Articles and Letters

Digital Disease Detection with Big Data

Samuel Kou, Harvard University
kou@stat.harvard.edu

Communicated by Susan A. Murphy

This article summarizes key ideas from the joint European Mathematical Society and Bernoulli Society lecture. Recent advances on using big data to detect and track infectious diseases are discussed along with insights we learned from them.

§1. Big Data and Disease Detection

The wide availability and growth of Internet and online platforms have profoundly transformed our society, from the daily lives of individuals, to the way business is run, to the interactions and communications between individuals, companies and governments. As people do Google search, use Facebook, Twitter, Instagram, etc., big data sets that collect the footprints of millions of online users are constantly generated. These big data contain information of the users' activities in nearly every aspect of life. They also offer the potential to transform decision making in industry, business, social policy and public health (Khoury and Ioannidis, 2014; Kim et al., 2014; McAfee and Brynjolfsson, 2012).

One area that has received recent attention is to use big data to track infectious diseases, which affect tens of millions of people worldwide. For instance, influenza causes about 500,000 death per year worldwide and about 3,000 to 50,000 per year in the US; dengue fever infects about 390 million people, causing up to 20,000 deaths per year worldwide. Accurate and reliable tracking and forecasting of infectious diseases can help public health officials and government agencies prepare and allocate resources for potential outbreaks, improve risk assessment and communication, issue warnings, and take preventive actions. Traditional disease surveillance tracks disease activity through patients' clinical visits or doctors' field diagnosis. However, owing to the time needed for processing and aggregating clinical information, the clinical-based surveillance often lags behind real time by weeks, which is far from optimal. In the case of influenza surveillance, the US Centers for Disease Control and Prevention (CDC)'s influenza-like illness report, which tracks influenza in the US, often have a delay of one to two weeks.

Big data generated from the Internet present the opportunity for *real-time* disease surveillance and tracking. For example, the surge of influenza (flu) related online search queries in a short time period, such as "flu symptoms", "flu treatment", "flu medicine", etc., can indicate a potential flu outbreak. The ubiquity of big data and that they track social behavior and trends in real time make it possible and attractive to build such a digital disease detection system.

§2. The Rise and Fall of Google Flu Trends

In November 2008, Google launched Google Flu Trends (GFT), which uses the volume of selected Google search terms to estimate current influenza-like illnesses (ILI) activity (Ginsberg et al., 2009). The introduction of GFT generated much excitement, and GFT was welcomed and identified by many as a good example of how big data would transform traditional predictive analysis (Helft, 2008). However, significant discrepancy between GFT's flu estimates and those measured by CDC started to emerge from May 2009. The discrepancy grew overtime, and by late August 2009, GFT underestimated the flu activity by more than 50% — CDC's true number was close to 5%, whereas GFT's estimate was about 2%. The subsequent revision of GFT (Cook et al., 2011) did not do much better. In January 2013, GFT overestimated the flu activity by more than 100% — CDC's true number was around 4.6%, whereas GFT's estimate was over 10%.

Close inspection of the original GFT algorithm reveals several limitations (Lazer et al., 2014; Santillana et al., 2014; Yang et al., 2015). First, GFT assumed that the relationship between total search volume and the proportion of people getting flu was static, but in reality people's search pattern changes over time, so does Google's search engine and the interaction between people and the search engine. Second, GFT did not take use of newly available ILI activity reports from CDC as a flu season evolves, even though CDC's reports contain crucial information of the severity of the flu season. Third, the idea of aggregating the search volumes of multiple query terms into a single variable as the predictor in the GFT model did not allow for changes in people's Internet search behavior over time to be accounted for (as the relative importance of

individual search terms changes over time). Fourth, GFT ignored the intrinsic time series properties of flu, including the seasonality of flu activity. Amid the promises and challenges, Google discontinued GFT in August 2015.

§3. ARGO for Digital Flu Detection

The mishap of GFT leads people to question the value and feasibility of digital disease detection systems (Butler, 2013). We started in 2014 to investigate if it is possible to build such a system that is capable of generating accurate and reliable real-time tracking of infectious diseases. In the case of flu, we found that it is in fact possible to build an accurate real-time digital flu detection system by using Internet search data, and introduced our model ARGO in Yang et al. (2015), which stands for AutoRegression with GOogle search data.

