement. For each tweet, we calculate the final label by adding up the five constituent labels provided by the workers (Yes = 1, No = –1, Can't tell = 0). In the event of a tie (0 score), we consider the tweet healthy in order to obtain a high-precision data set.

Designing HITs to elicit optimal responses from workers is a difficult problem (Mason and Suri 2012). Pricing HITs poorly can lead to workers not even considering a task; HITs that are too long can cause worker attrition, poorly or ambiguously worded HITs will lead to noisy data. Worker satisfaction is also an important "latent" factor, which should not be taken lightly. Many Mechanical Turk workers are members of communities that offer requester reviews, very similar to Amazon's product review system. As a result, requesters who are unresponsive or opportunistic will soon find it hard to get any HIT completed.

Given that tweets indicating foodborne illness are relatively rare, learning a robust language model poses considerable challenges (Japkowicz et al. 2000; Chawla, Japkowicz, and Kotcz 2004). This problem is called class imbalance and complicates virtually all machine learning. In the world of classification, models induced in a skewed setting tend to simply label all data as members of the majority class. The problem is compounded by the fact that the minority class members (sick tweets) are often of greater interest than the majority class.

We overcome class imbalance faced by nEmesis

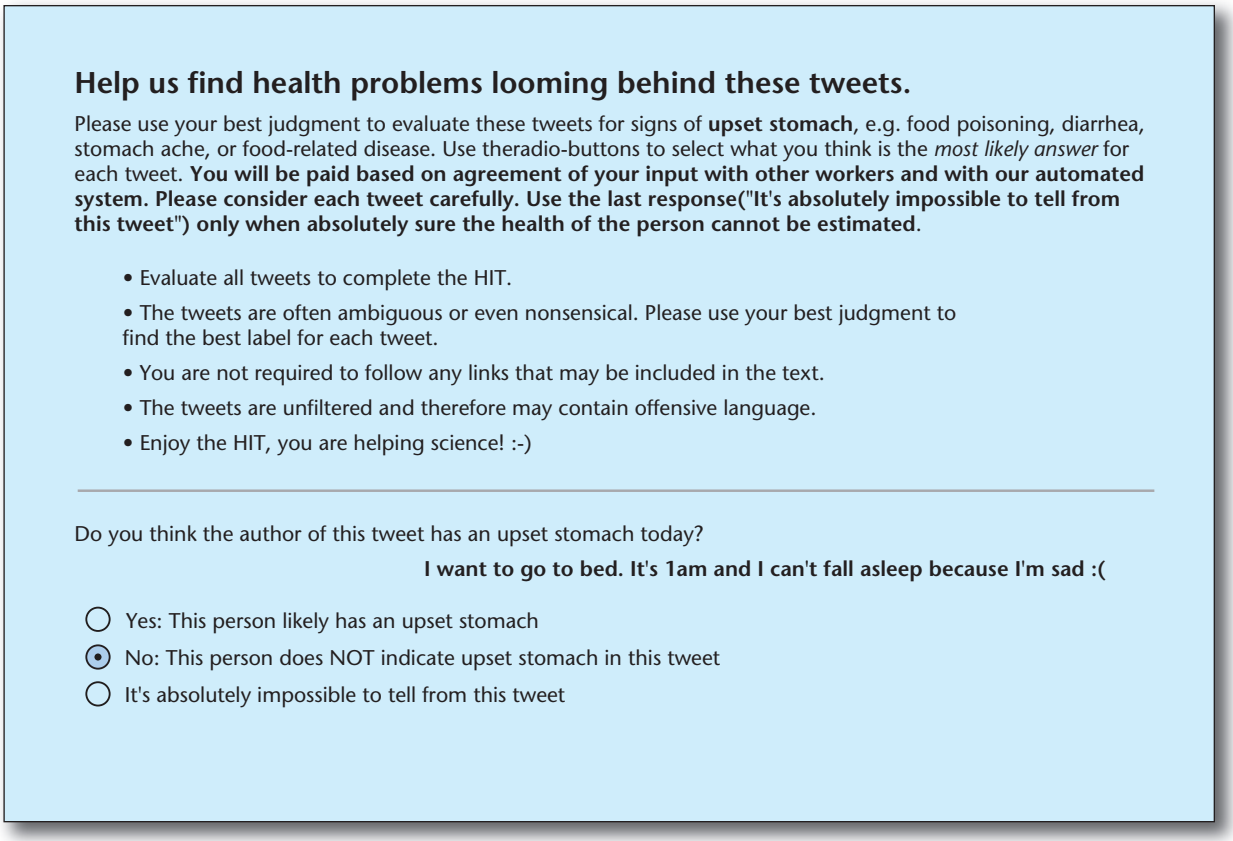

Figure 5. Example of a Mechanical Turk Task.

In this task, online workers are asked to label a given tweet. While tweets are often ambiguous, we encouraged workers to use their best judgment and try to polarize their answers. We found that when workers are presented with too many options, they tend to select "Can't tell" even when the text contains a strong evidence of illness.

through a combination of two techniques: human guided active learning, and learning a language model that is robust under class imbalance. We cover the first technique in this section and discuss the language model induction in the following section.

Previous research has shown that under extreme class imbalance, simply finding examples of the minority class and providing them to the model at learning time significantly improves the resulting model quality and reduces human labeling cost (Attenberg and Provost 2010). In this work, we leverage human guided machine learning — a novel learning method that considerably reduces the amount of human effort required to reach any given level of model quality, even when the number of negatives is many orders of magnitude larger than the number of positives (Sadilek et al. 2013). In our domain, the ratio of sick to healthy tweets is roughly 1 : 2500.

