Originally published 13 March; corrected 25 April 2014

www.sciencemag.org/content/343/6176/1203/suppl/DC1

Supplementary Materials for

The Parable of Google Flu: Traps in Big Data Analysis

David Lazer,* Ryan Kennedy, Gary King, Alessandro Vespignani

*Corresponding author. E-mail: d.lazer@neu.edu

Published 14 March 2014, *Science* **343**, 1203 (2014)
DOI: 10.1126/science.1248506

This PDF file includes

Materials and Methods
Supplementary Text
Figs. S1 to S26
Tables S1 to S5
References

Other Supplementary Material for this manuscript includes the following:

D. Lazer, R. Kennedy, G. King, A. Vespignani, "Replication data for: The Parable of Google Flu: Traps in Big Data Analysis" (2014); http://dx.doi.org/10.7910/DVN/24823 UNF:5:BJh9WzZQNEeSEpV3EWs+xg== IQSS Dataverse Network [Distributor].

Corrected: Author's submitted changes are now included. References in the list are renumbered correctly.

Table of Contents

Materials and Methods

1. Comparison of Time Series and GFT Results Across HHS Regions

As we noted in the Policy Forum, national-level flu trends can be well-approximated using lagged dependent variables and variables for seasonality. These approximations can be further improved by combining Google Flu Trends' (GFT's) estimates of influenza-like-illness (ILI) with a lagged dependent variable, 2- and 3-week lags of the difference between GFT and the Centers for Disease Control and Prevention (CDC), and the seasonality variables.

This section demonstrates that these results hold up in a region-by-region analysis. Table S1 compares errors across the entire period from 2003 to 2013 for the GFT model, the time series model (both with 07/03/2011, when GFT went off the rails, and with 09/06/2009, when the new GFT algorithm went live, as the out-of-sample start date). Table S2 compares the results in only the out-of-sample data for the post-09/06/2009 model. Table S3 compares the results in only the out-of-sample data for the post-07/03/2011 model. In the latter two tables, the comparison is against GFT after these respective dates.

The use of regional comparisons is closer to the method used by the original GFT, which was developed primarily in relation to the HHS regions (*13*). The results are generally consistent with the national results. GFT's errors, however, are not uniformly distributed across regions. For example, in the post-2011 period, GFT does remarkably well in regions 7 and 8, whereas it performs poorly in regions 6 and 10. Nevertheless, in almost every case, GFT is outperformed by the basic time series predictions and the combined model. Although not discussed in the Policy Forum, it is interesting to note that GFT started to estimate high regularly after the summer 2011 change in geocoding, which was aimed at providing more accurate geographic results for searches (see section 3 below). It is possible that this contributed to the algorithm's later increase in error (i.e., greater accuracy in geographic search attribution threw off an algorithm designed on the less-accurate attribution).

Replication code and data for these results are available in the folder labeled SOM1 (*36*) in the replication files. The replication data file is labeled ParableOfGFT(SOM1Replication).dta and the code is SOMpt1(Replication Code).do. Instructions for replication are included in the code file. The Stata statistical program was used for calculations.

Table S1. Regional Comparison of Models. RMSE, root mean squared error; MAE, mean absolute error. The lagged CDC model is specified as

$$flu_t = \alpha + \beta_1 flu_{t-2} + \beta_2 flu_{t-3} + \beta_3 flu_{t-4} + \sum_{i=1}^{52} \gamma_i \, week_{it}$$

where *flu* is the CDC estimate of percent doctors' visits for ILI, *week* is a binary variable indicating week of observation, and β and γ are estimated regression coefficients. The combination of GFT and CDC data is specified as

$$flu_t = \alpha + \beta_1 gflu_t + \beta_2 flu_{t-2} + \beta_3 (gflu_{t-2} - flu_{t-2}) + \beta_4 (gflu_{t-3} - flu_{t-3}) + \sum_{i=1}^{52} \gamma_i week_{it}$$

where *gflu* is the GFT estimate. (See Section 7 of SM for details.)

Region	Google Flu Trends	CDC Lag and Seasonality Model (Post-07/03/2011 Out-of-Sample)	Google Flu Trends Plus Lag and Seasonality Corrections (Post-07/03/2011 Out-of-Sample)	CDC Lag and Seasonality Model (Post-09/06/2009 Out-of-Sample)	Google Flu Trends Plus Lag and Seasonality Corrections (Post-09/06/2009 Out-of-Sample)
Region 1 (CT, ME, MA, NH, RI, VT)	RMSE = 0.772 MAE = 0.276	RMSE = 0.531 MAE = 0.297	RMSE = 0.532 MAE = 0.216	RMSE = 0.581 MAE = 0.292	RMSE = 0.538 MAE = 0.219
Region 2 (NJ, NY)	RMSE = 0.876 MAE = 0.527	RMSE = 0.615 MAE = 0.392	RMSE = 0.474 MAE = 0.309	RMSE = 0.640 MAE = 0.397	RMSE = 0.487 MAE = 0.313
Region 3 (DE, DC, MD, PA, VA, WV)	RMSE = 0.832 MAE = 0.556	RMSE = 0.682 MAE = 0.439	RMSE = 0.513 MAE = 0.337	RMSE = 0.717 MAE = 0.448	RMSE = 0.526 MAE = 0.351
Region 4 (AL, FL, GA, KY, MS, NC, SC, TN)	RMSE = 0.695 MAE = 0.380	RMSE = 0.491 MAE = 0.310	RMSE = 0.353 MAE = 0.203	RMSE = 0.500 MAE = 0.316	RMSE = 0.358 MAE = 0.207
Region 5 (IL, IN, MI, MN, OH, WI)	RMSE = 0.738 MAE = 0.379	RMSE = 0.576 MAE = 0.342	RMSE = 0.390 MAE = 0.214	RMSE = 0.599 MAE = 0.344	RMSE = 0.400 MAE = 0.221
Region 6 (AR, LA, NM, OK, TX)	RMSE = 1.280 MAE = 0.732	RMSE = 0.881 MAE = 0.570	RMSE = 0.728 MAE = 0.441	RMSE = 0.913 MAE = 0.582	RMSE = 0.760 MAE = 0.445
Region 7 (IA, KS, MO, NE)	RMSE = 0.894 MAE = 0.508	RMSE = 0.774 MAE = 0.442	RMSE = 0.492 MAE = 0.281	RMSE = 0.812 MAE = 0.453	RMSE = 0.507 MAE = 0.289
Region 8 (CO, MT, ND, SD, UT, WY)	RMSE = 0.610 MAE = 0.342	RMSE = 0.555 MAE = 0.319	RMSE = 0.434 MAE = 0.248	RMSE = 0.595 MAE = 0.323	RMSE = 0.452 MAE = 0.250
Region 9 (AZ, CA, HI, NV)	RMSE = 0.971 MAE = 0.623	RMSE = 0.659 MAE = 0.435	RMSE = 0.525 MAE = 0.344	RMSE = 0.672 MAE = 0.445	RMSE = 0.532 MAE = 0.364
Region 10 (AK, ID, OR, WA)	RMSE = 1.050 MAE = 0.610	RMSE = 0.818 MAE = 0.520	RMSE = 0.710 MAE = 0.450	RMSE = 0.839 MAE = 0.538	RMSE = 0.717 MAE = 0.473

Table S2. Regional Comparison of Models Post-09/06/2009.

The lagged CDC model is specified as

$$flu_t = \alpha + \beta_1 flu_{t-2} + \beta_2 flu_{t-3} + \beta_3 flu_{t-4} + \sum_{i=1}^{52} \gamma_i \, week_{it}$$

where *flu* is the CDC estimate of percent doctors' visits for ILI, *week* is a binary variable indicating week of observation, and β and γ are estimated regression coefficients. The combination of GFT and CDC data is specified as

$$flu_t = \alpha + \beta_1 gflu_t + \beta_2 flu_{t-2} + \beta_3 (gflu_{t-2} - flu_{t-2}) + \beta_4 (gflu_{t-3} - flu_{t-3}) + \sum_{i=1}^{52} \gamma_i week_{it}$$

where *gflu* is the GFT estimate. (See Section 7 of SM for details.)

Region	**Google Flu Trends**	**CDC Lag and Seasonality Model (Post-09/06/2009 Out-of-Sample)**	**Google Flu Trends Plus Lag and Seasonality Corrections (Post-09/06/2009 Out-of-Sample)**
Region 1 (CT, ME, MA, NH, RI, VT)	RMSE = 1.152 MAE = 0.375	RMSE = 0.741 MAE = 0.325	RMSE = 0.788 MAE = 0.291
Region 2 (NJ, NY)	RMSE = 1.113 MAE = 0.617	RMSE = 0.727 MAE = 0.417	RMSE = 0.569 MAE = 0.346
Region 3 (DE, DC, MD, PA, VA, WV)	RMSE = 1.029 MAE = 0.689	RMSE = 0.858 MAE = 0.492	RMSE = 0.603 MAE = 0.381
Region 4 (AL, FL, GA, KY, MS, NC, SC, TN)	RMSE = 0.960 MAE = 0.590	RMSE = 0.477 MAE = 0.323	RMSE = 0.392 MAE = 0.231
Region 5 (IL, IN, MI, MN, OH, WI)	RMSE = 0.988 MAE = 0.499	RMSE = 0.666 MAE = 0.385	RMSE = 0.520 MAE = 0.275
Region 6 (AR, LA, NM, OK, TX)	RMSE = 1.584 MAE = 0.824	RMSE = 0.993 MAE = 0.591	RMSE = 0.904 MAE = 0.479
Region 7 (IA, KS, MO, NE)	RMSE = 1.033 MAE = 0.611	RMSE = 0.941 MAE = 0.522	RMSE = 0.651 MAE = 0.355
Region 8 (CO, MT, ND, SD, UT, WY)	RMSE = 0.740 MAE = 0.373	RMSE = 0.728 MAE = 0.357	RMSE = 0.546 MAE = 0.266
Region 9 (AZ, CA, HI, NV)	RMSE = 1.308 MAE = 0.872	RMSE = 0.630 MAE = 0.408	RMSE = 0.592 MAE = 0.385
Region 10 (AK, ID, OR, WA)	RMSE = 1.256 MAE = 0.729	RMSE = 0.748 MAE = 0.473	RMSE = 0.668 MAE = 0.437

Table S3. Regional Comparison of Models Post-07/03/2011.

The lagged CDC model is specified as

$$flu_t = \alpha + \beta_1 flu_{t-2} + \beta_2 flu_{t-3} + \beta_3 flu_{t-4} + \sum_{i=1}^{52} \gamma_i \, week_{it}$$

where *flu* is the CDC estimate of percent doctors' visits for ILI, *week* is a binary variable indicating week of observation, and β and γ are estimated regression coefficients. The combination of GFT and CDC data is specified as

$$flu_t = \alpha + \beta_1 gflu_t + \beta_2 flu_{t-2} + \beta_3 (gflu_{t-2} - flu_{t-2}) + \beta_4 (gflu_{t-3} - flu_{t-3}) + \sum_{i=1}^{52} \gamma_i week_{it}$$

where *gflu* is the GFT estimate. (See Section 7 of SM for details.)

Region	**Google Flu Trends**	**CDC Lag and Seasonality Model (Post-07/03/2011 Out-of-Sample)**	**Google Flu Trends Plus Lag and Seasonality Corrections (Post-07/03/2011 Out-of-Sample)**
Region 1 (CT, ME, MA, NH, RI, VT)	RMSE = 1.526 MAE = 0.499	RMSE = 0.345 MAE = 0.216	RMSE = 1.004 MAE = 0.351
Region 2 (NJ, NY)	RMSE = 1.322 MAE = 0.623	RMSE = 0.483 MAE = 0.312	RMSE = 0.552 MAE = 0.324
Region 3 (DE, DC, MD, PA, VA, WV)	RMSE = 1.320 MAE = 0.951	RMSE = 0.594 MAE = 0.390	RMSE = 0.611 MAE = 0.372
Region 4 (AL, FL, GA, KY, MS, NC, SC, TN)	RMSE = 1.117 MAE = 0.610	RMSE = 0.452 MAE = 0.308	RMSE = 0.435 MAE = 0.253
Region 5 (IL, IN, MI, MN, OH, WI)	RMSE = 1.284 MAE = 0.658	RMSE = 0.459 MAE = 0.305	RMSE = 0.606 MAE = 0.301
Region 6 (AR, LA, NM, OK, TX)	RMSE = 2.013 MAE = 1.091	RMSE = 0.737 MAE = 0.486	RMSE = 0.894 MAE = 0.484
Region 7 (IA, KS, MO, NE)	RMSE = 0.524 MAE = 0.363	RMSE = 0.628 MAE = 0.398	RMSE = 0.714 MAE = 0.344
Region 8 (CO, MT, ND, SD, UT, WY)	RMSE = 0.550 MAE = 0.329	RMSE = 0.358 MAE = 0.242	RMSE = 0.512 MAE = 0.233
Region 9 (AZ, CA, HI, NV)	RMSE = 1.441 MAE = 0.728	RMSE = 0.539 MAE = 0.352	RMSE = 0.659 MAE = 0.375
Region 10 (AK, ID, OR, WA)	RMSE = 1.613 MAE = 0.977	RMSE = 0.464 MAE = 0.354	RMSE = 0.744 MAE = 0.437

2. Autocorrelation and Partial Autocorrelation analysis of CDC % ILI data and GFT errors

As noted in the Policy Forum, the signs of autocorrelation in the CDC's data and in GFT's errors are apparent from simple examination of the time series charts. A more formal method for picturing this are correlograms and partial correlograms. The autocorrelations and 95% confidence intervals (95% CI) (gray shaded region) for the raw CDC percent visits for influenza-like illness (ILI) data are shown in Fig. S1. These results are very highly autocorrelated up to nearly eight lags. The partial correlogram (Fig. S2) shows that the partial autocorrelation extends for about two lags.