ARGO starts from a hidden Markov structure, postulating that (i) the prediction target, the CDC's ILI percentage, which measures the percentage of patients having flu like symptoms in a given week, follows an autoregressive (AR) model with lag N, after logit-transformation. Let p_t denote CDC's ILI percentage that we want to predict at week t (ahead of the time-delayed official report). Then ARGO assumes that $y_t = \log(p_t/(1-p_t))$ follows

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

ARGO also postulates that (ii) the vector of normalized search volumes of flu related query terms on Google at time t depends only on the flu activity at the same time. This assumption reflects the intuition that flu occurrence leads people to search flu related information online. Let $X_{i,t}$ be the log-transformed normalized Google search volume of query term i at week t, and let $\boldsymbol{X}_t = (X_{1,t}, X_{2,t}, \ldots, X_{K,t})$, where K is the total number of query terms under consideration. Then ARGO assumes that

$$\boldsymbol{X}_t \mid y_{1:t} \sim \mathcal{N}_K\left(\boldsymbol{\mu}_x + y_t\boldsymbol{\beta}, \boldsymbol{Q}\right).$$

In our study, we used more than 100 search query terms ($K \geq 100$) and obtained the relative search volume of each query term from the publicly available *Google Trends* website, which gives us $\boldsymbol{X}_t$ in real time. The query terms included "symptoms of flu", "treating flu", "flu duration", "flu vs cold", "flu contagious", etc. See Yang et al. (2015) for the detailed list of query terms.

For predicting y_t at week t given the time delayed CDC reports and the up-to-date (relative) Google search volumes of the query terms, we calculate the predictive distribution $f(y_t|y_{1:(t-1)}, \boldsymbol{X}_{1:t})$, which is normal with mean linear in $y_{(t-N):(t-1)}$ and $\boldsymbol{X}_t$. This leads to the predictive equation

$$\hat{y}_t = \mu_y + \sum_{j=1}^{N} \alpha_j y_{t-j} + \sum_{i=1}^{K} \beta_i X_{i,t}.$$

To train the ARGO model, we take $N = 52$, which captures the within-year seasonality of flu activity. The coefficients μ_y, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$ are obtained by minimizing

$$\sum_t (y_t - \mu_y - \sum_{j=1}^{N} \alpha_j y_{t-j} - \sum_{i=1}^{K} \beta_i X_{i,t})^2 + \lambda_\alpha \|\boldsymbol{\alpha}\|_1 + \lambda_\beta \|\boldsymbol{\beta}\|_1,$$

where λ_α and λ_β are hyper-parameters. The training of ARGO has several features. (a) A two-year moving window that immediately precedes the desired date of estimation is used for the training period. This moving window estimation is to capture the most recent changes in people's search patterns, reflecting the fact that the search pattern and search engine both evolve over time. (b) In the two-year moving window, since there are more independent variables (> 150) than the number of observations ($= 104$), L_1 penalty is used, which serves to select the most useful Google search queries for estimation. (c) The estimation dynamically incorporates the most recent information from the CDC reports as they become available. (d) The time series terms in the predictive equation helps capture the long-term cyclic information (seasonality) from past flu activity.

Figure 1 shows (in red) the prediction of ARGO for the flu activity (measured by the CDC's ILI activity level) for the time period of March 2009 to July 2015, compared to the ground truth (in black): the CDC-reported ILI activity level, published typically one or two weeks later. Also shown in Figure 1 (in green) is the GFT estimates (showing the latest updated GFT estimates) for the same time period. The lower panel shows the prediction error. ARGO's prediction stayed close to the ground truth throughout the period. It evidently outperformed GFT. Zooming in on the individual years, ARGO significantly outperformed GFT in each flu season. In fact, it also significantly outperforms other alternative methods as detailed in Yang et al. (2015). The prediction result demonstrates ARGO as an effective method to harness information from Internet searches to provide accurate and reliable real-time tracking of flu.

§4. ARGO for Tracking Dengue

Dengue is a mosquito-borne disease that threatens an estimated 3.9 billion people in 128 countries with an estimated 390 million infections each year. To reduce dengue mortality and morbidity, timely identification of outbreaks is critical, as it can inform and help preventative measures, such as mosquito population control and mosquito bite prevention. For dengue surveillance, governments traditionally rely on hospital-based reporting, a method that is often lagged and limited with frequent post-hoc revisions, due to communication inefficiencies among local and national agencies and the time needed to aggregate information to the state level.

