Launching the next generation of digital disease surveillance tools

Mauricio Santillana

*Faculty member, CHIP, Boston Children’s Hospital Informatics Program*

*Associate, Harvard Institute for Computational and Applied Sciences*

*Instructor, Harvard Medical School*
The promise of big data in public health

GOOGLE FLU TRENDS
Google.org Flu Trends

Epidemiological information available 2-3 weeks ahead of traditional clinical tracking systems

![Graphs showing ILI percentages estimated by the model and provided by CDC](chart)

**Figure 3**: ILI percentages estimated by our model (black) and provided by CDC (red) in the Mid-Atlantic region, showing data available at four points.

- **Data available as of February 4, 2008**
- **Data available as of March 3, 2008**
- **Data available as of March 31, 2008**
- **Data available as of May 12, 2008**

Each candidate search query was evaluated nine times, once per region, using the search data originating from a particular major city if within the United States.

Our database of queries contains 50 million Google web search logs. Google web search logs can provide one of the most timely, broad reaching influenza monitoring systems available today.

While traditional systems require 1-2 weeks to gather and process surveillance data, our estimates are current each day. As with other syndromic surveillance systems, the data are most useful as a means to spur further investigation andcollection of direct measures of disease activity.

Methods

In the query selection process, we fit per-query models using all weeks between September 28, 2003 and March 11, 2007 (inclusive) for which CDC reported a non-zero ILI percentage, yielding 128 training points for each region. For each region, we obtained 36 different models to test each of the candidate queries. We used a distributed computing framework for this project.

None of the queries in our project's database can be excluded. Using the internet protocol (IP) address associated with a Google search user; we don't combine linguistic variations, synonyms, cross-language translations, misspellings, or original web search logs older than 9 months are being excluded. Using the internet protocol (IP) address associated with a particular individual. Our project's database includes the ILI percentage, yielding 128 training points for each region. With four cross-validation folds per region, we obtained 36 different models to test each of the candidate queries. We used a distributed computing framework for this project.

At Google, we recognize that privacy is important. None of the queries in our project's database can be linked to a particular individual's search history. Our project's database is anonymous in accordance with Google's Privacy Policy and exclusively used for research purposes. Our project's database contains data, but retains no information about the identity, IP address, or specific physical location of any user. Furthermore, any identifiable data are most useful as a means to spur further investigation andcollection of direct measures of disease activity.

Each candidate search query was evaluated nine times, once per region, using the search data originating from a particular major city if within the United States. Results are freely available online at http://www.google.org/flutrends.
Letter

Nature 457, 1012-1014 (19 February 2009) | doi:10.1038/nature07634; Received 14 August 2008; Accepted 13 November 2008; Published online 19 November 2008; Corrected 19 February 2009

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

1. Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA
2. Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA

Correspondence to: Matthew H. Mohebbi¹ Correspondence and requests for materials should be addressed to J.G. or M.H.M. (Email: flutrends-support@google.com).
Very promising retrospective comparison!

In April 2009, Dr. Brilliant said it epitomized the power of Google’s vaunted engineering prowess to make the world a better place, and he predicted that it would save untold numbers of lives.
Google Flu Trends

launched in November 2008
Real-time performance, first year...

Big errors seen during H1N1 pandemic (off-season)

To some extent GFT was good at predicting seasons: fall-winter, not flu!
What next?
need to remove (not useful) search terms

Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html
What next?

need to remove (not useful) search terms

Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html
What next? need to remove (not useful) terms.

Big discrepancies again!

Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html
When Google got flu wrong.
nature.com/news/when-google-got-flu-wrong.
Snowden And The Challenge Of Intelligence: The Practical Case Against The NSA's Big Data

"We should soon be able to keep track of most activities on the surface of the earth, day or night, in good weather or bad.