In each human guided learning iteration, nEmesis samples representative and informative examples to be sent for human review. As the focus is on the minority class examples, we sample 90 percent of tweets for a given labeling batch from the top 10 percent of the most likely sick tweets (as predicted by our language model). The remaining 10 percent is sampled uniformly at random to increase diversity. We use the HITs described above to obtain the labeled data.

In parallel with this automated process, we hire workers to actively find examples of tweets in which the author indicates he or she has an upset stomach. We asked them to paste a direct link to each tweet they find into a text box. Workers received a base pay of 10 cents for accepting the task, and were motivated by a bonus of 10 cents for each unique relevant tweet they provided. Each wrong tweet resulted in a 10 cent deduction from the current bonus balance of a worker. Tweets judged to be too ambiguous were neither penalized nor rewarded. Overall, we have posted 50 HITs that resulted in 1971 submitted tweets (mean of 39.4 per worker). Removing duplicates yielded 1176 unique tweets.

As a result, we employ human workers that "guide" the classifier induction by correcting the system when it makes erroneous predictions, and proactively seek-

ing and labeling examples of the minority classes. Thus, people and machines work together to create better models faster. This combination of human guided learning and active learning in a loop with a machine model has been shown to lead to significantly improved model quality (Sadilek et al. 2013).

In a postmortem, we have manually verified submitted tweets and 97 percent were correct sick tweets. This verification step could also be crowdsourced. Since searching for relevant tweets is significantly more time consuming than simply deciding if a given tweet contains a good example of sickness, future work could explore multitiered architecture, where a small number of workers acting as "supervisors" verify data provided by a larger population of "assistants." Supervisors as well as assistants would collaborate with an automated model, such as the support vector machine (SVM) classifier described in this paper, to perform search and verification tasks.

Language Model

Harnessing human and machine intelligence in a unified way, we develop an automated language model that detects individuals who likely suffer from a foodborne disease, on the basis of their online Twitter communication.

Support vector machines are an established method for classifying high-dimensional data (Cortes and Vapnik 1995). We train a linear binary SVM by finding a hyperplane with the maximal margin separating the positive and negative data points. Class imbalance, where the number of examples in one class is dramatically larger than in the other class, complicates virtually all machine learning. For SVMs, prior work has shown that transforming the optimization problem from the space of individual data points to one over pairs of examples yields significantly more robust results (Joachims 2005).