Encouraged by the promising result of using ARGO to track flu, we extended it for dengue tracking in Yang et al. (2017a), and tested it on producing near real-time estimates of dengue cases in five countries/states: Brazil, Mexico, Singapore, Thailand and Taiwan. The basic idea, similar to the case of flu tracking, is that dengue-related Google search volumes can indicate the ups and downs of dengue activity. The five countries/states were chosen to explore the applicability of ARGO in a diverse set of situations, as the five have different ecology, size, population, economic de-

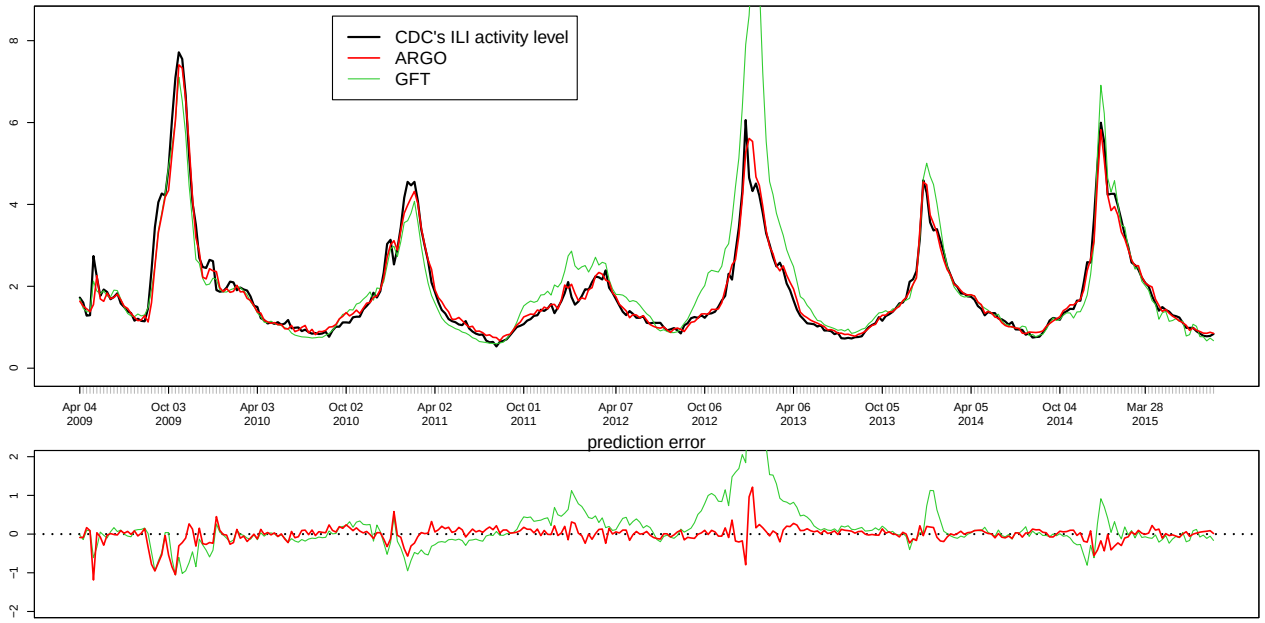

Figure 1: ARGO prediction of the flu level (red) in comparison to the ground truth, CDC's ILI activity level (black), as well as the latest updated estimates from GFT (green). The lower panel shows the estimation error.

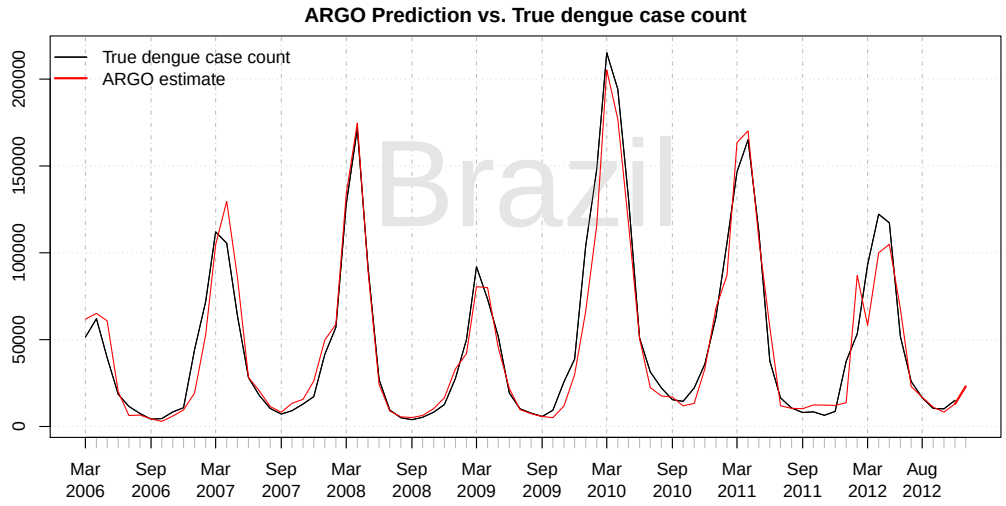

Figure 2: ARGO one-month-ahead prediction of the dengue case counts (red) in Brazil, compared to the official counts (black).

velopment level, Internet penetration (Singapore and Taiwan over 75% versus Thailand's 27%), and Google market share (Brazil, Mexico and Thailand over 90% versus Taiwan's 42% in 2012). We apply ARGO to forecast the monthly dengue case counts in the five countries/states, as the official dengue counts are only available in three countries at the monthly level.