Perhaps the most interesting plot here is Fig. S3, however, which shows that the GFT model produces highly autocorrelated errors. This means that we can predict quite accurately how much GFT will miss by in time *t* by looking at how much it misses by in times $t - 1$, $t - 2$, etc. This is the intuition that leads us to use the lagged error along with GFT's data to produce our most accurate model.

Replication of data for this figure is in the Manuscript file in the replication file, labeled ParableOfGFT(Replication).dta (*36*). Code is located in the SOM2 folder with the code SOM2(Replication Code).do. Calculations done in Stata.

Fig. S1. Correlogram of CDC % ILI Data

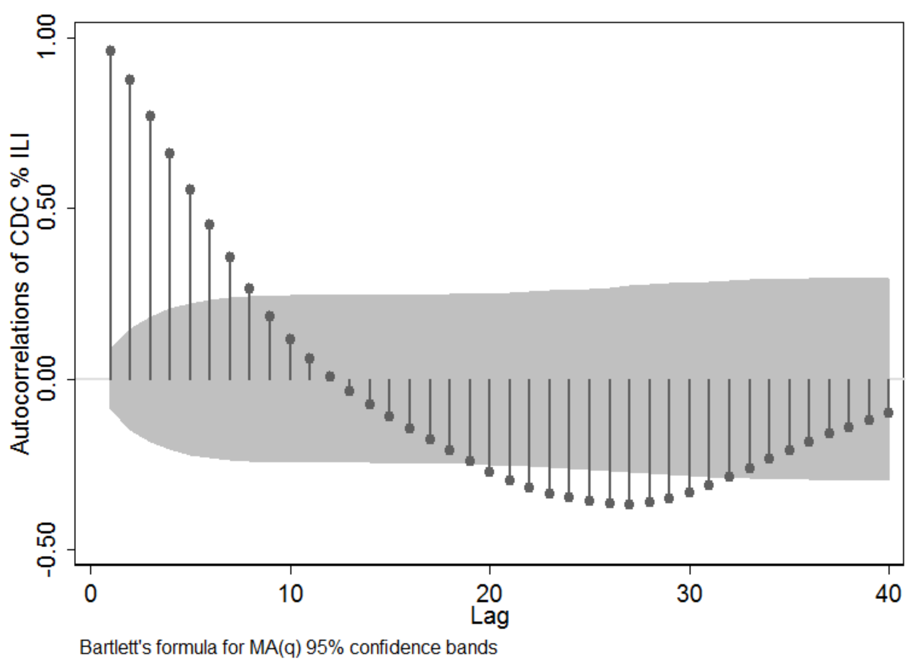


Fig. S2. Partial Correlogram of CDC % ILI Data

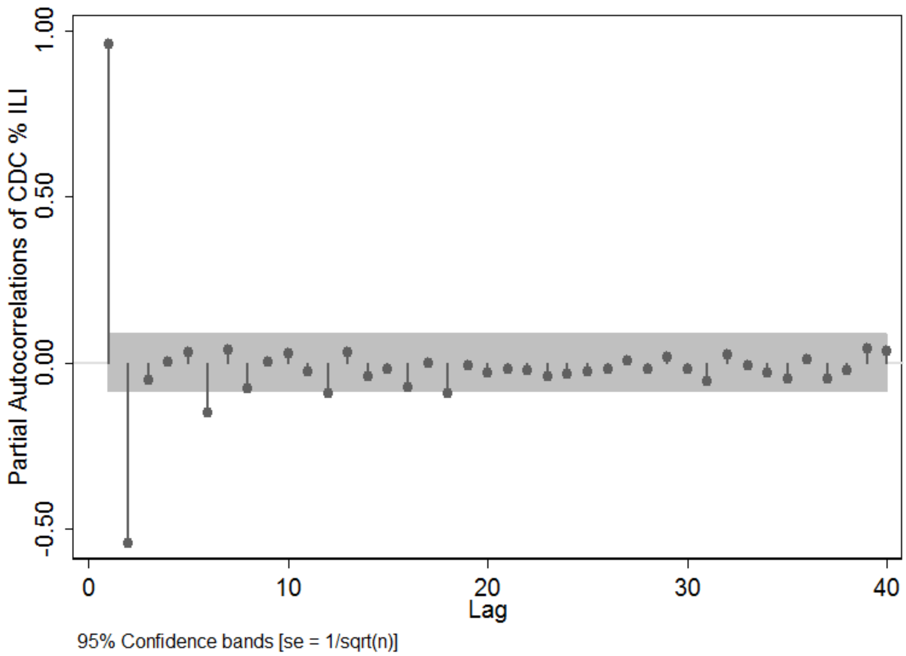
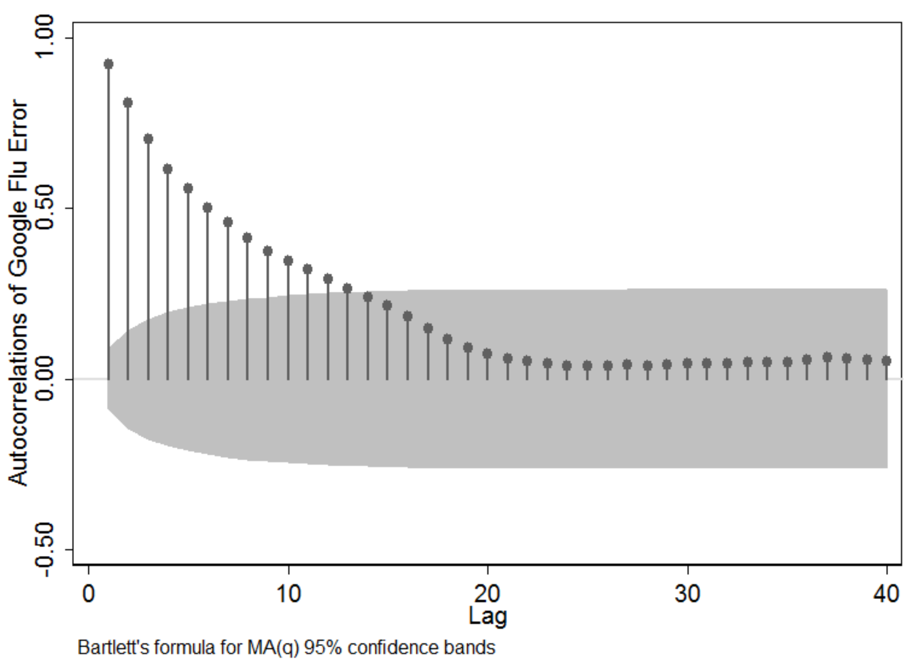


Fig. S3. Correlogram of Google Flu Trends Errors

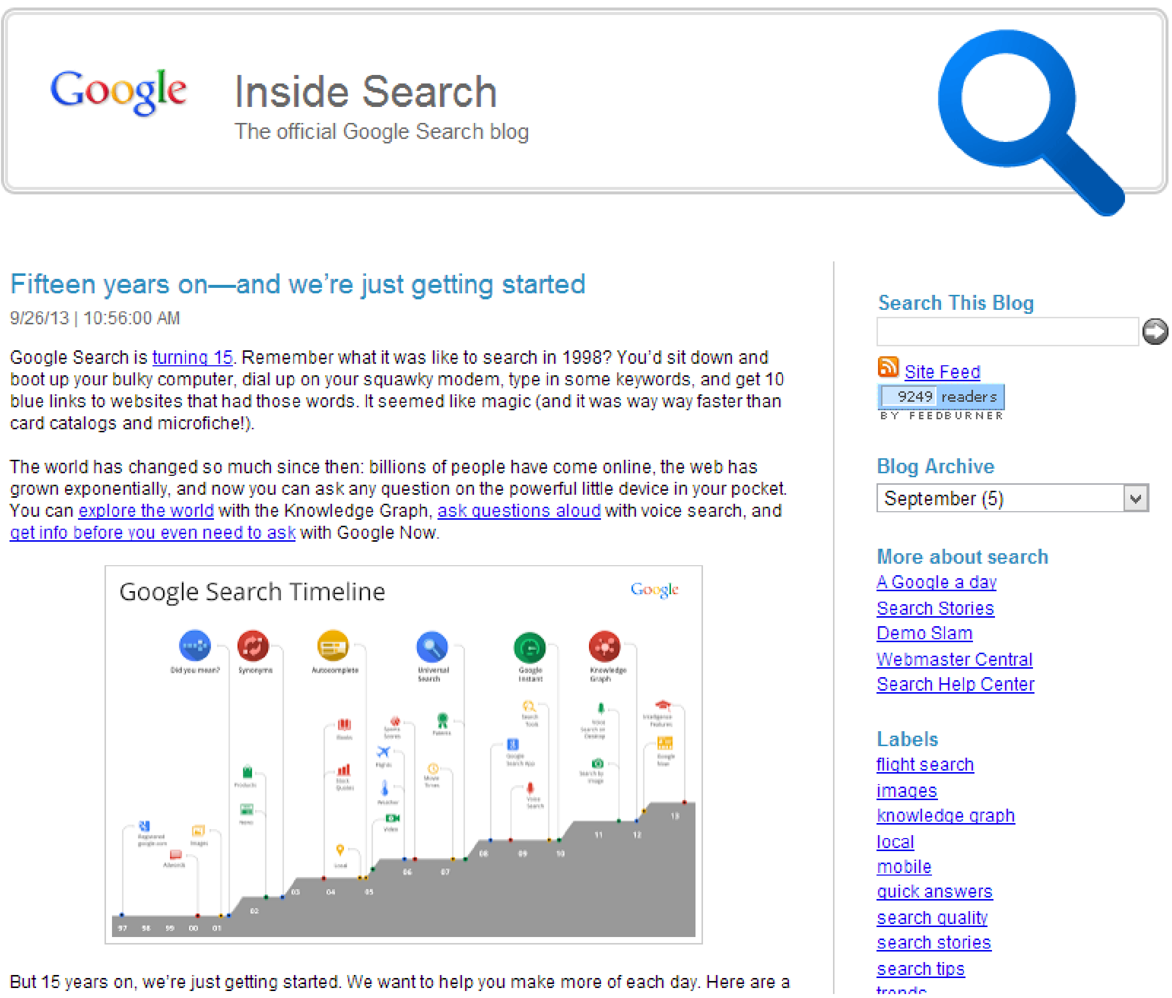


3. Tracking Google's Search Algorithm: An Analysis of Google's Official Search Blog

As we note in the Policy Forum, a full accounting of the evolution of Google's search algorithm is nearly impossible—hundreds of changes are made to the algorithm each year, some major and some minor. We can, however, track major changes using Google's official search blog, which has recorded major events for Google since May 2011 (http://insidesearch.blogspot.com/).

The first thing to note from this record is how much Google's search algorithm changes over time. As Fig. S4 points out, this is a point of pride for Google, as they note in their most recent blog post, celebrating their 15th anniversary.

Fig. S4. Google Blog Discussing How Much Things Have Changed in 15 Years



Source: http://insidesearch.blogspot.com/2013/09/fifteen-years-onand-were-just-getting.html (*37*)

Health-related searches have been a particular area of emphasis for Google. At the official Google blog, the company announced (Fig. S5) that it is consistently running experiments to improve the services provided to its users. In this case, the company was using a short poll to determine the primary reasons why people use different health-related search terms in order to provide users with more relevant search results. We suspect that this determination to provide useful information had the unexpected side effect of increasing certain searches and throwing off GFT's statistics.