We use the trained SVM language model to predict how likely each tweet indicates foodborne illness. The model is trained on 8000 tweets, each independently labeled by five human annotators as described above. As features, the SVM uses all uni-gram, bigram, and tri-gram word tokens that appear in the training data at least twice. For example, a tweet "My tummy hurts" is represented by the following feature vector:

{my, tummy, hurts, my tummy, tummy hurts, my tummy hurts}

Prior to tokenization, we convert all text to lower case and strip punctuation. Additionally, we replace mentions of user identifiers (the "@" tag) with a special @ID token, and all web links with a @LINK token. We do keep hashtags (such as #upsetstomach), as those are often relevant to the author's health state, and are particularly useful for disambiguation of short or ill-formed messages.

Training the model associates a real-valued weight to each feature. The score the model assigns to a new tweet is the sum of the weights of the features that appear in its text. There are more than 1 million features; figure 2 lists the 20 most significant positive and negative features. While tweets indicating illness are sparse and our feature space has a very high dimensionality, with many possibly irrelevant features, support vector machines with a linear kernel have been shown to perform very well under such circumstances (Joachims 2006, Sculley et al. 2011, Paul and Dredze 2011a). Evaluation of the language on a held-out test set of 10,000 tweets shows 0.75 precision and 0.96 recall. The high recall is critical because evidence of illness is very scarce.

System Architecture

nEmesis consists of several modules that are depicted at a high-level in figure 3. Here we describe the architecture in more detail. We implemented the entire system in Python, with NoSQL data store running on Google Cloud Platform. Most of the code base implements data download, cleanup, filtering, snapping (for example, "at a restaurant"), and labeling ("sick" or "healthy"). There is also a considerable model-learning component described in the previous two sections.

Downloader

This module runs continuously and asynchronously with other modules, downloading all geo-coded tweets based upon the bounding box defined for the Las Vegas Metro area. These tweets are then persisted to a local database in JSON format.

Tracker

For each unique Twitter user that tweets within the bounding box, this module continues to download all of their tweets for two weeks, independent of location (also using the official Twitter API). These tweets are also persisted to local storage in JSON format.

Snapper

The responsibility of this module is to identify Las Vegas area tweets that are geo-coded within 50 meters of a food establishment. It leverages the Google Places API, which serves precise location for any given venue. We built an in memory spatial index that included each of those locations (with a square boundary based on the target distance we were looking for). For each tweet, nEmesis identifies a list of Google Places in the index that overlapped with the tweet based on its lat/long. If a given tweet had one or more location matches, the matching venues are added as an array attribute to the tweet.

Labeler

Each tweet in the data store is piped through our SVM model that assigns it an estimate of probability of foodborne illness. All tweets are annotated and saved back into the data store.