Can Nate Silver’s Data Culture Lead Us Out of the NSA + Public Data Scare?

ryan cox | september 18th
Lessons learned
Let’s work on a short exercise to understand how Google Flu Trends used to work...
Supervised machine learning examples:

Given the number of Google searches associated to the term “dengue”, and given the number of confirmed cases of dengue in Mexico from 2004 to 2006 (Training period), can we estimate how many people will most likely get dengue based on the number of searches during the subsequent years?
4. Least squares in Public Health. (30 points)
Dengue fever is a virus-caused disease that is spread by mosquitoes that affects millions of people in tropical environments around the Globe. In this problem, you are asked to construct a simple version of the digital disease detection tool: “Google Dengue Trends” for Mexico. For this, you will download the spreadsheet Dengue trends AM 111.xls from the course website. The first column in the spreadsheet represents the date (in months, from 2004-2011), the second column represents the number of Google searches of the term “dengue” in Mexico, in a given month. The third column represents the number of cases of Dengue in Mexico, as reported by the Mexican Ministry of Health. You may use Matlab or Excel for this problem.

(a) Plot the number of cases of Dengue as a function of time.

(b) For the training period 2004-2006 (36 months), find the best line that explains the number of cases of Dengue as a function of the number of searches of the term “dengue”. You should do this by solving the least squares problem, and you should obtain the value of the y-intercept and the slope.

(c) Use the equation of the line you obtained in (b) and plot the number of cases as a function of the number of searches of the term “dengue”, predicted by your method during the training period. Compare your results to the plot in (a) for such time period.

(d) For the prediction or validation period 2007-2011, use the equation of the line you obtained in (b) to predict the number of the dengue cases as a function of the number of searches of the term “dengue” from 2007-2011. Plot your predictions and compare them to the actual number of cases.

(d) Discuss your results. Could you improve this modeling approach? If so, how?
Did you get it to work?
Using Google searches to track diseases statically

begin

%% Load data %%
CDC=load(CDC ILI Data)   \hspace{1cm} \textbf{(One column of values)}
X=load(Google search Data)   \hspace{1cm} \textbf{(Multiple columns of values)}

%% initialize output array %%
Y=zeros(1: end.of.predictions)   \hspace{1cm} \textbf{(Initialize array to store predictions)}

%% train model and produce predictions %%
CDC $\leftarrow$ standardize(CDC)   \hspace{1cm} \textbf{(Perhaps use a transform: z-score, logit)}
X $\leftarrow$ standardize(X)   \hspace{1cm} \textbf{(Perhaps use a transform: z-score, logit)}
model=LASSO.routine.fit(CDC[1 : training] $\sim$ X[1 : training]) \hspace{1cm} \textbf{(Training: in-sample model)}

Y[1 : training]=
  LASSO.routine.predict(model, X[1 : training]) \hspace{1cm} \textbf{(In-sample predictions)}

Y[training + 1 : end.of.predictions]=
  LASSO.routine.predict(model, X[training + 1 : end.of.predictions]) \hspace{1cm} \textbf{( Produce out-of-sample predictions )}

end
Supervised machine learning examples:

Static approach, fixed training set
How could the previous approach be improved with the given information?
Using Google searches to track diseases dynamically

begin
%% Load data %%
CDC=load(CDC ILI Data) (ONE COLUMN OF VALUES)
X=load(Google search Data) (MULTIPLE COLUMNS OF VALUES)

%% initialize output arrays %%
Y=zeros(1:end.of.predictions) (INITIALIZE ARRAY TO STORE PREDICTIONS)
coefficients=zeros(1:end.of.predictions) (INITIALIZE ARRAY TO STORE COEFFS)

%% train models and produce out-of-sample predictions %%
for i = training : end.of.predictions
   CDC ← standardize(CDC) (Perhaps use a transform: z-score, logit)
   X ← standardize(X) (Perhaps use a transform: z-score, logit)
   model=LASSOroutine.fit(CDC[1 : i] ∼ X[1 : i]) (TRAINING: IN-SAMPLE MODEL)
   coefficients(i) ← model(coefficients)
   Y(i + 1)=LASSOroutine.predict(model, X(i + 1)) (PRODUCE OUT-OF-SAMPLE PREDICTIONS)
   if(i == training)
      Y[1:i]=LASSOroutine.predict(model, X[1:i]) IN-SAMPLE PREDICTIONS
   end
end
end
Supervised machine learning examples:

Static approach, fixed training set
Supervised machine learning examples:

Dynamic approach, letting the training set expand as more information becomes available

Mexican Secretariat reported cases (Potentially shifted 2 weeks)
GDT (48 months training period)
How could the previous approach be improved?
Assumptions in Google Flu Trends:

1. Number of (influenza-like) ill people proportional to number of total searches of (Influenza-like illnesses) related terms
Assumptions in Google Flu Trends:

1. Number of (influenza-like) ill people proportional to number of total searches of (Influenza-like illnesses) related terms.