Fig. S5. An Example of Google's Consistent Experiments with Health Searches

Google | Official Blog

Insights from Googlers into our products, technology, and the Google culture

Understanding health-related searches

Posted: Wednesday, May 13, 2009

+1 1 Tweet 1 Like 4

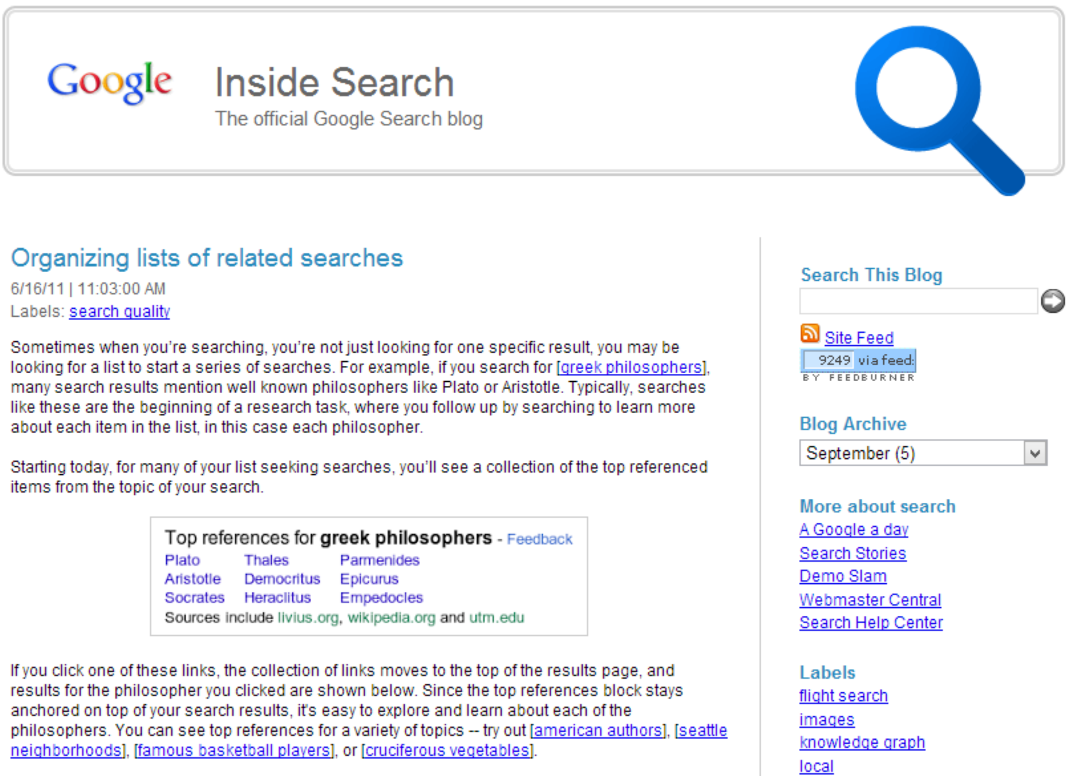
As we blogged last summer, there are lots of experiments running on Google web search all over the world. Today we've started a temporary experiment that some people might find interesting: we're researching how Google users search the Internet when they or someone they know is feeling sick.

Understanding how people search when they're feeling sick is an important problem to solve, as it can help improve projects like Google Flu Trends, which uses aggregated search data to detect influenza epidemics. Statistics gathered in this experiment may also help Google deliver more relevant search results in the future. For example, someone who searches for [arthritis pain] to understand why an aging parent is experiencing joint pain might want to learn about nearby health facilities and potential treatments, whereas somebody who searches for [arthritis pain] because she is doing a research project might want results about how common arthritis is and what its risk factors are. Rather than make educated guesses about how many users are searching because they're sick, we're running this experiment to collect real statistics. This is not a permanent change, but a short-term experiment. A small percentage of random health-related searches will trigger the poll question.

Source: http://googleblog.blogspot.com/2009/05/understanding-health-related-searches.html (*38*)

An example of the changes made by Google, and also a likely culprit for changing search behavior, is the addition of the related search feature in mid-June 2011, a few weeks before GFT started persistently missing high. This addition provides users with the top referenced searches related to their original search term. For example, a search for “flu” will bring up a number of additional recommended searches about diagnosis and treatment of the flu. As we discuss below, increasing searches for flu treatments are a likely reason for GFT’s growing inaccuracy. The announcement of this feature is shown in Fig. S6.

Fig. S6. Google Introduces List of Related Searches Feature



Source: http://insidesearch.blogspot.com/2011/06/organizing-lists-of-related-searches_16.html (*21*)

Another likely culprit was introduced in February 2012. The health search box was introduced as a way for consumers to search particular symptoms, like “fever” and “cough” and find potential diagnoses of these conditions. The announcement of this technology (Fig. S7) is followed by the sample search given by Google to demonstrate its use (Fig. S8). In Fig. S9, we demonstrate this in a more general context of a search for “fever” and “runny nose.” As the reader can see, the top two results are for the flu and the common cold. This seems a likely reason why searches like “cold vs flu” and “cold or flu” seem to spike up in the 2012–2013 flu season.

Fig. S7. Google Announcement of Improved Health Searches

Google Inside Search
The official Google Search blog

Improving health searches, because your health matters
2/13/12 | 9:15:00 AM

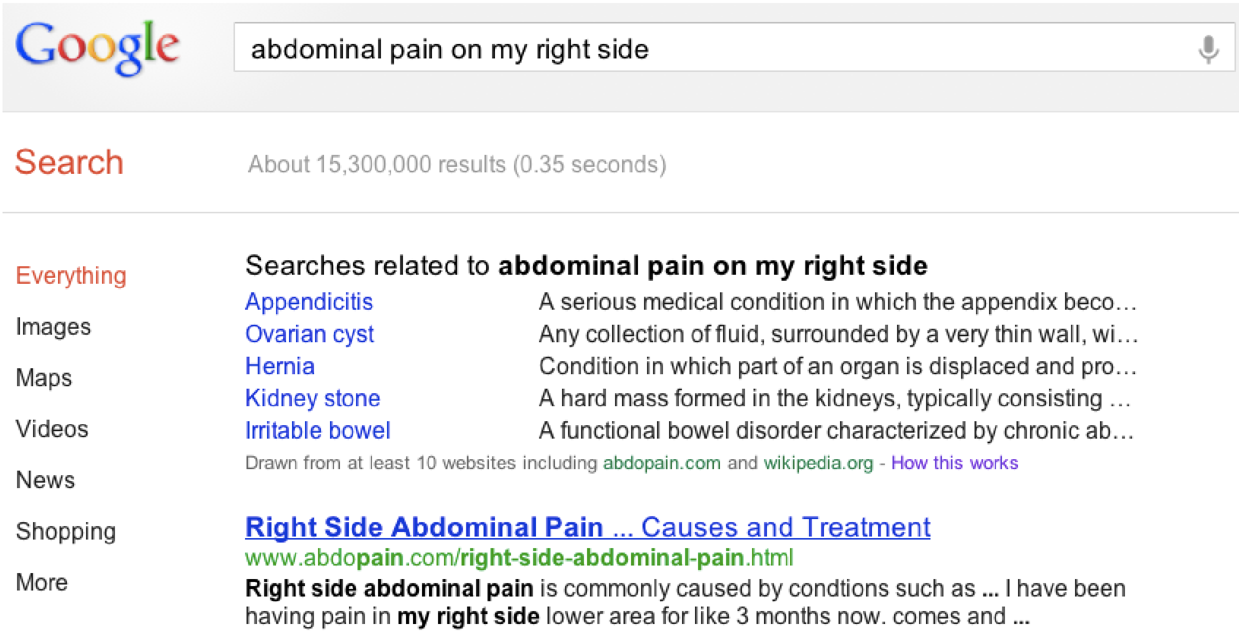
Every day, people search on Google for health information. Many of these searches relate to symptoms they or their loved ones may be experiencing. You might be trying to understand why you've had a headache every morning for a week or why your child has a tummy ache all of a sudden. Our data shows that a search for symptoms is often followed by a search for a related condition.

To make the process easier, now when you search for a symptom or set of symptoms, you'll often see a list of possibly related health conditions that you can use to refine your search. The list is generated by our algorithms that analyze data from pages across the web and surface the health conditions that appear to be related to your search.

For example, if you search for [abdominal pain on my right side], you'll be able to quickly see some potentially related conditions and learn more about them by clicking on the links in the list.

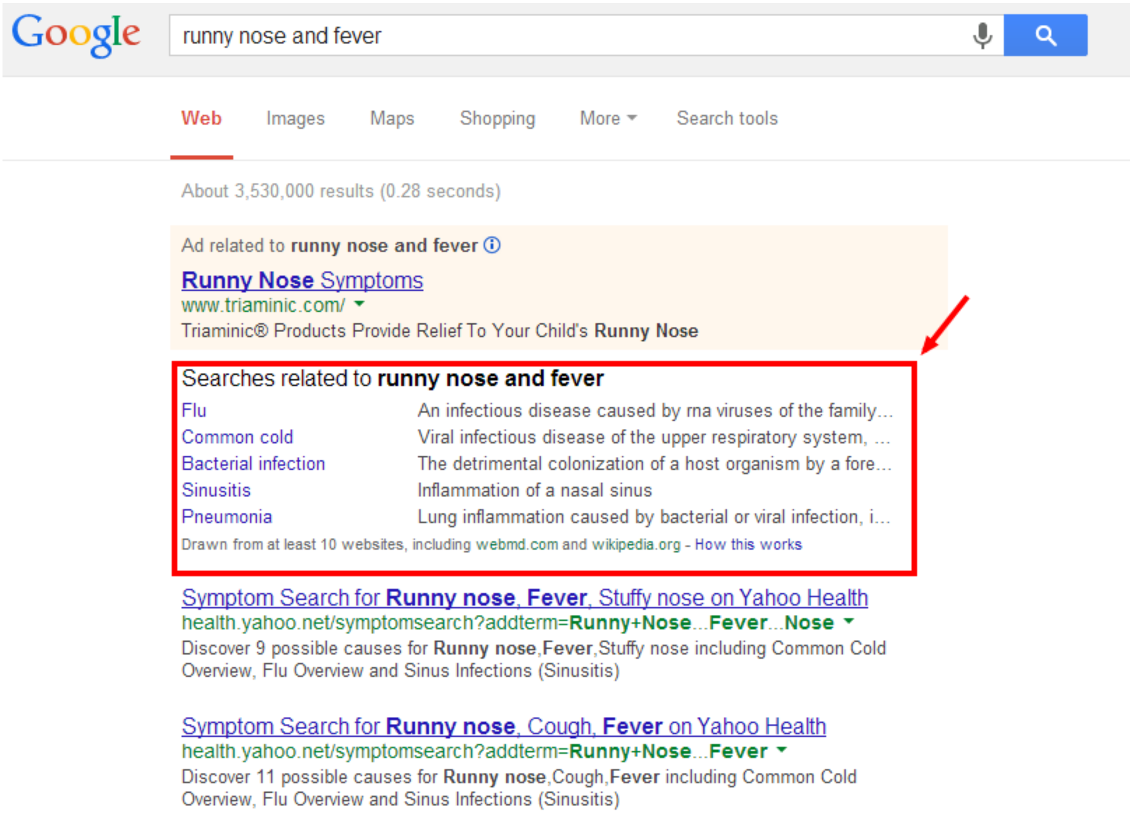
Source: http://insidesearch.blogspot.com/2012/02/improving-health-searches-because-your.html (*22*)

Fig. S8. Google Sample of New Health Search Box



Source: http://insidesearch.blogspot.com/2012/02/improving-health-searches-because-your.html (*22*)