Let $y_t = \log(c_t + 1)$ be the log-transformed dengue case counts c_t at time t, and $X_{k,t}$ the log-transformed (relative) Google search volume of query term k at time t. The query terms we used included "dengue symptoms", "dengue fever", "mosquito bites", etc. (see Yang et al. (2017a)). The hidden Markov structure of the ARGO model gives

$$y_t = \mu_y + \sum_{j\in J} \alpha_j y_{t-j} + \sum_{k\in K} \beta_k X_{k,t} + \epsilon_t,\ \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

where J is the set of auto-regressive lags, and K is the set of Google query terms. For forecasting the monthly dengue case counts, we took $J = \{1, \ldots, 12\} \cup \{24\}$, i.e., the most recent 12 months plus the month exactly two years ago. The coefficients μ_y, $\boldsymbol{\alpha} = \{\alpha_j : j \in J\}$, and $\boldsymbol{\beta} = \{\beta_k : k \in K\}$ are obtained by minimizing over a two-year moving window

$$\sum_t \left(y_t - \mu_y - \sum_{j\in J} \alpha_j y_{t-j} - \sum_{k\in K} \beta_k X_{k,t} \right)^2 + \sum_{j\in J} \lambda_{\alpha_j} |\alpha_j| + \sum_{k\in K} \lambda_{\beta_k} |\beta_k|$$

where λ_{α_j} and λ_{β_k} are regularization hyper-parameters.

Figure 2 shows (in red) the ARGO prediction of dengue case counts in Brazil, one month ahead of the official case counts (shown in black) for the period of March 2006 to December 2012. A close agreement between ARGO prediction and the true counts is seen. In fact, ARGO outperformed other alternative methods as well. The full results and methodology details (including the specification of the hyper-parameters) are described in Yang et al. (2017a). The encouraging results show that the ARGO modeling framework can be used to improve the tracking of dengue activity in multiple locations around the world and that it can be a useful tool to help governments/public health agencies to prepare for and cope with potential dengue outbreaks.

§5. Road Ahead in the Big Data World

The results of using ARGO to track flu and dengue suggests its versatility. It can be potentially applied to track other infectious diseases, such as Zika, malaria, yellow fever and Chikungunya. As long as a sizable proportion of the population do Internet search, in principle the aggregated Internet search information can be utilized to track disease activities in real time. ARGO can be deployed for such purposes. Such tools that harness information from big data generated from the Internet can be particularly helpful for less developed countries where government-led hospital-based disease surveillance systems are lacking or ineffective.

ARGO can be generalized to other temporal or spacial scales. It can also incorporate other sources of information. One such source is electronic health records. Over the last two decades many hospitals and medical centers have adopted electronic health records (EHR) to give clinicians faster and easier access to retrieve, enter and modify patient information. The cloud-based EHR systems facilitate real-time retrieval of aggregated disease information. Yang et al. (2017b) extended ARGO to include EHR as well as Internet search data for flu tracking: the predictors $\boldsymbol{X}_t$ in the ARGO model now include variables derived from nationally aggregated flu-related patient visit counts from a cloud-based EHR system, in addition to Google search volumes. Further error reduction was achieved (Yang et al., 2017b).

The emergence of big data from online or cloud systems offers the potential for real-time tracking of social or public health events. Equally important is the development of statistical/mathematical models and methods that are capable to effectively extract information from the digital data sources and produce accurate and reliable predictions. In fact, GFT was criticized not because people do not believe the value of Internet search data, but because the predictions from GFT were misleading due to its methodological flaws to process the valuable information. Effective use of big data raises many interesting and challenging methodological questions. If not handled properly, they can lead to very inaccurate results. The failure to predict the outcomes of Brexit and 2016 US presidential election despite the large amount of data from social media is a vivid reminder.

Big data present big opportunities for predictive analysis and decision making, but only with proper and rigorous methods and reasoning can the potential be unleashed.