Figure 1: An evaluation of how many top-scoring queries to include in the ILI-related query fraction. Maximal performance at estimating out-of-sample points during cross-validation was obtained by summing the top 45 search queries. A steep drop in model performance occurs after adding query 81, which is "oscar nominations".
Assumptions in Google Flu Trends:

2. Relationship between search volume and proportion of (influenza like) ill people is **static** (during a given year).
**Assumptions** in Google Flu Trends:

2. Relationship between search volume and proportion of (influenza like) ill people is **static** (during a given year).

**Consequences:** Model needed constant supervision by human experts

a. **Human experts** needed to **assess** relevance of individual search terms,

b. **Human Experts** needed to **recalculate** relationship between total number of searches and ill people, and

c. It is bound to **deliver poor predictions** at some point in the near future!
We proposed an alternative method and tested it using low quality input from Google Correlate in January 2013.
(with D. Wendong Zhang)

New model:

1. Each search term may contribute to prediction of ILI rate separately (**multi-variate approach**)

2. Relationship between search volume for each individual term and proportion of ill people is **dynamic** and should be found using supervised machine learning optimization techniques.

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

Every week the multiplicative coefficients (\(\beta\)’s) would be automatically updated by expanding the training set (labeled data) as new information from the CDC became available.
# Top correlated terms to CDC-reported data from 1/2004-3/2009 (using Google Correlate)

<table>
<thead>
<tr>
<th>Rank</th>
<th>Term</th>
<th>Rank</th>
<th>Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>influenza type a</td>
<td>35</td>
<td>is the flu contagious</td>
</tr>
<tr>
<td>2</td>
<td>bronchitis</td>
<td>36</td>
<td>flu in children</td>
</tr>
<tr>
<td>3</td>
<td>influenza a</td>
<td>37</td>
<td>fever flu</td>
</tr>
<tr>
<td>4</td>
<td>symptoms of pneumonia</td>
<td>38</td>
<td>take action tour</td>
</tr>
<tr>
<td>5</td>
<td>flu incubation</td>
<td>39</td>
<td>flu remedies</td>
</tr>
<tr>
<td>6</td>
<td>influenza incubation</td>
<td>40</td>
<td>flu report</td>
</tr>
<tr>
<td>7</td>
<td>flu contagious</td>
<td>41</td>
<td>nasal congestion</td>
</tr>
<tr>
<td>8</td>
<td>influenza contagious</td>
<td>42</td>
<td>fever reducer</td>
</tr>
<tr>
<td>9</td>
<td>flu incubation period</td>
<td>43</td>
<td>sinus infections</td>
</tr>
<tr>
<td>10</td>
<td>tussinex</td>
<td>44</td>
<td>rhode island wrestling</td>
</tr>
<tr>
<td>11</td>
<td>benzonatate</td>
<td>45</td>
<td>symptoms of influenza</td>
</tr>
<tr>
<td>12</td>
<td>influenza symptoms</td>
<td>46</td>
<td>castaway bay</td>
</tr>
<tr>
<td>13</td>
<td>a influenza</td>
<td>47</td>
<td>coral by the sea</td>
</tr>
<tr>
<td>14</td>
<td>sinus</td>
<td>48</td>
<td>cold or flu</td>
</tr>
<tr>
<td>15</td>
<td>pneumonia</td>
<td>49</td>
<td>respiratory infection</td>
</tr>
<tr>
<td>16</td>
<td>flu fever</td>
<td>50</td>
<td>take action</td>
</tr>
<tr>
<td>17</td>
<td>flu duration</td>
<td>51</td>
<td>respiratory flu</td>
</tr>
<tr>
<td>18</td>
<td>taste of chaos</td>
<td>52</td>
<td>soweto gospel</td>
</tr>
<tr>
<td>19</td>
<td>bronchitis symptoms</td>
<td>53</td>
<td>soweto gospel choir</td>
</tr>
<tr>
<td>20</td>
<td>symptoms of bronchitis</td>
<td>54</td>
<td>illinois wrestling</td>
</tr>
<tr>
<td>21</td>
<td>how long does the flu last</td>
<td>55</td>
<td>how long is the flu contagious</td>
</tr>
<tr>
<td>22</td>
<td>symptoms of the flu</td>
<td>56</td>
<td>cold symptoms</td>
</tr>
<tr>
<td>23</td>
<td>taste of chaos tour</td>
<td>57</td>
<td>the taste of chaos</td>
</tr>
<tr>
<td>24</td>
<td>influenza incubation period</td>
<td>58</td>
<td>is bronchitis</td>
</tr>
<tr>
<td>25</td>
<td>sinus infection</td>
<td>59</td>
<td>upper respiratory</td>
</tr>
<tr>
<td>26</td>
<td>flu recovery</td>
<td>60</td>
<td>afrin</td>
</tr>
<tr>
<td>27</td>
<td>chaos tour</td>
<td>61</td>
<td>painful cough</td>
</tr>
<tr>
<td>28</td>
<td>type a influenza</td>
<td>62</td>
<td>laprepsoccer</td>
</tr>
<tr>
<td>29</td>
<td>flu symptoms</td>
<td>63</td>
<td>upper respiratory infection</td>
</tr>
<tr>
<td>30</td>
<td>tessalon</td>
<td>64</td>
<td>amoxicillin</td>
</tr>
<tr>
<td>31</td>
<td>type a flu</td>
<td>65</td>
<td>ski harness</td>
</tr>
<tr>
<td>32</td>
<td>treat the flu</td>
<td>66</td>
<td>robitussin dm</td>
</tr>
<tr>
<td>33</td>
<td>treating the flu</td>
<td>67</td>
<td>treating flu</td>
</tr>
<tr>
<td>34</td>
<td>how to treat the flu</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In October 2012 Google decided to discontinue updating Google Correlate (it was Jan 2013)
In October 2012 Google decided to discontinue updating Google Correlate (it was Jan 2013)