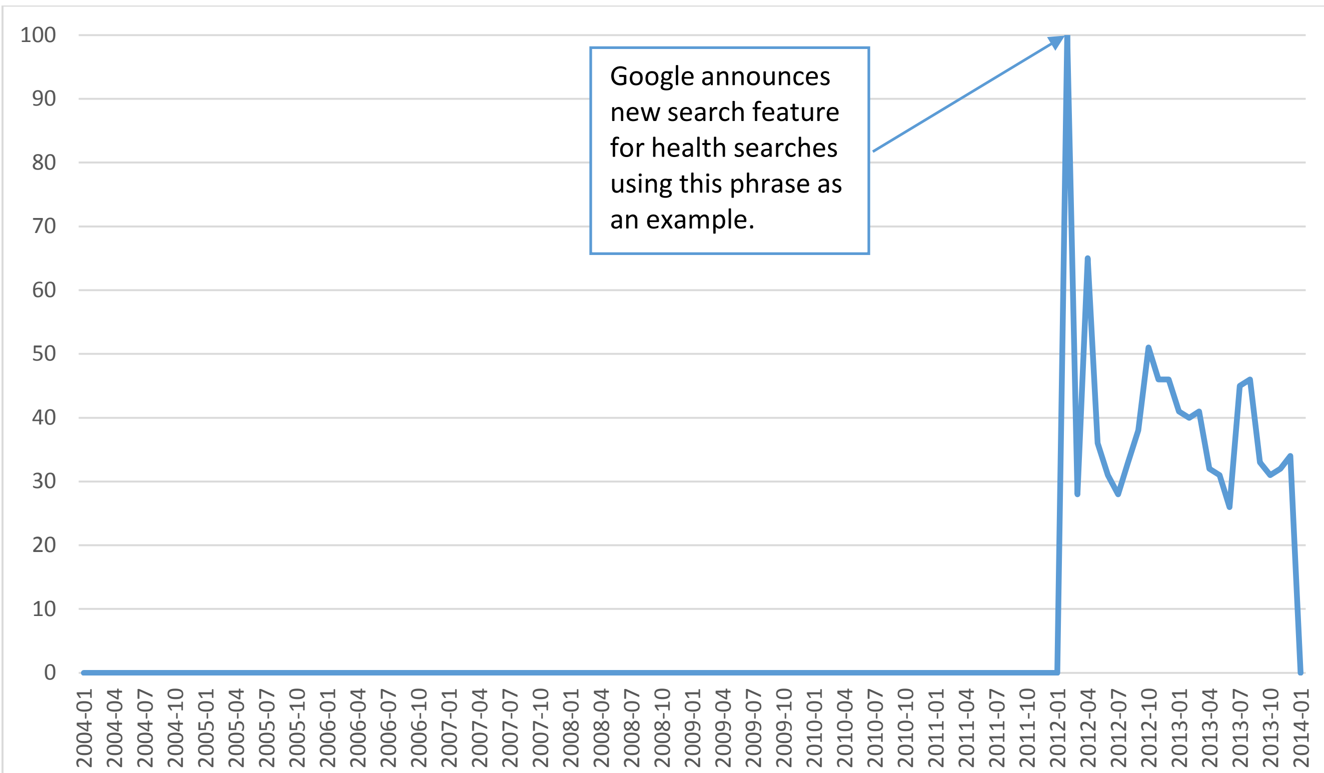
Fig. S9. Google Search for "runny nose" and "fever"



Note: Search for "runny nose and fever" (see search line) conducted from Boston, MA, in October 2013. Results last replicated on 26 February 2014. Searches were conducted using both Google Chrome and Mozilla Firefox browsers.

This particular announcement also had an added effect—it demonstrates the ability for Google itself to influence the relative prevalence of search term use. As one might suspect, searches for "abdominal pain in my right side" are relatively rare. The relative prevalence of this search, however, sparks markedly right after the announcement using this term as an example. This spike is documented in Fig. S10.

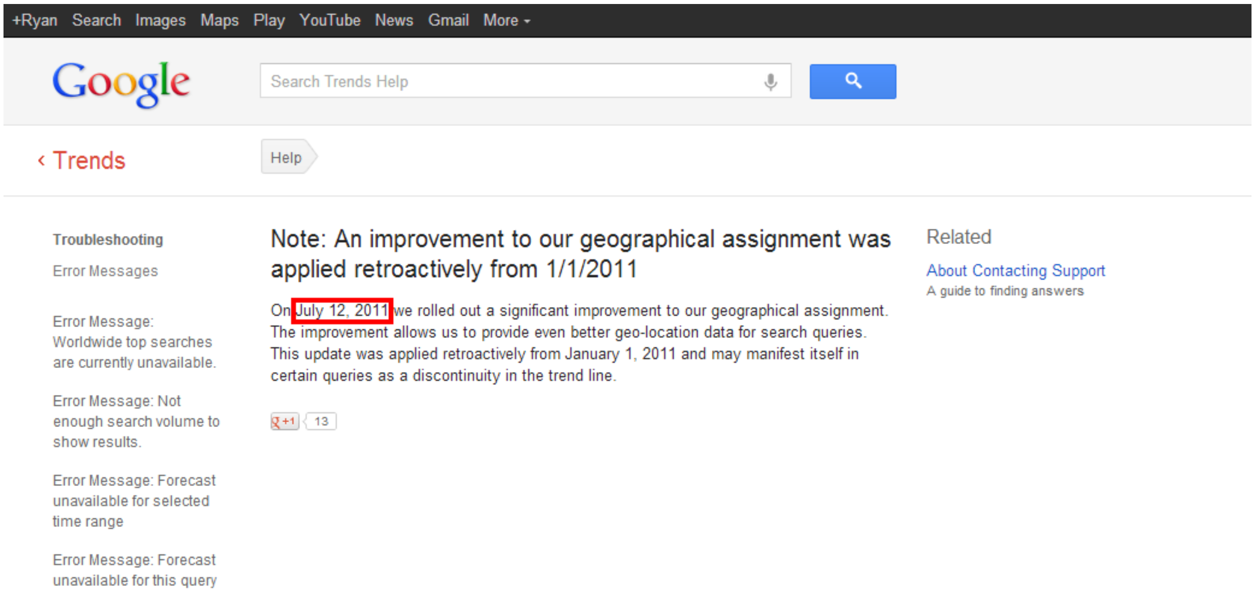
Fig. S10. Searches for "abdominal pain on my right side." Google Trends results are based on the proportion of searches for the phrase relative to total searches. The reported results are scaled so that 100 represents the highest relative volume of searches for the phrase. All others are reported as the percentage relative to the highest search volume. The peak, in this case, is more than four standard deviations above the average for this search term. The peak is reached in February 2012. Google announced the new search feature on 13 February 2012 (see above).



Source: Google Trends (http://www.google.com/trends/) (*39*). Downloaded data available in replication materials (*36*).

These are far from the only changes that might have affected GFT's performance. As noted above, the performance of GFT after 2011 was markedly different between regions. This is important, as the model was originally developed based on regional flu time series, rather than the national time series [p. 2 of (*13*)]. Note that in July 2011, Google announced a substantial improvement in its geolocation abilities for search that were retroactively applied to all searches after 1 January 2011. This change was not documented on the official search blog but was noted in certain searches in Google Trends. We document this in Fig. S11.

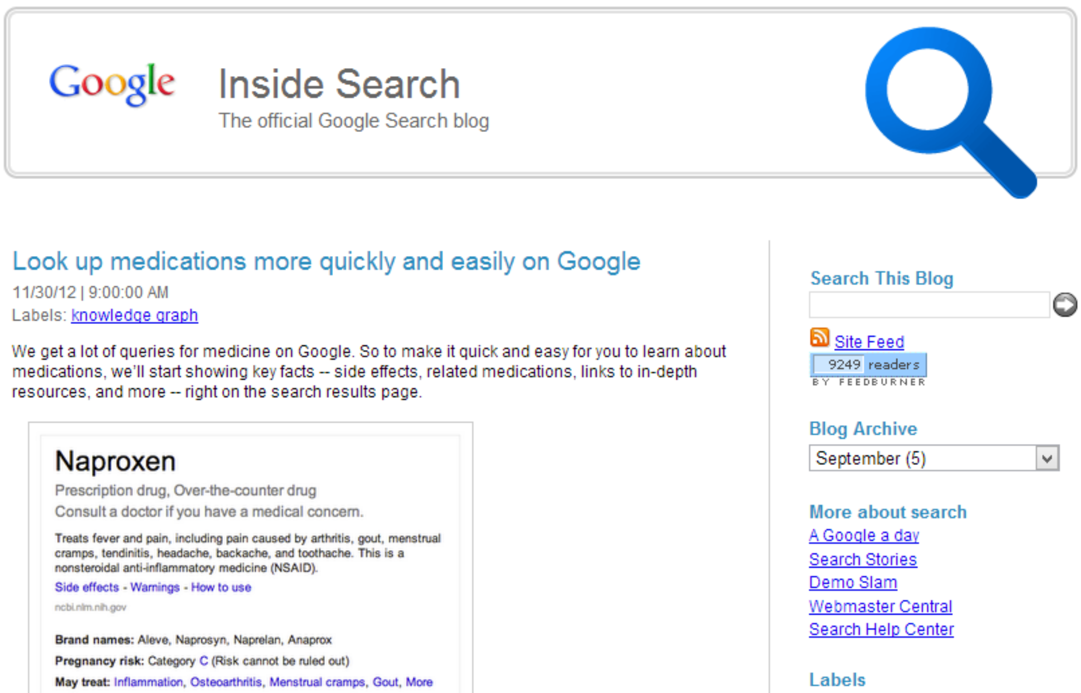
Fig. S11. Documentation of Change in Geographic Resolution in 2011. As indicated by the red highlight, this change was originally implemented on July 12, 2011. This note was originally posted at https://support.google.com/trends/?hl=en&rd=2#topic=4365599 (*40*) but has subsequently been removed.



Documentation of this change can be found in (*41*) and http://www.epiphanysearch.co.uk/blog/2011/07/warning-google-makes-insights-for-search-data-useless/ (42).

Even smaller changes can also affect how GFT will work in the future. At the end of 2012, Google announced new features to help people with finding information on different medications, their uses, and potential interactions. From a business and customer service perspective, the goal is to increase the usage of Google for information on medications, but given that "robitussin" is listed as one of the example search terms for GFT by Google (*14*) and that there are likely other treatments present, an increase in the relative frequency of searches for medication might influence GFT results. We document this change in Fig. S12.

Fig. S12. Google Feature to Look Up Medication Information



Source: http://insidesearch.blogspot.com/2012/11/look-up-medications-more-quickly-and.html (*43*)

The last three figures present some summary issues related to Google search. Fig. S13 provides a visual summary of many of the changes made in the last few years to how Google presents results. As the reader can see, a standard search for "flu" produces a number of new choices that may affect consumer behavior. Among these are a noticeable side bar on influenza that includes common searches as well as treatment and symptom information, a set of in-depth articles of varying quality, links to posts by Google+ contacts, and the related searches option discussed above.

Fig. S13. Examples of Other Google Search Changes for "flu." Search for "flu" (see above search line) conducted from Boston, MA, in October 2013. Results last replicated on 26 February 2014. Searches were conducted using both Google Chrome and Mozilla Firefox browsers.

