Google Flu Trends
7 January 2016
Google Flu Trends

https://www.google.org/flutrends/ca/#CA
Flu Data

Proportion of Influenza-Like-Illnesses (PILI):
proportion of doctor visits where patient has an
“influenza-like illness”

ILINet in U.S., compiled and published by CDC.

Takes time to record, prepare, and publish
(week or two). Might like to have results faster.

Google Flu tries to predict what the CDC data will be.
Google Flu Trends Model (Classic Edition)

Query fraction (QF): proportion of Google searches that match a given Query

E.g., 16 per 100000 searches are the word “fever”

50 million distinct queries, unknown number of searches (billions)
A little EDA
Step 1: Find terms whose QF correlates with PILI

Worldwide non-commercial space launches correlates with Sociology doctorates awarded (US)
Step 2: Keep the best 45

• Most terms are secret; categories are:
  Influenza Complication, Cold/Flu Remedy, General Influenza Symptoms, Term for Influenza, Specific Influenza Symptom, Symptoms of an Influenza Complication, Antibiotic Medication, General Influenza Remedies, Symptoms of a Related Disease, Antiviral Medication, Related Disease

• [symptoms of bronchitis], [pnumonia]*, [fever], [early signs of the flu], [robitussin], [influenza a], [amoxicillin], [strep throat]
Step 3: Build Model

- Let $S$ be the sum of the QF of the 45 terms over a 1-week period.
- Let $Y$ be the PILI for next week, provided by CDC.
- Find $w$ and $c$ so that

$$\text{logit}(Y) \approx w \times \text{logit}(S) + c$$

as much as possible, on $Y$ and $S$ from historical data.

- $c$ determines predicted “baseline ILI” if $S = 0$, $w$ determines how fast predicted $Y$ grows with $S$.
- To predict $Y$ in the future, stick the learned $w$ and $c$, and the new $S$, into the model, get result.
Step 2: Keep the best 45

- Why 45?
  - Chosen to maximize accuracy on unseen data
Flu Trends Ups and Downs

Figure 1. Time series plots of ILINet data and original and updated GFT estimates.


doi:10.1371/journal.pone.0023610.g001

Google Flu Trends Performance during H1N1
Flu Trends Ups and Downs

<table>
<thead>
<tr>
<th>Year</th>
<th>Google Flu Trends</th>
<th>CDC Data</th>
<th>Flu Near You</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>4.5</td>
<td>3.0</td>
<td>2.5</td>
</tr>
<tr>
<td>2012</td>
<td>3.5</td>
<td>2.5</td>
<td>2.0</td>
</tr>
<tr>
<td>2013</td>
<td>5.0</td>
<td>4.5</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Google's algorithms overestimated peak flu levels this year.
Step 2: Keep the best 45

• Why 45?
  
  • Chosen to maximize accuracy on unseen data
  
  • Other search queries in the top 100, not included in our model, included topics like “high school basketball”
Why

“high school basketball”? 
changes in aggregated search counts on Google during the composition and performance throughout 2009 and discuss through September 2009. We compare the two models' December 2009 and retrospective estimates from July 2003 produce both prospective estimates of ILI from September–ILINet data from April–September, 2009 and was used to updated model launched on September 24, 2009 incorporated prospective estimates of ILI activity for the 2008–2009 flu [8]. The original U.S. GFT model was used to produce and quickly spreading to the United States and around the world influenza A (H1N1) [pH1N1], emerged, beginning in Mexico hypothesis. non-seasonal flu outbreaks, there was no way to test this searching for similar flu-related terms; however, in the absence of expected GFT to detect it as long as Google users continued virus emerge and cause the same symptoms as seasonal flu, we information. Thus, an open question was whether GFT could a change in the terminology used to search online for health behavior could occur during an outbreak or pandemic, resulting in consistency of online health-seeking behavior [7]. Such a shift in GFT has expressed concern that it may be limited by the outbreaks occurred [6]. Previous commentary on the utility of other seasons' data. The correlations between ILINet and GFT model-fitting and then used to test the model estimated from the scale [3]. One season of influenza data was held out during that are influenza-related using a linear model on the log-odds related was estimated from the proportion of Google queries related queries was chosen using a sequential correlation-based method, and the proportion of outpatient visits that are ILI-related queries. A set of influenza-counts were kept for every query in each state. No information issued by a Google search user. Separate aggregate weekly

### Table 1. Comparison of relative query category volume in original and updated United States GFT models.

<table>
<thead>
<tr>
<th>Query Category</th>
<th>Sample Query</th>
<th>Original Model Relative Category Volume</th>
<th>Updated Model Relative Category Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symptoms of an influenza complication</td>
<td>[symptoms of bronchitis]</td>
<td>6%</td>
<td>11%</td>
</tr>
<tr>
<td>Influenza complication</td>
<td>[pnumonia]*</td>
<td>42%</td>
<td>6%</td>
</tr>
<tr>
<td>Specific influenza symptom</td>
<td>[fever]</td>
<td>6%</td>
<td>39%</td>
</tr>
<tr>
<td>General influenza symptoms</td>
<td>[early signs of the flu]</td>
<td>2%</td>
<td>30%</td>
</tr>
<tr>
<td>Cold/flu remedy</td>
<td>[robitussin]</td>
<td>12%</td>
<td>4%</td>
</tr>
<tr>
<td>Term for influenza</td>
<td>[influenza a]</td>
<td>&lt;1%</td>
<td>3%</td>
</tr>
<tr>
<td>Antibiotic medication</td>
<td>[amoxicillin]</td>
<td>12%</td>
<td>0%</td>
</tr>
<tr>
<td>Related disease</td>
<td>[strep throat]</td>
<td>16%</td>
<td>&lt;1%</td>
</tr>
</tbody>
</table>

*Search users often misspell the word *pneumonia*. doi:10.1371/journal.pone.0023610.t001
\[ \logit(Y) \approx w \times \logit(S) + c \]
What causes flu-related searches?