We used even lower quality data from Google Trends to test methodology in 2012-2013 recent flu season
What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?

Mauricio Santillana, PhD, MS, D. Wendong Zhang, MA, Benjamin M. Althouse, PhD, ScM, John W. Ayers, PhD, MA

First week after being published online, it became the second most read paper in journal’s history! (After a paper published in 1998)
Relationship of Childhood Abuse and Household Dysfunction to Many of the Leading Causes of Death in Adults: The Adverse Childhood Experiences (ACE) Study

Vincent J Felitti, Robert F Anda, Dale Nordenberg, David F Williamson, Alison M Spitz, Valerie Edwards, Mary P Koss, James S Marks

Vol. 14, Issue 4
Published in issue: May, 1998

Abstract | Full-Text HTML | PDF

What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?

Mauricio Santillana, D. Wendong Zhang, Benjamin M. Althouse, John W. Ayers

Vol. 47, Issue 3
Published online: July 1, 2014

Abstract | Full-Text HTML | PDF
Figure 1. The alternative model outperforms Google Flu Trends

\[ \text{logit}[I(t)] = \sum_{i=1}^{n} a_i(t) \text{logit}[Q_i(t)] + e, \]
Google Flu Trends promises are overstated, researchers say

New study finds way to improve Google Flu Trends accuracy threefold - but says systems must be more open
Google Flu Trends promises are overstated, researchers say

New study finds way to improve Google Flu Trends accuracy threefold - but says systems must be more open

Researchers Suggest Fixes to Google Flu Trends Analytics

A new study concludes that “revising the inner plumbing” of the Google Flu Trends disease surveillance system can improve the accuracy of forecasts for the severity of a flu season.
Finding real value in big data for public health

Date: July 2, 2014
Source: San Diego State University
Summary: Media reports of public health breakthroughs from big data have been largely oversold, according to a new study. But don't throw away that data just yet. The authors maintain that the promise of big data can be fulfilled by tweaking existing methodological and reporting standards. In the study, the research team demonstrate this by revising the inner plumbing of the Google Flu Trends (GFT) digital disease surveillance system, which was heavily criticized last year (see here and here) after producing erroneous forecasts.