Aggregation Pipelines

We use Map Reduce framework on Google App

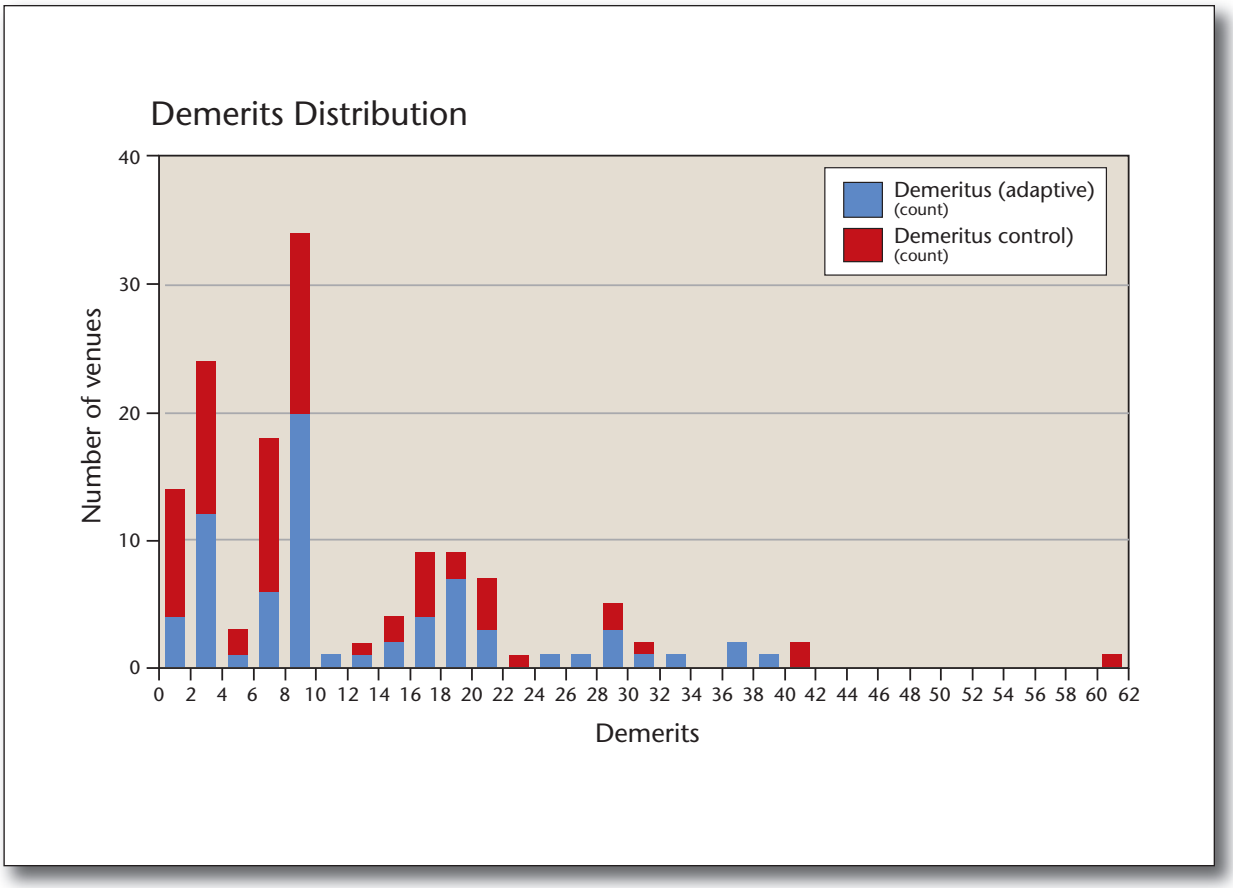

Figure 6. Histogram of the Inspection Results.

The adaptive inspections are blue (light gray), and the control inspections are red (dark gray). The horizontal axis is the number of demerits where the bucket size is 2, and the vertical axis is the number of venues.

Engine to support custom aggregation pipeline. It updates statistics about each venue (number of sick tweets associated with that venue, etc.).

Web Interface

The health professionals interact with nEmesis through a web application shown in figure 1. All modules described above work together to produce a unified view that lists most likely offending venues along with supporting evidence. This allows inspectors to make informed decisions how to allocate their resources. The application was written using a combination of Python for the data access layer and AngularJS for the front-end.

Developing the SVM model took 3 engineer-months. The backend modules above (Downloader through Labeler) took 2 engineer-months, and the Web Interface took an additional engineer-month.

Results and Discussion

Figure 6 is a histogram of the inspection results. There are clearly more control restaurants (red) that passed inspection with flying colors — zero or one demerit. The adaptive inspections (blue) appear to cluster toward the right — more demerits — but a careful statistical analysis is necessary to determine if this is really the case. We use paired Mann-Whitney-Wilcoxon tests to calculate the probability that the distribution of demerits for adaptive inspection is stochastically greater than the control distribution (Mann and Whitney 1947). This test can be used even if the shapes of the distributions are nonnormal and different, which is the case here. The test shows that adaptive inspections uncover significantly more demerits: nine versus six per inspection (p-value of 0.019).

Note that the result would have been even stronger if not for an outlier in the control group, a single control restaurant that received a score of 62 for egregious violations. Even including this outlier, however, we have very strong statistical evidence that adaptive inspections are effective.