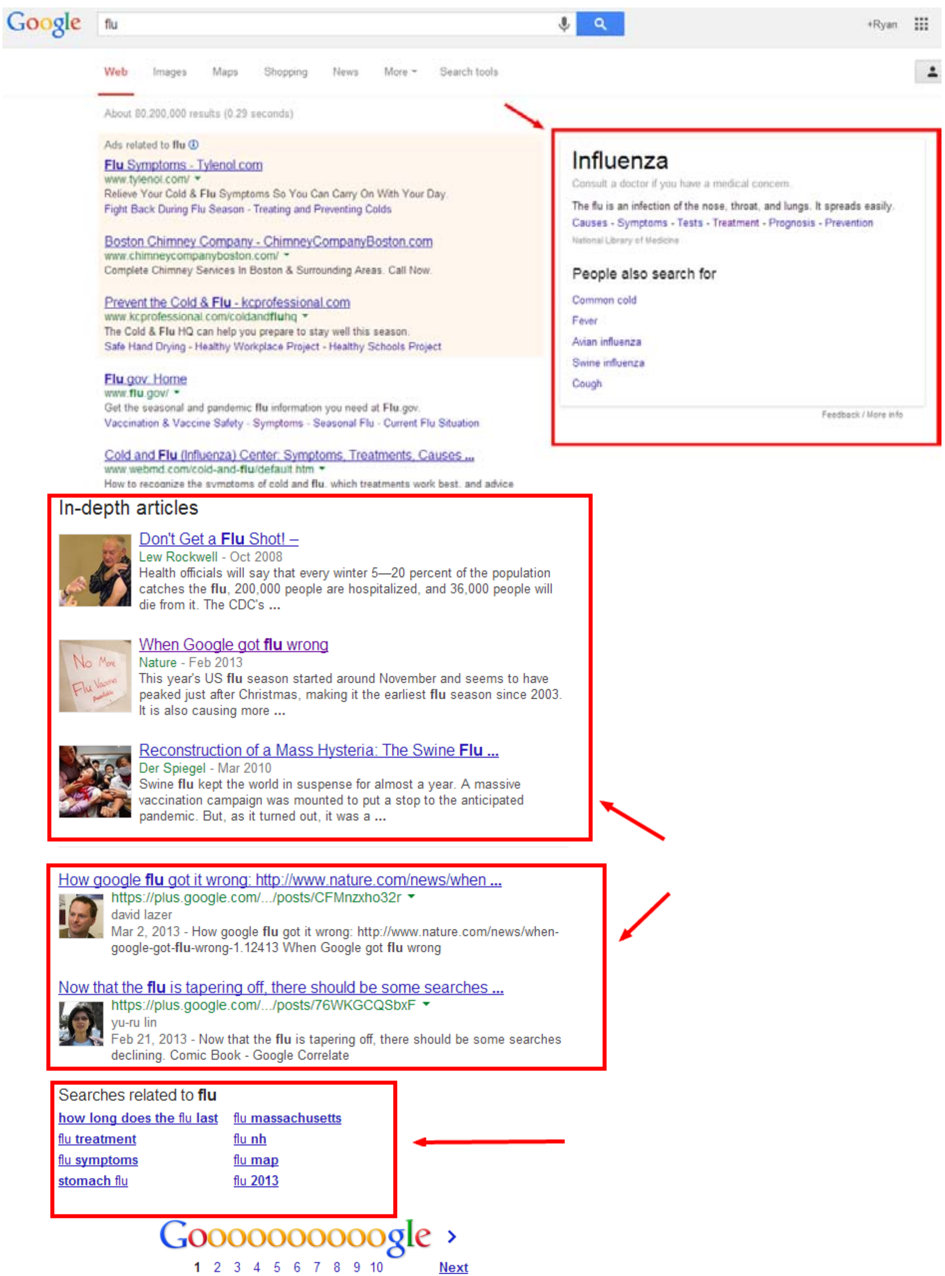
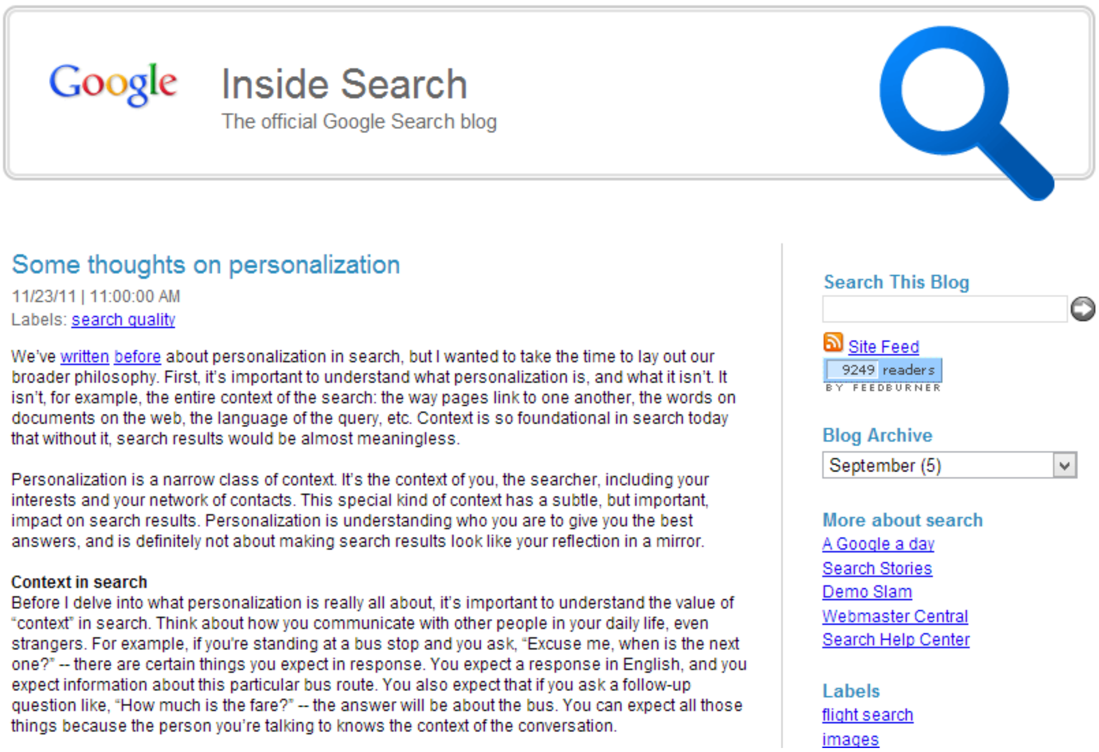


Figure S14 deals with search personalization, a topic for which the academic community needs a much greater understanding. This is an excerpt from a blog post about Google's personalization. It discusses how they attempt to use context in order to improve searches. Such context is likely to change the results and the types of searches recommended for consumers.

Fig. S14. Google Discusses Personalization



Google Inside Search
The official Google Search blog

Some thoughts on personalization
11/23/11 | 11:00:00 AM
Labels: search quality

We've written before about personalization in search, but I wanted to take the time to lay out our broader philosophy. First, it's important to understand what personalization is, and what it isn't. It isn't, for example, the entire context of the search: the way pages link to one another, the words on documents on the web, the language of the query, etc. Context is so foundational in search today that without it, search results would be almost meaningless.

Personalization is a narrow class of context. It's the context of you, the searcher, including your interests and your network of contacts. This special kind of context has a subtle, but important, impact on search results. Personalization is understanding who you are to give you the best answers, and is definitely not about making search results look like your reflection in a mirror.

Context in search
Before I delve into what personalization is really all about, it's important to understand the value of "context" in search. Think about how you communicate with other people in your daily life, even strangers. For example, if you're standing at a bus stop and you ask, "Excuse me, when is the next one?" -- there are certain things you expect in response. You expect a response in English, and you expect information about this particular bus route. You also expect that if you ask a follow-up question like, "How much is the fare?" -- the answer will be about the bus. You can expect all those things because the person you're talking to knows the context of the conversation.

Search This Blog
Site Feed
9249 readers
BY FEEDBURNER
Blog Archive
September (5)
More about search
A Google a day
Search Stories
Demo Slam
Webmaster Central
Search Help Center
Labels
flight search
images

Source: http://insidesearch.blogspot.com/2011/11/some-thoughts-on-personalization.html (*44*)

Finally, Fig. S15 is included to point out that Google may not only experience “blue team” dynamics in its search patterns. Indeed, some companies hire engineers specifically to reconstruct Google’s search algorithm so that they can be at the top of search results. There are now companies who, as part of their marketing services, will try to change placement on search results, and, of course, Google also sells this space with its sponsored content pattern. The item shown in this illustration is the introduction of hot searches. Much like Twitter’s trending hashtags, this would seem to be an area ripe for manipulation by campaigns and companies interested in getting media attention for their brand.

Fig. S15. Google Announcement of Search Trends Feature

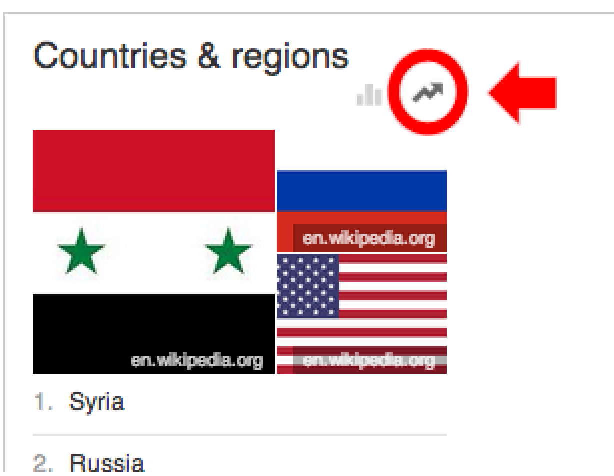
New ways to explore what's trending on Google

9/12/13 | 8:55:00 AM

Whether you're looking for trending celebrities, a monthly recap of what's hot, or power tools to make your own discoveries about what's piquing the world's curiosity - today you'll find new features in Google Trends to make it easier to explore hot topics in Google Search.

Trending Top Charts.
In May we added a new feature to Google Trends called "Top Charts," where you can explore real-world people, places and things ranked by overall search interest in the United States (with more countries coming soon). These "Most Searched" lists span dozens of areas from athletes to cities to cocktails. We've heard great feedback from people who want "Trending" lists -- not just what's most searched overall, but what's spiking compared with usual search volumes. Starting today, you can explore these new Trending Top Charts for a number lists across entertainment, sports, politics and more.

For example, while it may come as no surprise that the United States is the most searched country among people in the U.S., it's more interesting that Syria and Russia were the two top trending countries last month. To see the new "Trending" charts, click the arrow icon at the top of any supported Top Chart.



Source: http://insidesearch.blogspot.com/2013/09/new-ways-to-explore-whats-trending-on.html (*45*)

4. Search Dynamics – Which Search Terms Likely Led to GFT's Inaccuracy and the Problem of Replicating GFT's Results.

We start with the last topic listed in the title—the difficulty replication GFT's results. We attempted in vain to find the 45 search terms that were utilized by GFT. They were not listed in the supplemental materials for any of the published articles on GFT, nor were they available from any online sources that we could locate. The original *Nature* article (*13*) listed 12 categories of search terms, but only listed examples of search terms that were not utilized (e.g., high school basketball). A later article, published by *PLOS One*, did give some examples of the search terms utilized (*14*). The eight examples listed were symptoms of bronchitis, pnumonia [sic], fever, early signs of the flu, robitussin, influenza a, amoxicillin, and strep throat (p. 2). We plotted the time trend for these terms in Fig. S16 using results from Google Trends. It is notable that, with the exception of "influenza a," none of these examples correlated well with either the GFT of the CDC data. Although it is possible that this is a result of the aggregation mechanism used for GFT, the fact that the technology designed to allow others to access the power that made GFT produces results so at odds with the information provided by the authors gave us pause. We cannot be sure if these are really examples used to produce GFT or if the examples were purposefully misleading. A later article discussing GFT's 2013 update seems to suggest that they were indeed purposefully misleading (*15*).

Fig. S15. Google Trend Searches for Terms Indicated in *PLOS One* Article (*14*). Data were acquired by typing search terms into Google Trends (http://www.google.com/trends/) (*39*) and downloading the associated data. The downloaded data are available in the SOM/SOM4/FigS15 folder of the replication materials.

:

Also troubling is that none of those sample terms, at least in their verbatim form, came up when we used Google Correlate to find the most related terms with different time series. Table S3 lists the top correlated terms found by Google Correlate for the GFT data, the GFT data post-2011, the CDC data, and the CDC data pre-2009 [what is received when one clicks the link "…match the pattern of actual flu activity (this is how we made Google Flu Trends!)"] (*46*). As one can note from this table, none of the eight example terms are listed in the top 50 for any of these time series. The table also lists the relative ranks of the different terms and a rough categorization that we put together for analysis. This provides the first clues for which terms were responsible for GFT's errors.

Table S3. List of Search Terms and Classification. Search terms compiled from Google Correlate (https://www.google.com/trends/correlate) (*46*) using national-level time series data. The "replication" of GFT using Google Correlate (CDC pre-2009) is located at (https://www.google.com/trends/correlate/search?e=id:20xKcnNqHrk&t=weekly#) (*47*). Data used for this table are available in the replication materials in folder SOM/SOM4/TabS3.

Search Text	**Rank for GFT**	**Rank for CDC**	**Rank for CDC Pre-2009**	**Rank for GFT Post-2011**	**Classification of Search**
influenza type a		1	1		Term for Influenza
flu duration		2	3	37	General Information
flu fever		3	5	39	Flu Symptoms
treating flu		4	23		Remedies/Treatments
braun thermoscan		5	40		Flu Diagnosis
fever flu		6	33		Flu Symptoms
flu recovery		7	10	34	General Information
flu vs. cold		8	24		Flu Diagnosis
cold or flu	2	9	11	18	Flu Diagnosis
treating the flu		10			Remedies/Treatments
Oscillococcinum	29	11	34		Related Diseases
flu versus cold		12	49		Flu Diagnosis
flu remedies	6	13	35	23	Remedies/Treatments
cold versus flu		14			Flu Diagnosis
human temperature		15	26		Flu Diagnosis
contagious flu	17	16	31		General Information
type a influenza		17	46		Term for Influenza
flu or cold	3	18	13	48	Flu Diagnosis
flu contagious	1	19	4	8	General Information
Thermoscan		20			Flu Diagnosis
flu cough		21			Flu Symptoms
influenza incubation period		22			General Information
duration of flu	40	23			General Information