Related Topics
- Health & Medicine
  - Health Policy
  - Public Health Education
- Computers & Math
  - Computers and Internet
  - Computer Modeling
- Science & Society
  - Public Health
  - Surveillance

Related Articles
- Public health
- Data mining

A graph depicting Google Flu Trends.
Study: Methodology Changes Improve Google Flu Trend Accuracy

Monday, July 7, 2014

The accuracy of Google Flu Trends' disease surveillance system can be improved through simple changes in three different methodologies used by the system, according to a new study published in the American Journal of Preventive Medicine, Health Data Management reports (Goedert, Health Data Management, 7/7).

- Public Health

A graph depicting Google Flu Trends.
Big Data’s Potential in Public Health: Revisiting Google Flu Trends

July 7, 2014  Written by: Dan Gray  1 Reply

The accuracy of Google Flu Trends’ disease surveillance system can be improved through simple changes in three different methodologies used by the system, according to a new study published in the American Journal of Preventive Medicine, Health Data Management reports (Goedert, Health Data Management, 7/7).

A new study concludes that “revising the input surveillance system can improve the accuracy of Google Flu Trends.” 

Related Articles

> Public health
> Data mining

A graph depicting Google Flu Trends.
Google Flu Trends gets a brand new engine

Posted: Friday, October 31, 2014

Posted by Christian Stefansen, Senior Software Engineer

Each year the flu kills thousands of people and affects millions around the world. So it's important that public health officials and health professionals learn about outbreaks as quickly as possible. In 2008 we launched Google Flu Trends in the U.S., using aggregate web searches to indicate when and where influenza was striking in real time. These models nicely complement other survey systems—they're more fine-grained geographically, and they're typically more immediate, up to 1-2 weeks ahead of traditional methods such as the CDC's official reports. They can also be incredibly helpful for countries that don't have official flu tracking. Since launching, we've expanded Flu Trends to cover 29 countries, and launched Dengue Trends in 10 countries.

The original model performed surprisingly well despite its simplicity. It was retrained just once per year, and typically used only the 50 to 300 queries that produced the best estimates for prior seasons. We then left it to perform through the new season and evaluated it at the end. It didn't use the official CDC data for estimation during the season—only in the initial training.
Google Flu Trends heavily criticized in a paper published by Alex’s research team
1. Lagged (CDC-based) models capable of outperforming GFT.

2. GFT + lagged CDC can outperform GFT (recalibrating importance of GFT)

3. Google search engine itself changed 86 times in June and July 2012 potentially leading to changes in Google search results (independent variable)

4. Feedbacks (recommended search terms depend on previous searches)
Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang\textsuperscript{a}, Mauricio Santillana\textsuperscript{b,c,1}, and S. C. Kou\textsuperscript{a,1}

\textsuperscript{a}Department of Statistics, Harvard University, Cambridge, MA 02138; \textsuperscript{b}School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and \textsuperscript{c}Computational Health Informatics Program, Boston Children’s Hospital, Boston, MA 02115

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

Accurate real-time tracking of influenza outbreaks helps public health officials make timely and meaningful decisions that could save lives. We propose an influenza tracking model, ARGO (AutoRegression with GOogle search data), that uses publicly available online search data. In addition to having a rigorous statistical foundation, ARGO outperforms all previously available Google-search-based tracking models, including the latest version of Google Flu Trends, even though it uses only low-quality search data as input from publicly available Google Trends and Google Correlate websites. ARGO not only incorporates the seasonality in influenza epidemics but also captures changes in people’s online search behavior over time. ARGO is also flexible, self-correcting, robust, and scalable, making it a potentially powerful tool that can be used for real-time tracking of other social events at multiple temporal and spatial resolutions.