Chi-squared test at the level of discrete letter grades (as noted earlier, 0–10 is an A, 11–20 is a B, 21–39 is a C, and 40+ is an F), also show a significant skew

toward worse grades in adaptive inspections. The most important distinction, however, is between restaurants with minor violations (grades A and B) and those posing considerable health risks (grade C and worse). nEmesis uncovers 11 venues in the latter category, whereas control finds only 7, a 64 percent improvement.

All of our data, suitably anonymized to satisfy Twitter's terms of use, is available upon request to other researchers for further analysis.

CDC studies show that each outbreak averages 17.8 afflicted individuals and 1.1 hospitalizations (CDC 2013). Therefore we estimate that adaptive inspections saved 71 infections and 4.4 hospitalizations over the three-month period. Since the Las Vegas health department performs more than 35,000 inspections annually, nEmesis can prevent over 9126 cases of foodborne illness and 557 hospitalizations in Las Vegas alone. This is likely an underestimate as an adaptive inspection can catch the restaurant sooner than a normal inspection. During that time, the venue continues to infect customers.

Adaptive inspections yield a number of unexpected benefits. nEmesis alerted SNHD to an unpermitted seafood establishment. This business was flagged by nEmesis because it uses a comprehensive list of food venues independent of the permit database. An adaptive inspection also discovered a food handler working while sick with an influenza-like disease. Finally, we observed a reduced amount of foodborne illness complaints from the public and subsequent investigations during the experiment. Between January 2, 2015, and March 31, 2015, SNHD performed 5 foodborne illness investigations. During the same time frame the previous year, SNHD performed 11 foodborne illness investigations. Over the last 7 years, SNHD averaged 7.3 investigations during this three-month time frame. It is likely that nEmesis alerted the health district to food safety risks faster than traditional complaint channels, prior to an outbreak.

Given the ambiguity of online data, it may appear hopeless to identify problematic restaurants fully automatically. However, we demonstrate that nEmesis uncovers significantly more problematic restaurants than current inspection processes. This work is the first to directly validate disease predictions made from social media data. To date, all research on modeling public health from online data measured accuracy by correlating aggregate estimates of the number of cases of disease based on online data and aggregate estimates based on traditional data sources (Grassly, Fraser, and Garnett 2005; Brownstein, Wolfe, and Mandl 2006; Ginsberg et al. 2008; Golder and Macy 2011; Sadilek et al. 2013). By contrast, each prediction of our model is verified by an inspection following a well-founded professional protocol. Furthermore, we evaluate nEmesis in a controlled double-blind experiment, where predictions are verified in the order of hours.

Finally, this study also showed that social-media-driven inspections can discover health violations that could never be found by traditional protocols, such as unlicensed venues. This fact indicates that it may be possible to adapt the nEmesis approach for identifying food safety problems in noncommercial venues, ranging from school picnics to private parties. Identifying possible sources of foodborne illness among the public could support more targeted and effective food safety awareness campaigns.

The success of this study has led the Southern Nevada Health District to win a CDC grant to support the further development of nEmesis and its permanent deployment statewide.

Acknowledgements

This research was partly funded by NSF grants 1319378 and 1516340; NIH grant 5R01GM108337-02; and the Intel ISTC-PC.

References

Achrekar, H.; Gandhe, A.; Lazarus, R.; Yu, S.; and Liu, B. 2012. Twitter Improves Seasonal Influenza Prediction. *Proceedings of the Fifth Annual International Conference on Health Informatics*. Setubal, Portugal: Institute for Systems and Technologies of Information, Control and Communication.

Anderson, R., and May, R. 1979. Population Biology of Infectious Diseases: Part I. *Nature* 280(5721): 361.

Attenberg, J., and Provost, F. 2010. Why Label When You Can Search?: Alternatives to Active Learning for Applying Human Resources to Build Classification Models Under Extreme Class Imbalance. In *Proceedings of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 423–432. New York: Association for Computing Machinery.