Search Text	Rank for GFT	Rank for CDC	Rank for CDC Pre-2009	Rank for GFT Post-2011	Classification of Search
cold vs flu	8	24	39	24	Flu Diagnosis
influenza a		25			Term for Influenza
low body temperature		26			Flu Symptoms
flu headache		27		42	Flu Symptoms
flu complications		28			Complications
flu stomach		29			Flu Symptoms
cold and flu symptoms		30			Flu Symptoms
flu and fever		31		31	Flu Symptoms
cold vs. flu		32			Flu Diagnosis
treatment for flu	9	33	25		Remedies/Treatments
treatment of flu		34			Remedies/Treatments
ear thermometer		35			Flu Diagnosis
how long does the flu last?		36			General Information
flu in children		37			General Information
influenza incubation		38			General Information
flu length		39			General Information
length of flu		40			General Information
type a flu		41			Term for Influenza
getting over the flu		42		29	Remedies/Treatments
treat flu	7	43	16		Remedies/Treatments
Robitussin ac		44			Remedies/Treatments
treatment for the flu	48	45			Remedies/Treatments
influenza symptoms		46	38		General Information
what is influenza		47			General Information
flu care		48			Remedies/Treatments
expectorant		49			Remedies/Treatments
cold symptoms		50			Related Diseases
flu germs	33				Term for Influenza
cure flu	27				Remedies/Treatments
how to treat flu	18				Remedies/Treatments
how to get rid of the flu	49			1	Remedies/Treatments
get rid of the flu	36		45	2	Remedies/Treatments
fever reducer			22		Remedies/Treatments
i have the flu	37		47	26	General Information
flu treatment			20		Remedies/Treatments
dangerous fever			27		Flu Symptoms
remedies for the flu	12			7	Remedies/Treatments

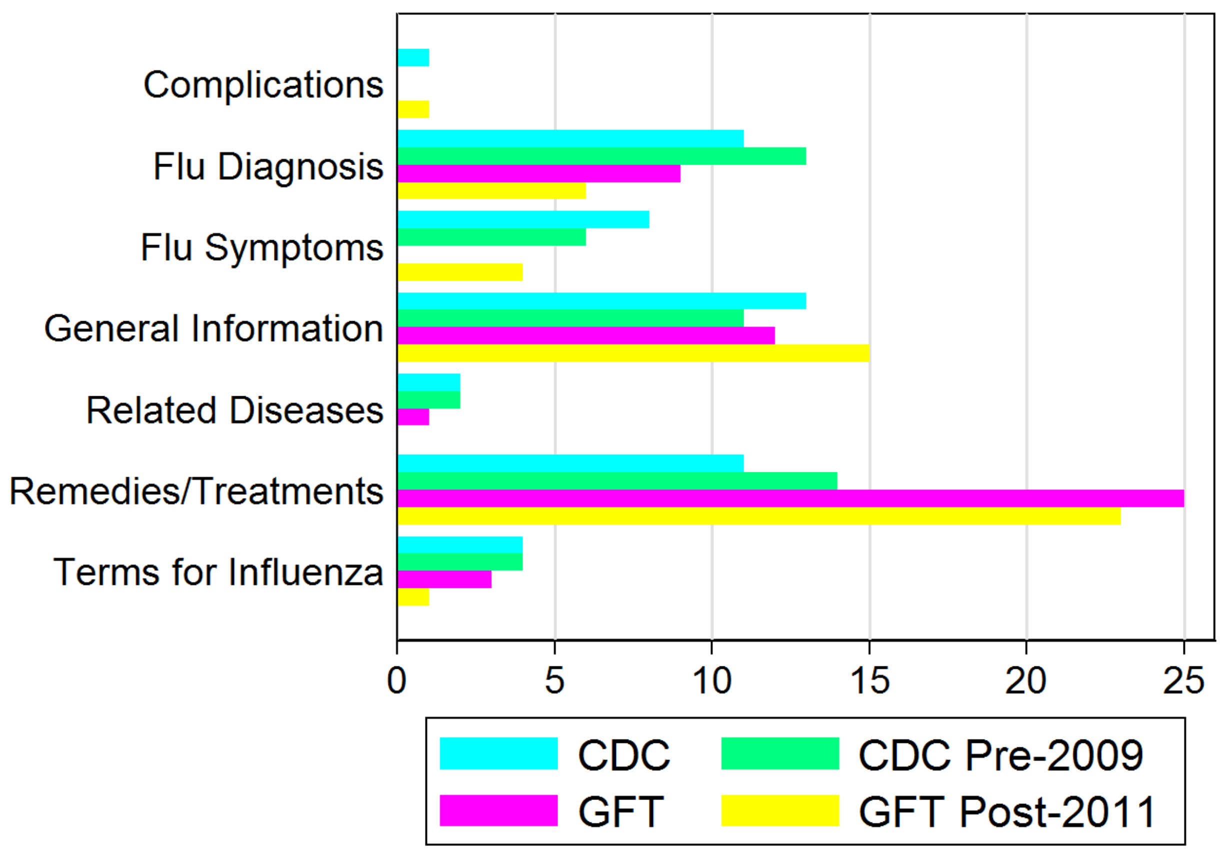
Search Text	Rank for GFT	Rank for CDC	Rank for CDC Pre-2009	Rank for GFT Post-2011	Classification of Search
medicine for the flu	25			5	Remedies/Treatments
how long does it take to get the flu				33	General Information
how to cure the flu	38			19	Remedies/Treatments
flu in toddlers				27	General Information
symptoms of flu	13		2		General Information
treat the flu	4		6	36	Remedies/Treatments
bronchitis			50		Related Diseases
reduce fever			19		Remedies/Treatments
when is the flu contagious	32			28	General Information
flu a				35	Term for Influenza
the flu	20		28		Term for Influenza
home remedies for flu				14	Remedies/Treatments
flu medicine	19		12	3	Remedies/Treatments
influenza a and b	23		30		Term for Influenza
medicine for flu	34				Remedies/Treatments
taking temperature			48		Flu Diagnosis
cure the flu	16			13	Remedies/Treatments
signs of the flu			8		Flu Diagnosis
flu and cold	10				Flu Diagnosis
do i have the flu	28				Flu Diagnosis
flu treatments			37		Remedies/Treatments
how long is the flu contagious	46				General Information
is flu contagious	30		15	38	General Information
how long does the flu last			32		General Information
fight the flu	44				Remedies/Treatments
normal body			14		Flu Diagnosis
home remedies for the flu	39			11	Remedies/Treatments
best flu medicine				20	Remedies/Treatments
flu relief				44	Remedies/Treatments
signs of flu			42		Flu Diagnosis
symptoms of influenza	41				Flu Diagnosis
how to treat the flu	5		7	40	Remedies/Treatments
flu swab				10	Flu Diagnosis
how long is flu contagious			36	47	General Information
what to do for the flu	42				Remedies/Treatments
how long are you contagious				9	General Information
how long am i contagious				15	General Information

Search Text	**Rank for GFT**	**Rank for CDC**	**Rank for CDC Pre-2009**	**Rank for GFT Post-2011**	**Classification of Search**
body temperature			17		Flu Symptoms
flu home remedies	24			21	Remedies/Treatments
normal body temperature			44		Flu Symptoms
get rid of flu				4	Remedies/Treatments
over the counter flu medicine	31			50	Remedies/Treatments
the flu virus	43				General Information
how to get rid of flu				16	Remedies/Treatments
is the flu contagious	35		18	32	Remedies/Treatments
how long do flu symptoms last				41	General Information
cold and flu	15			43	Flu Diagnosis
pregnant with the flu				22	Complications
incubation period for the flu	47				General Information
exposure to flu	50				General Information
how long does flu last	21		43		General Information
flu vs cold	11		21		Flu Diagnosis
natural flu remedies	45				Remedies/Treatments
symptoms of the flu			9		Flu Diagnosis
am i contagious				6	General Information
how long flu last				45	General Information
difference between cold and flu	22				Flu Diagnosis
flu how long				25	General Information
flu test				49	Flu Diagnosis
how to get over the flu	26			17	Remedies/Treatments
best medicine for flu				46	Remedies/Treatments
remedies for flu	14		29	12	Remedies/Treatments
fever cough			41		Flu Symptoms
flu without fever				30	Flu Symptoms

.

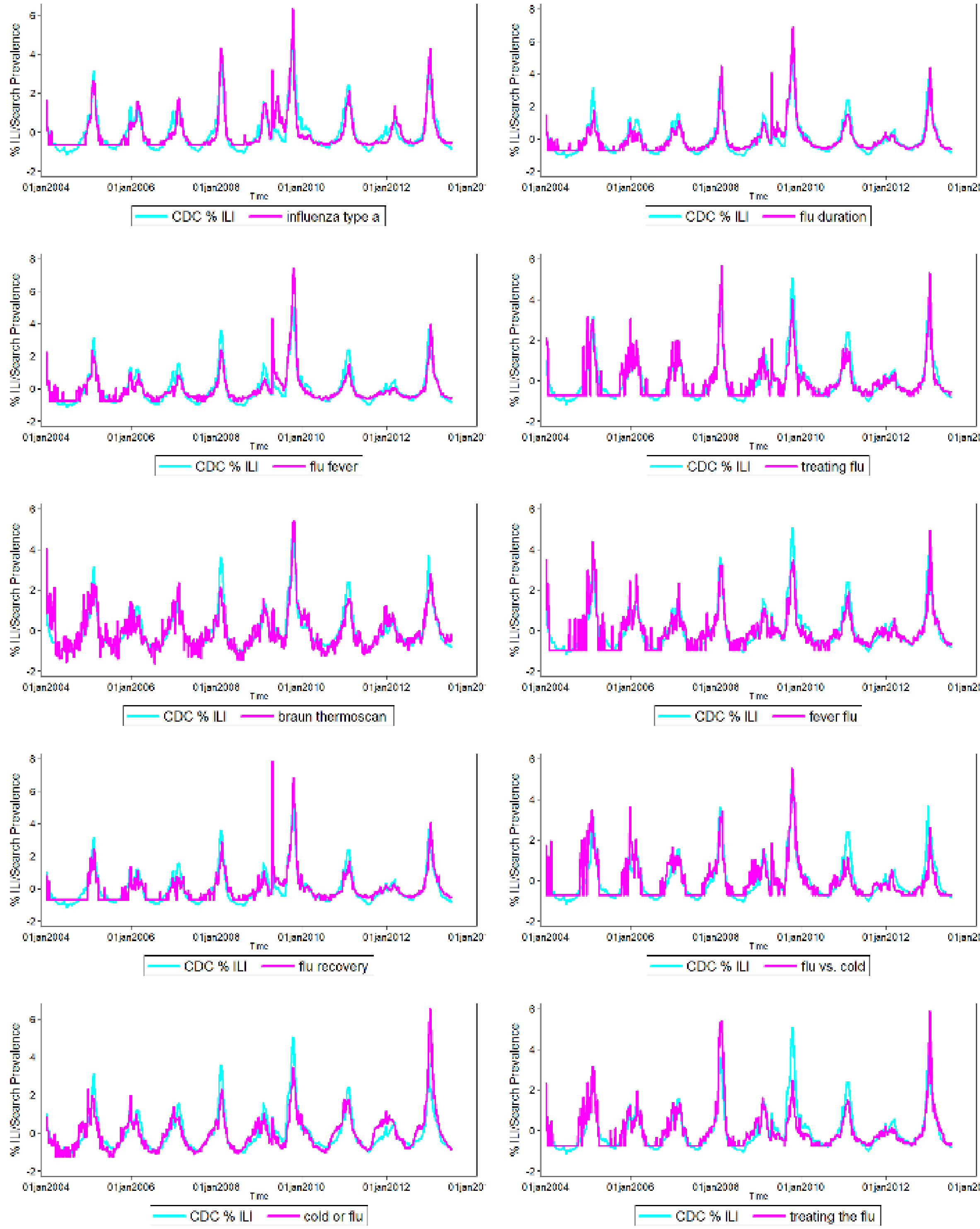
One notable pattern that is immediately apparent is that the number and rank of terms dealing with "treatment for flu" and its variants are much higher for the GFT data than for the CDC data. Indeed, Fig. S16 shows that the number of terms in the category of "Remedies/Treatments" that fall within the top 50 for GFT time series is about double the number of treatment related terms that fall in the top 50 for the CDC time series. This would seem to suggest that an increased prevalence of searches for remedies and treatments was a likely culprit in throwing GFT's count off.

Fig. S16. Category of Top-50 Search Terms From Google Correlate



Further evidence for this hypothesis can be found by looking at the pattern for the top-10 terms correlated with the CDC's full time series of data (Fig. S17). Most of the terms that fall into the top 10 continue to resemble the CDC's data throughout the time period, including after 2011, when GFT started to persistently miss high. Two types of terms are the exception to this rule. The first are terms related to remedies and treatments for the flu. Searches for "treating flu" and "treating the flu" are both very high for the 2012–2013 flu season. The other search result that is noticeably high for this period is the search for "cold or flu." Looking back on the table above, there are a large number of variations of "cold or flu," "flu or cold," etc. that seem to show up with the GFT data and not with the CDC data. After searching through a number of additional terms, an increase in these two search terms seems the most likely explanation for GFT's increasing errors. As we note in the Policy Forum , trying to pin down the exact causes is impossible without at least having the original search terms—and this is why we emphasize replication—but these two seem very likely candidates for explaining what happened.