CDC’s ILI reports have a delay of 1–3 wk due to the time for processing and aggregating clinical information. This time lag is far from optimal for decision-making purposes. To alleviate this information gap, multiple methods combining climate, demographic, and epidemiological data with mathematical models have been proposed for real-time estimation of flu activity (18, 21–25). In recent years, methods that harness Internet-based information have also been proposed, such as Google (1), Yahoo (2), and Baidu (3) Internet searches, Twitter posts (4), Wikipedia article views (5), clinicians’ queries (6), and crowdsourced self-reporting mobile apps such as Influenzanet (Europe) (26), Flutracking (Australia) (27), and Flu Near You (United States) (28). Among them, GFT has received the most attention and has inspired subsequent digital disease detection systems (3, 8,
We assume there is a Hidden Markov model

\[ Y_{1:N} \rightarrow Y_{2:(N+1)} \rightarrow \cdots \rightarrow Y_{(T-N+1):T} \]
\[ \downarrow \quad \downarrow \quad \downarrow \quad \downarrow \]
\[ X_N \rightarrow X_{N+1} \rightarrow \cdots \rightarrow X_T \]

Our formal mathematical assumptions are

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

(assumption 2) \( X_t | y_t \sim \mathcal{N}_K(\mu_x + y_t \beta, Q) \)

(assumption 3) conditional on \( y_t, X_t \) is independent of \( \{y_l, X_l : l \neq t\} \)
<table>
<thead>
<tr>
<th></th>
<th>Whole period</th>
<th>Off-season flu</th>
<th>Regular flu seasons (week 40 to week 20 next year)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RMSE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARGO</td>
<td>0.637</td>
<td>0.655</td>
<td>0.618</td>
</tr>
<tr>
<td>GFT (Oct 2014)</td>
<td>2.213</td>
<td>0.773</td>
<td>1.110</td>
</tr>
<tr>
<td>Santillana et al. (2014)</td>
<td>0.909</td>
<td>0.945</td>
<td>0.864</td>
</tr>
<tr>
<td>AR(3)</td>
<td>0.955</td>
<td>0.813</td>
<td>0.794</td>
</tr>
<tr>
<td>Naive</td>
<td>1.000 (0.354)</td>
<td>1.000 (0.600)</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>MAE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARGO</td>
<td>0.680</td>
<td>0.607</td>
<td>0.588</td>
</tr>
<tr>
<td>GFT (Oct 2014)</td>
<td>1.828</td>
<td>0.777</td>
<td>1.260</td>
</tr>
<tr>
<td>Santillana et al. (2014)</td>
<td>1.035</td>
<td>0.793</td>
<td>0.977</td>
</tr>
<tr>
<td>AR(3)</td>
<td>0.920</td>
<td>0.777</td>
<td>0.787</td>
</tr>
<tr>
<td>Naive</td>
<td>1.000 (0.206)</td>
<td>1.000 (0.425)</td>
<td>1.000</td>
</tr>
<tr>
<td><strong>Correlation</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARGO</td>
<td>0.984</td>
<td>0.984</td>
<td>0.988</td>
</tr>
<tr>
<td>GFT (Oct 2014)</td>
<td>0.874</td>
<td>0.989</td>
<td>0.968</td>
</tr>
<tr>
<td>Santillana et al. (2014)</td>
<td>0.970</td>
<td>0.959</td>
<td>0.982</td>
</tr>
<tr>
<td>AR(3)</td>
<td>0.963</td>
<td>0.968</td>
<td>0.971</td>
</tr>
<tr>
<td>Naive</td>
<td>0.960</td>
<td>0.951</td>
<td>0.954</td>
</tr>
<tr>
<td><strong>Corr. of increment</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARGO</td>
<td>0.744</td>
<td>0.796</td>
<td>0.793</td>
</tr>
<tr>
<td>GFT (Oct 2014)</td>
<td>0.706</td>
<td>0.863</td>
<td>0.702</td>
</tr>
<tr>
<td>Santillana et al. (2014)</td>
<td>0.671</td>
<td>0.782</td>
<td>0.688</td>
</tr>
<tr>
<td>AR(3)</td>
<td>0.386</td>
<td>0.585</td>
<td>0.569</td>
</tr>
<tr>
<td>Naive</td>
<td>0.438</td>
<td>0.602</td>
<td>0.570</td>
</tr>
</tbody>
</table>
New flu tracker uses Google search data better than Google
Unlike defunct Flu Trends, the model is self-correcting and close to reality.