Brennan, S.; Sadilek, A.; and Kautz, H. 2013. Towards Understanding Global Spread of Disease from Everyday Interpersonal Interactions. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*. Menlo Park, CA: AAAI Press

Broniatowski, D. A., and Dredze, M. 2013. National and Local Influenza Surveillance Through Twitter: An Analysis of the 2012–2013 Influenza Epidemic. *PLoS ONE* 8(12): e83672. doi: 10.1371/journal.pone.0083672.

Brownstein, J.; Wolfe, C.; and Mandl, K. 2006. Empirical Evidence for the Effect of Airline Travel on Inter-Regional Influenza Spread in the United States. *PLoS Medicine* 3(10): e401. dx.doi.org/10.1371/journal.pmed.0030401

Brownstein, J. S.; Freifeld, B. S.; and Madoff, L. C. 2009. Digital Disease Detection — Harnessing the Web for Public Health Surveillance. *New England Journal of Medicine* 260(21): 2153–2157.

CDC. 2013. *Surveillance for Foodborne Disease Outbreaks United States, 2013: Annual Report*. Technical Report, Centers for Disease Control and Prevention National Center for Emerging and Zoonotic Infectious Diseases. Atlanta, GA: Centers for Disease Control and Prevention.

Chawla, N.; Japkowicz, N.; and Kotcz, A. 2004. Editorial: Special Issue on Learning from Imbalanced Data Sets. *ACM SIGKDD Explorations Newsletter* 6(1): 1–6.

Chen, P.; David, M.; and Kempe, D. 2010. Better Vaccination Strategies for Better People. In *Proceedings of the 11th ACM Conference on Electronic Commerce,* 179–188. New York: Association for Computing Machinery.

Chunara, R.; Andrews, J.; and Brownstein, J. 2012. Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *The American Journal of Tropical Medicine and Hygiene* 86(1): 39–45.

Cortes, C., and Vapnik, V. 1995. Support-Vector Networks. *Machine Learning* 20(3): 273–297.

Culotta, A. 2010. Towards Detecting Influenza Epidemics by Analyzing Twitter Messages. Paper presented at the First Workshop on Social Media Analytics, July 25–28, Washington DC.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting Depression

via Social Media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media,* 128–137. Palo Alto, CA: AAAI Press.

Eubank, S.; Guclu, H.; Anil Kumar, V.; Marathe, M.; Srinivasan, A.; Toroczkai, Z.; and Wang, N. 2004. Modelling Disease Outbreaks in Realistic Urban Social Networks. *Natur*e 429(6988): 180–184.

FDA. 2012. *Bad Bug Book*. U.S. Food and Drug Administration, 2nd ed. Silver Spring, MD: U.S. Food and Drug Administration.

Ginsberg, J.; Mohebbi, M.; Patel, R.; Brammer, L.; Smolinski, M.; and Brilliant, L. 2008. Detecting Influenza Epidemics Using Search Engine Query Data. *Nature* 457(7232): 1012–1014.

Golder, S., and Macy, M. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science* 333(6051): 1878–1881.

Grassly, N.; Fraser, C.; and Garnett, G. 2005. Host Immunity and Synchronized Epidemics of Syphilis Across the United States. *Nature* 433(7024): 417–421.

Grenfell, B.; Bjornstad, O.; and Kappey, J. 2001. Travelling Waves and Spatial Hierarchies in Measles Epidemics. *Nature* 414(6865): 716–723.

Harrison, C.; Jorder, M.; Stern, H.; Stavinsky, F.; Reddy, V.; Hanson, H.; Waechter, H.; Lowe, L.; Gravano, L.; and Balter, S. 2014. Using a Restaurant Review Website to Identify Unreported Complaints of Foodborne Illness. *Morbidity and Mortality Weekly Report* 63(20): 441–445.

Heymann, D. L. 2004. C*ontrol of Communicable Diseases Manual: A Report of the American Public Health Association* 18th edition. Washington, DC: American Public Health Association.

Japkowicz, N., et al. 2000. Learning from Imbalanced Data Sets: A Comparison of Various Strategies. In *Learning from Imbalanced Data Sets: Papers from the AAAI Workshop*. Technical Report WS-00-05. Palo Alto, CA: AAAI Press.