Fig. S17. Top-10 Google Correlate Search Terms with CDC Data. Data acquired from the best correlated search terms with CDC percent ILI data as reported by Google Correlate (https://www.google.com/trends/correlate) (*46*). Replication data available in SOM/SOM4/FigS17 folder of replication materials.



Another factor that may have contributed to the error is that the search terms used may not be prevalent in the first place. Again, taking the authors at their word, we explore the relative prevalence of the eight example search terms from the *PLOS One* article (*14*). Using "fever," by far the most prominent term, as the baseline, we can see in Fig. S18 that most of these terms are searched relatively rarely. What this means is that, although they may have had a high correlation, they are unlikely to be robust to changes in search patterns. A relatively small change in their frequency may have a large effect on their relative frequency. Even the most searched example term, "fever," is still a relatively rare search compared to general searches about "flu" (Fig. S19). The general search for "flu" should also give us pause when talking about the common explanation for GFT's miss—that there was massive media attention given to the flu in 2012–2013, which resulted in more flu-related searches. First, searches for "flu" peaked in 2009, which is what one might expect given that this is when the swine flu panic took place. We should also note that not all flu-related searches spiked in 2012–2013, only a select few did. And these select few do not seem to be a random sample of the overall search terms. This does not seem to fit well with the media panic story.

Fig. S18. Relative Frequency of Search Terms from GFT's *PLOS One* Article (*14*). Data derived from search prevalence of terms as reported by Google Trends (www.google.com/trends/) (*39*). Data in SOM/SOM4/FigS15 folder of replication materials.

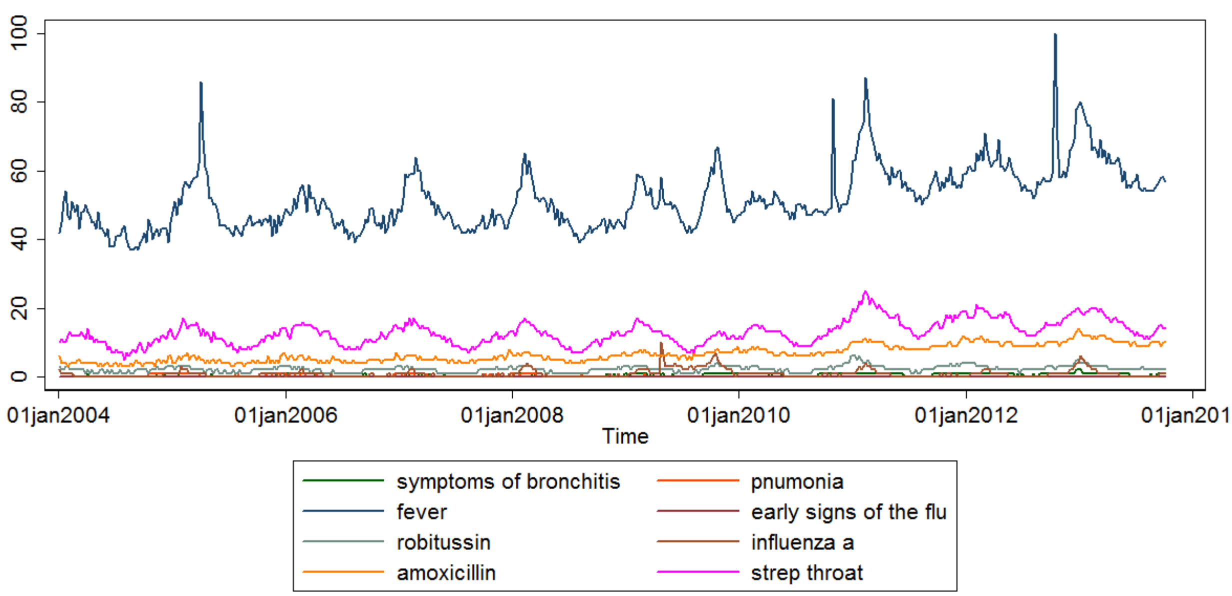
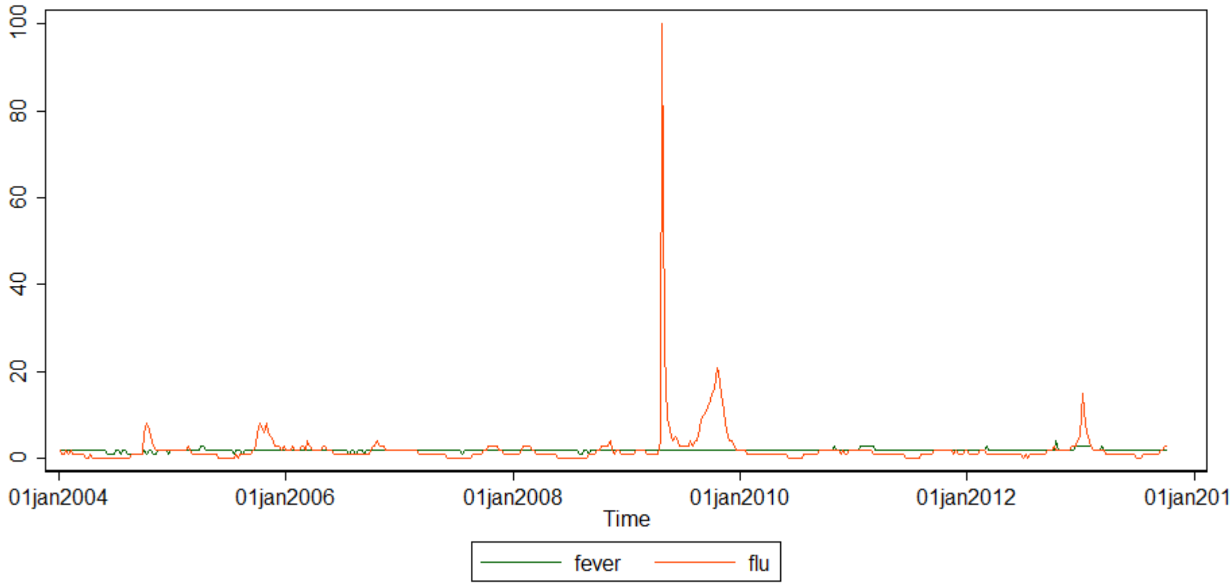


Fig. S19. Relative Frequency of Searches for "Fever" Versus "Flu." Data derived from search prevalence of "fever" and "flu" as reported by Google Trends (www.google.com/trends/) (*39*). Data available in the SOM/SOM4/FigS19 folder of the replication materials.



5. A Note on GFT's 2008 Model

In the Policy Forum, we made reference to GFT's 2008 algorithm, which was substantially modified after the 2009 influenza A H1N1 pandemic. The data from this algorithm are no longer available from Google Flu. A previous study made use of data downloaded in August 2009 to conduct an analysis of the algorithm's performance and documents its problems during the 2009 pandemic. Today, however, the only available data are from March 2009 as made available through the Internet archive (aka "The Wayback Machine"). It can be downloaded from the following link:
http://web.archive.org/web/20090303211715/http://www.google.org/about/flutrends/download.html (*48*).

We make these data available, along with the global estimates, in the replication materials, so that they can be preserved. They can be found in folder /SOM/SOM5/.

6. Lag Models Using CDC Data 3 and 4 Weeks Out

One criticism that could be leveled against both our models using lagged CDC data and those that have been analyzed previously is that there is unmodeled data "vintaging" within the CDC data (i.e., the data are subsequently corrected by the reporting agency and the originally reported data are usually not made available). Similar vintaging is common in economic measures, which are regularly retrospectively revised.

If early errors in CDC data reporting are randomly distributed, then this would not pose much of a problem for our results. Nonetheless, it is worthwhile for us to conduct robustness checks using longer lags to make sure we are not receiving artificially high results because of using already corrected data.

Unsurprisingly, the best-fitting models with only 3+ week lags are a little different from those in the Policy Forum. For this section, our lagged CDC variable model is specified as

$$flu_t = \alpha + \beta_1 flu_{t-3} + \beta_2 flu_{t-4} + \beta_3 flu_{t-5} + \sum_{i=1}^{52} \gamma_i \, week_{it}$$

where *flu* is the CDC estimated percent of doctors' visits for ILI and *week* is coded 1 if the week is number *n* in a year.

The combination of GFT and the CDC data is specified as

$$flu_t = \alpha + \beta_1 gflu_t + \beta_2 flu_{t-3} + \beta_3 flu_{t-4} + \beta_4 (gflu_{t-3} - flu_{t-3}) + \sum\nolimits_{i=1}^{52} \gamma_i week_{it}$$

where *gflu* is the GFT estimate of ILI.

Again, the regression models were initially specified using only data before the September 2009 launch of the new GFT and were tested on data produced subsequently. The out-of-sample predictions are made on a rolling basis, with predictions for the next time step (t + 1) based on estimates from data on all previous time periods ($t_i \leq t$).

Despite using further out lags, these models still outperform the GFT data by itself. For the out-of-sample time period, MAE for the GFT data, by itself, is 0.486. For the model using only lagged CDC data, the MAE is 0.412. For the model, using a combination of GFT's estimates and the lagged CDC data, the MAE is 0.303, or about a 38% decrease in error.

Code for replicating these results can be found in the SOM/SOM6 folder of the replication materials. The code is in the file SOM6 (Replication Code).do. The data is located in the Manuscript folder and is ParableOfGFT(Replication).dta (*36*). All calculations were done in Stata.

7. Full Results and Sensitivity Tests

For space reasons, full results of the models using lagged CDC data and combinations of the CDC data and GFT data were not included. In this section, we present not only the models underlying the graphs in the Policy Forum, but also conduct a general sensitivity test to show how sensitive, if at all, these results are to different specifications. The general sensitivity test should reassure readers about potential over-fitting due to the inclusion of higher-order autoregressive variables or about difficulties in comparing models with different numbers of features.

In the table below, we compare the following model specifications:

$$flu_t = gflu_t \tag{1}$$

$$flu_t = \alpha + \beta_1 flu_{t-2} \tag{2}$$

$$flu_t = \alpha + \beta_1 gflu_t + \beta_2 (gflu_{t-2} - flu_{t-2}) \tag{3}$$

$$flu_t = \alpha + \beta_1 flu_{t-2} + \textstyle\sum_{i=1}^{52} \gamma_i \, week_{it} \tag{4}$$

$$flu_t = \alpha + \beta_1 flu_{t-2} + \beta_2 flu_{t-3} + \textstyle\sum_{i=1}^{52} \gamma_i \, week_{it} \tag{5}$$

$$flu_t = \alpha + \beta_1 flu_{t-2} + \beta_2 flu_{t-3} + \beta_3 flu_{t-4} + \textstyle\sum_{i=1}^{52} \gamma_i \, week_{it} \tag{6}$$

$$flu_t = \alpha + \beta_1 gflu_t + \beta_2 (gflu_{t-2} - flu_{t-2}) + \textstyle\sum_{i=1}^{52} \gamma_i week_{it} \tag{7}$$

$$flu_t = \alpha + \beta_1 gflu_t + \beta_2 flu_{t-2} + \beta_3 (gflu_{t-2} - flu_{t-2}) + \textstyle\sum_{i=1}^{52} \gamma_i week_{it} \tag{8}$$

$$flu_t = \alpha + \beta_1 gflu_t + \beta_2 flu_{t-2} + \beta_3 (gflu_{t-2} - flu_{t-2}) + \beta_4 (gflu_{t-3} - flu_{t-3}) + \textstyle\sum_{i=1}^{52} \gamma_i week_{it} \tag{9}$$

where *flu* is the CDC estimate of ILI and *gflu* is the GFT estimate of ILI, *week* is a binary variable indicating week of observation, and β and γ are estimated regression coefficients. Specifications 1, 6 and 9 are the ones presented in Fig. S20.