by Beth Mole - Nov 9, 2015 3:35pm EST
La revanche du big data : Harvard plus forte que Google pour prédire la grippe

Des chercheurs de la prestigieuse université américaine ont conçu un modèle statistique deux fois plus efficace que la méthode Google. Le géant de l'Internet avait fermé cet été son projet, dont les prédictions avaient tourné au flop.
Un nuevo modelo que combina información epidemiológica y búsquedas de Google es capaz de predecir los brotes de gripe una o dos semanas antes que los métodos clínicos tradicionales. El modelo podrá servir para mejorar la toma de decisiones, como la distribución de personal y recursos hospitalarios en regiones que más lo necesiten.
Google, estadística y ‘big data’ para cazar brotes de gripe

Un nuevo modelo que combina información epidemiológica y búsquedas de Google es capaz de predecir los brotes de gripe una o dos semanas antes que los métodos clínicos tradicionales. El modelo podrá servir para mejorar la toma de decisiones, como la distribución de personal y recursos hospitalarios en regiones que más lo necesiten.
Let’s work on writing our own version of **ARGO**  
**Step-by-step**

1. Download Google searches input file from the website “Google correlate”
2. Download the CDC data (gold standard) from course website
3. Repeat dengue exercise, this time make it multi-variables (build a static and dynamic version: Santillana et al, 2014, AJPM)
4. Add historical information in the form of autoregressive terms

5. Let’s make sure you succeed on Wed morning.

6. If you need assistance, please download the ARGO package from course website
Using Google searches to track diseases dynamically

begin

%% Load data %%
CDC = load(CDC ILI Data)  (ONE COLUMN OF VALUES)
X = load(Google search Data)  (MULTIPLE COLUMNS OF VALUES)

%% initialize output arrays %%
Y = zeros(1:end.of.predictions)  (INITIALIZE ARRAY TO STORE PREDICTIONS)
coefficients = zeros(1:end.of.predictions)  (INITIALIZE ARRAY TO STORE COEFFS)

%% train models and produce out-of-sample predictions %%
for i = training : end.of.predictions
    CDC ← standardize(CDC)  (PERHAPS USE A TRANSFORM:Z-SCORE, LOGIT)
    X ← standardize(X)  (PERHAPS USE A TRANSFORM:Z-SCORE, LOGIT)
    model = LASSOroutine.fit(CDC[1:i] ~ X[1:i])  (TRAINING: IN-SAMPLE MODEL)
    coefficients(i) ← model(coefficients)
    Y(i+1) = LASSOroutine.predict(model, X(i+1))  (PRODUCE OUT-OF-SAMPLE PREDICTIONS)
    if (i == training)
        Y[1:i] = LASSOroutine.predict(model, X[1:i])  IN-SAMPLE PREDICTIONS
    end
end
end
And on Aug 20\textsuperscript{th}, 2015
Google discontinues Flu Trends indefinitely!

The Next Chapter for Flu Trends

Posted: Thursday, August 20, 2015

Instead of maintaining our own website going forward, we’re now going to empower institutions who specialize in infectious disease research to use the data to build their own models. Starting this season, we’ll provide Flu and Dengue signal data directly to partners including Columbia University’s Mailman School of Public Health (to update their dashboard), Boston Children’s Hospital/Harvard, and Centers for Disease Control and Prevention (CDC) Influenza Division. We will also continue to make historical Flu and Dengue estimate data available for anyone to see and analyze.
Google Flu Trends calls out sick, indefinitely

Google will pass along search queries related to the flu to health organizations so they can develop their own prediction models

By Fred O'Connor | Follow
IDG News Service | Aug 20, 2015 2:07 PM PT

Google discontinues Flu Trends, starts offering data to researchers

By Jordan Novet | August 20, 2015 12:17 PM
TAGS: GOOGLE, GOOGLE FLU TRENDS
Our team at Boston Children’s Hospital now has access to Google’s search volumes, as one of the exclusive Google’s partners.

We are creating a new improved disease forecasting platform
Thanks to Sue Aman, Rachel Chorney, Jeff Andre, Andre Nguyen, John Brownstein and Healthmap team!
Thank you!

Contact: msantill@fas.harvard.edu