Joachims, T. 2005. A Support Vector Method for Multivariate Performance Measures. In *Proceedings of the Twenty-Second International Conference on Machine Learning,* 377–384. New York: Association for Computing Machinery.

Joachims, T. 2006. Training Linear Svms in Linear time. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 217–226. New York: Association for Computing Machinery.

Landis, J. R., and Koch, G. G. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33(1): 159–174.

Lane, N. D.; Miluzzo, E.; Lu, H.; Peebles, D.; Choudhury, T.; and Campbell, A. T. 2010. A Survey of Mobile Phone Sensing. *IEEE Communications Magazine* 48(9): 140–150.

Mann, H., and Whitney, D. 1947. On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other. *Annals of Mathematics and Statistics* 18(1): 50–60.

Mason, W., and Suri, S. 2012. Conducting Behavioral Research on Amazon's Mechanical Turk. *Behavior Research Methods* 44(1): 1–23.

Morris, J. G., and Potter, M. 2013. *Foodborne Infections and Intoxications,* 4th ed. Amsterdam: Elsevier Science.

Newman, M. 2002. Spread of Epidemic Disease on Networks. *Physical Review* E 66(1): 016128.

Paul, M., and Dredze, M. 2011a. A Model for Mining Public Health Topics from Twitter. Unpublished paper, Johns Hopkins University..

Paul, M., and Dredze, M. 2011b. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Palo Alto, CA: AAAI Press.

Sadilek, A., and Kautz, H. 2013. Modeling the Impact of Lifestyle on Health at Scale. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. New York: Association for Computing Machinery.

Sadilek, A.; Brennan, S.; Kautz, H.; and Silenzio, V. 2013. nEmesis: Which Restaurants Should You Avoid Today? In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing,* 138–146. Palo Alto, CA: AAAI Press.

Sadilek, A.; Kautz, H.; and Silenzio, V. 2012. Predicting Disease Transmission from Geo-Tagged Micro-Blog Data. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Scallan, E.; Hoekstra, R. M.; Angulo, F. J.; Tauxe, R. V.; Widdowson, M. A.; and Roy, S. L. 2011. Foodborne Illness Acquired in the United States — Major Pathogens. *Emerging Infectious Diseases*. 17(1): 7–15. doi: 10.3201/eid1701.P11101

Scharff, R. L. 2012. Economic Burden from Health Losses Due to Foodborne Illness in the United States. *Journal of Food Protection* 75(1): 123–131.

Sculley, D.; Otey, M.; Pohl, M.; Spitznagel, B.; Hainsworth, J.; and Yunkai, Z. 2011. Detecting Adversarial Advertisements in the Wild. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: Association for Computing Machinery.

Snow, J. 1855. *On the Mode of Communication of Cholera*. London: John Churchill.

Ugander, J.; Backstrom, L.; Marlow, C.; and Kleinberg, J. 2012. Structural Diversity in Social Contagion. *Proceedings of the National Academy of Sciences* 109(16): 5962–5966. Washington, DC: National academy of Sciences of the United States of America.

White, R., and Horvitz, E. 2008. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. Technical Report MSR-TR-2008-177, Microsoft Research. Appearing in *ACM Transactions on Information Systems,* 27(4), Article 23, November 2009, DOI 101145/1629096. 1629101.

Widener, M. J., and Li, W. 2014. Using Geolocated Twitter Data to Monitor the Prevalence of Healthy and Unhealthy Food References Across the US. *Applied Geography* 54(October): 189–197.

Adam Sadilek is a senior engineer at Google.

Henry Kautz is a professor in the Department of Computer Science at the University of Rochester.

Lauren DiPrete is a senior environmental health specialist at the Southern Nevada Health District.

Brian Labus is a visiting research assistant professor in the School of Community Health Sciences at the University of Nevada, Las Vegas.

Eric Portman is a consulting research engineer in greater Atlanta Georgia.

Jack Teitel is an undergraduate student at the University of Rochester.

Vincent Silenzio, M.D., is an associate professor at the School of Medicine and Dentistry at the University of Rochester.