Acknowledgement. The material summarized in this article is based on fruitful collaborations with Mauricio Santillana and Shihao Yang.

References

Butler D. (2013), "When Google got flu wrong," *Nature*, 494, 155–156.

Cook S., et al. (2011), "Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic," *PLOS One*, 6, e23610.

Ginsberg, J., et al. (2009), "Detecting influenza epidemics using search engine query data," *Nature*, 457, 1012–1014.

Helft M. (November 11, 2008), "Google uses searches to track flu's spread," *New York Times*.

Khoury, M., and Ioannidis, J. (2014), "Big data meets public health," *Science*, 346, 1054–1055.
Kim, G., Trimi, S., and Chung, J. (2014), "Big-data applications in the government sector," *Communications of the ACM*, 57, 78–85.
Lazer D., et al. (2014), "The parable of Google Flu: Traps in big data analysis," *Science*, 343, 1203–1205.
McAfee, A., and Brynjolfsson, E. (2012), "Big data: the management revolution," *Harvard business review*, 90, 60–68.
Santillana M., et al. (2014), "What can digital disease detection learn from (an external revision to) Google Flu Trends?" *Am. J. Prev. Med.*, 47, 341–347.
Yang, S., Santillana, M., and Kou, S.C., (2015), "Accurate estimation of influenza epidemics using Google search data via ARGO," *Proc. Natl. Acad. Sci. USA*, 112, 14473–14478.
Yang, S., et al. (2017a), "Advances in using Internet searches to track dengue," *PLOS Comput. Biol.*, 13, e1005607.
Yang, S., et al. (2017b), "Using electronic health records and Internet search information for accurate influenza forecasting," *BMC Infect. Dis.*, 17, 332.

Past Conferences, Meetings and Workshops

Sponsored and Co-Sponsored by Bernoulli Society for Mathematical Statistics and Probability

Statistics Meets Friends: Nov, 29–Dec, 1, 2017; Göttingen, Germany

The conference *Statistics Meets Friends* was held in Göttingen from Nov. 29th to Dec. 1st, 2017, at the Alte Mensa of the University of Göttingen. It was organized by the members of the scientific committee Timo Aspelmeier, Thorsten Hohage, Stephan Huckemann, Andrea Krajina, Tatyana Krivobokova, Johannes Schmidt-Hieber and Frank Werner, on the occasion of Axel Munk's 50th birthday under the motto "*from biophysics to inverse problems and back*", bridging the gap between mathematical statistics, inverse problems and biophysics, highlighting recent developments at their interfaces. There were approximately 100 participants and the conference featured 25 high quality invited talks by the following renowned scientists:

Rabindra N. Bhattacharya on "*Monotone random dynamical systems: Existence of steady states and convergence*", Peter Bühlmann on "*AAA*", Tony Cai on "*Rate-optimal perturbation bounds for singular subspaces with applications to high-dimensional data analysis*", Emmanuel Candès on "*The likelihood ratio test in high-dimensional logistic regression is not a chi-square*", Manfred Denker on "*Improving statistical decision procedures*", Holger Dette on "*Relevant change points in high dimensional time series*", Lutz Dümbgen on "*Simultaneous inference about features of densities and regression functions*", Alexander Egner on "*Optical nanoscopy and statistics: Towards the optimum resolution*", Markus Grasmair on "*Convergence rates for multiresolution based regularisation methods*", Helmut Grubmüller on "*Structure determination from single molecule X-ray scattering with three photons per image*", Markus Haltmeier on "*Compressed sensing and sparsity in photoacoustic tomography*", Marc Hoffmann on "*Nonparametric estimation of an inhomogeneous age-dependent model in a large population limit*", Chris Holmes on "*Probabilistic decision functions*", Hajo Holzmann on "*Inverse problems in econometrics*", Thomas Hotz on "*Statistics in circles*", Zakhar Kabluchko on "*Convex cones and statistics*", Bernard A. Mair on "*From positron emission tomography to potential theory and back*", Enno Mammen on "*Nonparametric estimation of locally stationary Hawkes processes*", Victor M. Panaretos on "*Nearly blind deconvolution of Gaussian processes*", Richard Samworth on "*Isotonic regression in general dimensions*", David O. Siegmund on "*Detection and estimation of local signals*", Sara van de Geer on "*On the asymptotic variance of the de-biased Lasso*", Aad van der Vaart on "*Credible sets for sparse models*", and Harrison Huibin Zhou on "*Theoretical and computational guarantees on meanfield variance Bayes method for community detection*".

Richard Nickl held the Ethel–Newbold-Prize Lec-