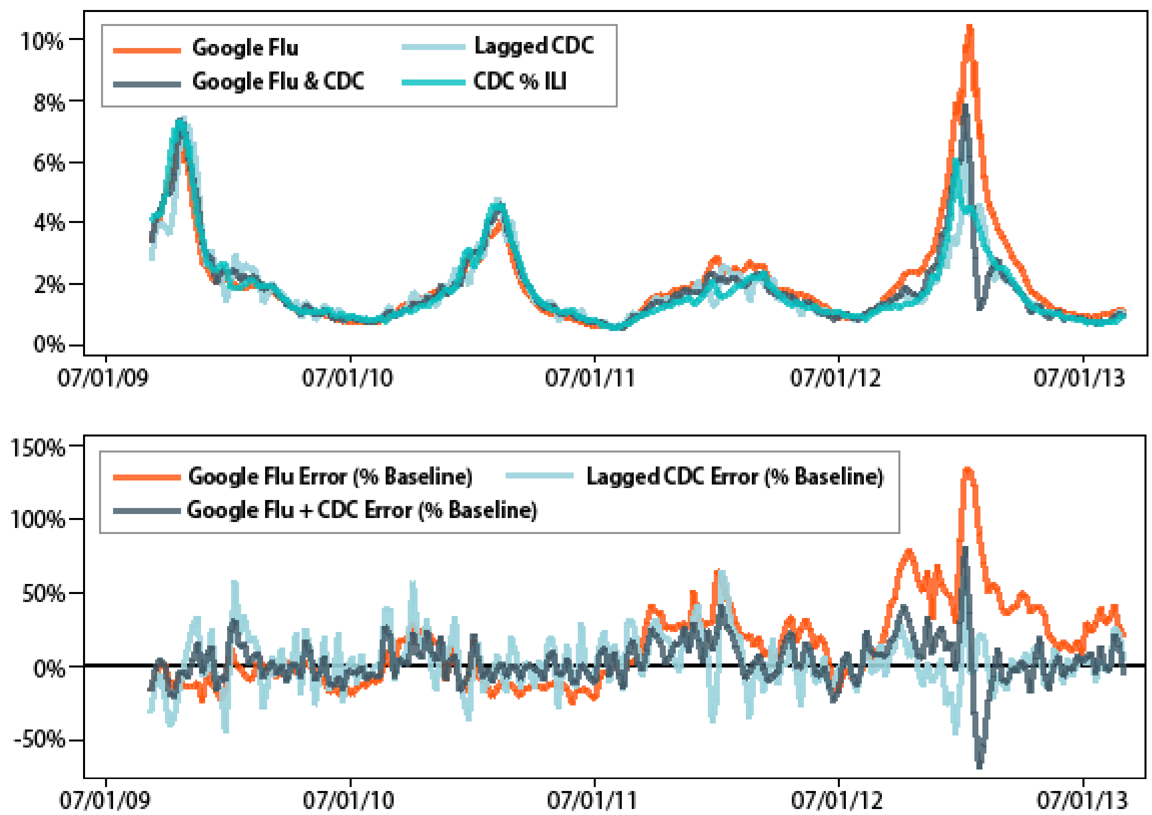


Fig. S20. CDC estimates of doctor visits for ILI compared with GFT, lagged estimates, and 52-week seasonality. (**Top**) CDC estimates of percent doctor visits for ILI overlaid on estimates by GFT, lagged CDC estimates, and 52-week seasonality variables, and a combination of GFT, lagged CDC estimates, lagged error of GFT estimates, and 52-week seasonality variables. **(Bottom**) Error (as percent of CDC baseline) for GFT, lagged CDC estimates, and the combination of GFT and CDC data.

Data and code for running all of these models and calculating their MAE and RMSE can be found in the /Manuscript/ folder of the replication materials and can be produced using the ParableOfGFT(Replication).dta data with the ParableOfGoogle Flu (Replication Code).do code file. The code for calculating statistical significance from this data can be found in the /SOM/SOM7/ folder of the replication materials [SOM7(Replication Code).do] (*36*). All calculations were done in Stata.

As was done for the Policy Forum, all estimates are originally trained on the data before the release of the 2009 GFT update (6 September 2009). From that point on, predictions are done on a rolling basis, similarly to the way such a system would likely be deployed in the real world. Predictions in time step $t + 1$ are based on estimates from data on all previous time periods, $t_i \leq t$.

As the reader can see clearly in the table below, all of the specifications are an improvement on the GFT estimate in the out-of-sample (after 6 September 2009) period. The differences between all other models shown in Eqs. (2) to (9) and the GFT model are also statistically significant.

The reader can also see that the models that combine the information from GFT with the information from the CDC (Eqs. 3, 7, 8, and 9) perform much better than their counterparts that use either the CDC data or the GFT data alone. Although the models presented in the Policy Forum are overspecified, they are not overfit, as their performance in this out-of-sample data is still an improvement over less highly specified models using the same data, although these differences are not always statistically significant.

Measures of statistical significance will vary based on the loss function for the errors. In the Policy Forum, we report MAE, which is resistant to outliers. Table S4 reports these differences along with a sign test for statistical significance. The sign test evaluates the hypothesis that the median difference between the out-of-sample errors is zero. Although this is not a very powerful test, it makes few assumptions about the distribution of the data and is resistant to outliers. Other scholars have suggested that this makes it an attractive significance test for comparing forecast errors across nonnested models [e.g., (*49*), pp. 254–255]. The test is also somewhat conservative in this context, as the results cannot be driven by a few large GFT misses during the last flu season.

A more common approach in regression forecasts is to report the RMSE and assume a squared error loss function (i.e., larger misses are more harmful than smaller misses). This is somewhat consistent with epidemiologists' concern with estimating the size of the peak of flu season, which is also where forecast errors are highest. In this situation, a rather intuitive, but somewhat cumbersome, test of the null hypothesis that the mean squared forecast errors of the two models are equal (an *F* test) is suggested by Ashley *et al*. (*50*–*52*). These results, along with the RMSE are reported in Table S5. This test's utility is somewhat limited by its assumption of squared error loss and its being only asymptotically justified (while the sign test is an exact finite-sample test), but the size of our out-of-sample cases is well above what Ashley (*52*) finds to be appropriate.

Table S4. Comparison of Models – MAE with Sign Test for Significance. Values in the cells are the difference in MAE between models with *P* values (two-tailed) based on a sign test in parentheses. Error is estimated for the out-of-sample period (post 6 September 2009) using the dynamic procedure outlined above.

	1	2	3	4	5	6	7	8	9
1	--	0.148 (0.072)	0.237 (0.000)	0.134 (0.000)	0.169 (0.001)	0.174 (0.000)	0.243 (0.000)	0.251 (0.000)	0.254 (0.000)
2		--	0.089 (0.000)	-0.014 (0.128)	0.021 (0.038)	0.027 (0.053)	0.096 (0.000)	0.103 (0.001)	0.106 (0.000)
3			--	-0.103 (0.019)	-0.068 (0.000)	-0.063 (0.000)	0.006 (0.782)	0.014 (0.489)	0.016 (0.580)
4				--	0.035 (0.782)	0.041 (1.000)	0.109 (0.013)	0.117 (0.027)	0.120 (0.027)
5					--	0.005 (0.000)	0.074 (0.000)	0.082 (0.000)	0.084 (0.000)
6						--	0.069 (0.000)	0.077 (0.000)	0.079 (0.000)
7							--	0.008 (0.213)	0.010 (0.489)
8								--	0.003 (0.072)
9									--

Table S5. Comparison of Models – RMSE with F-Test for Significance. Values in the cells are the difference in RMSE between models with *P* values (two-tailed) based on an *F* test in parentheses. Error is estimated for the out-of-sample period (post 6 September 2009) using the dynamic procedure outlined above.

	1	2	3	4	5	6	7	8	9
1	--	0.371 (0.000)	0.435 (0.000)	0.335 (0.000)	0.410 (0.000)	0.413 (0.000)	0.449 (0.000)	0.482 (0.000)	0.487 (0.000)
2		--	0.064 (0.072)	-0.036 (0.139)	0.039 (0.137)	0.041 (0.110)	0.078 (0.032)	0.111 (0.002)	0.116 (0.001)
3			--	-0.101 (0.007)	-0.026 (0.418)	-0.023 (0.463)	0.014 (0.261)	0.046 (0.000)	0.051 (0.000)
4				--	0.075 (0.001)	0.078 (0.001)	0.115 (0.003)	0.147 (0.000)	0.152 (0.000)
5					--	0.003 (0.416)	0.039 (0.239)	0.072 (0.023)	0.077 (0.014)
6						--	0.037 (0.269)	0.069 (0.026)	0.074 (0.016)
7							--	0.033 (0.000)	0.038 (0.000)
8								--	0.005 (0.176)
9									--

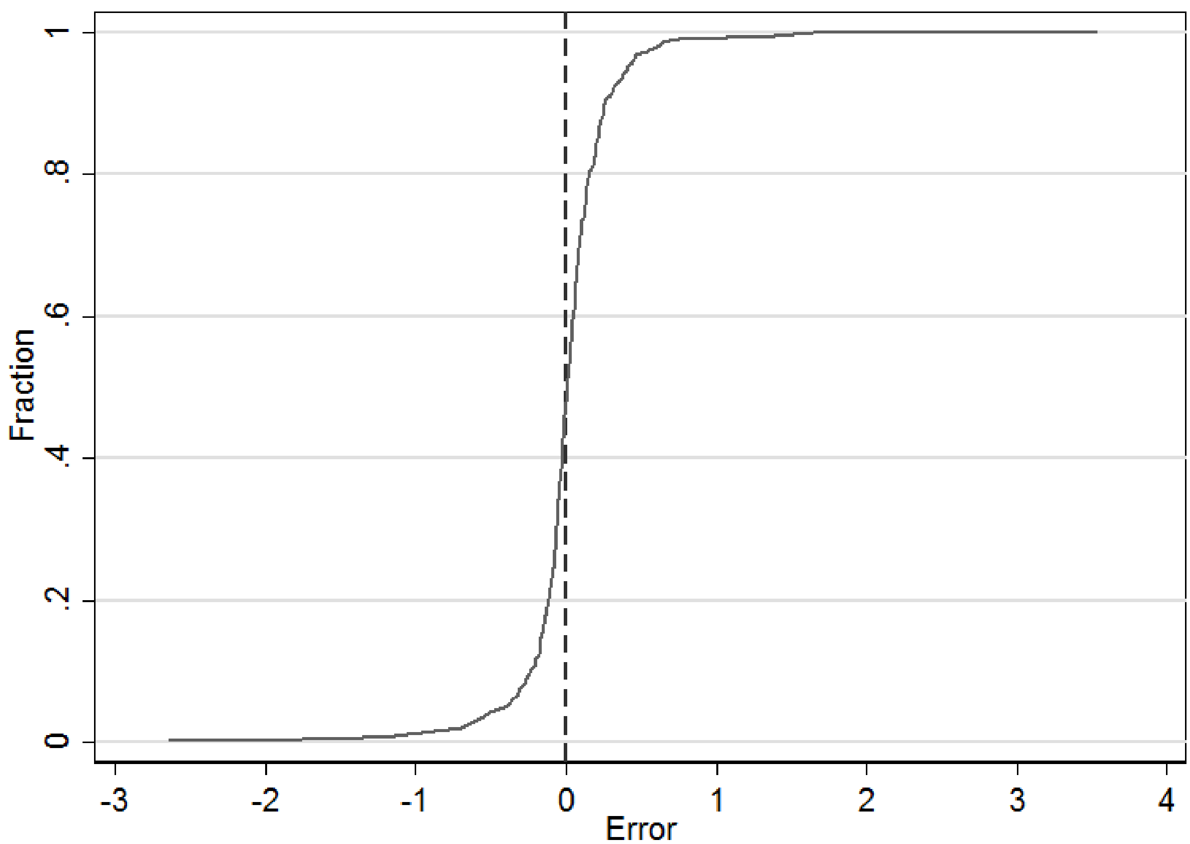
8. Comparison With Nonlinear Models

In this section of the SOM, we utilize a more flexible, nonlinear, technique for modeling the CDC ILI estimate. Linear models have dominated this particular area of analysis [see e.g., *4*, *13*, *14*)], and have clear advantages in simplicity and interpretability. Linear models, however, may be insufficient in the presence of even simple nonlinearities (*53*). This section serves as a check on whether we can gain substantial ground through use of simple nonlinear models.

Replication for all of these results can be done using the code in the /SOM/SOM8/ folder of the replication materials. The code SOM8(Replication Code).do will allow the user to replicate the error plots in conjunction with the ParableOfGFT(Replication).dta data file in the /Manuscript/ folder using Stata (*36*). The neural network models were run using the "nnet" package in the R statistical programming environment (*54*). This code is also available in the /SOM/SOM8/ folder and is labelled SOMpt8(Replication Code).r (*36*).

We begin by conducting the sanity checks suggested by Weigend and Gershenfeld (*53*), p. 44; see also Smith (*55*). The first check is to look at the distribution of errors to see if they are uniform or if there are a few large and unusual outliers. A cumulative distribution (Fig. S21) is plotted by the size of errors from the combination of GFT and CDC data presented in the Policy Forum in the out-of-sample data. While this seems to show the pattern Weigend and Gershenfeld are concerned about (very small errors along with a few large outliers), these outliers are very rare.

Fig. S21. Cumulative Distribution of Errors from Combined GFT and CDC Model



Similarly, Fig. S22 and S23 show the errors plotted against the actual CDC ILI data and the predicted values from the model, respectively. Here again, the errors look relatively uniform with a few relatively large (though not as large as for GFT by itself) exceptions.

Fig. S22. Plot of Errors from Combined GFT and CDC Model Against Actual CDC ILI Values

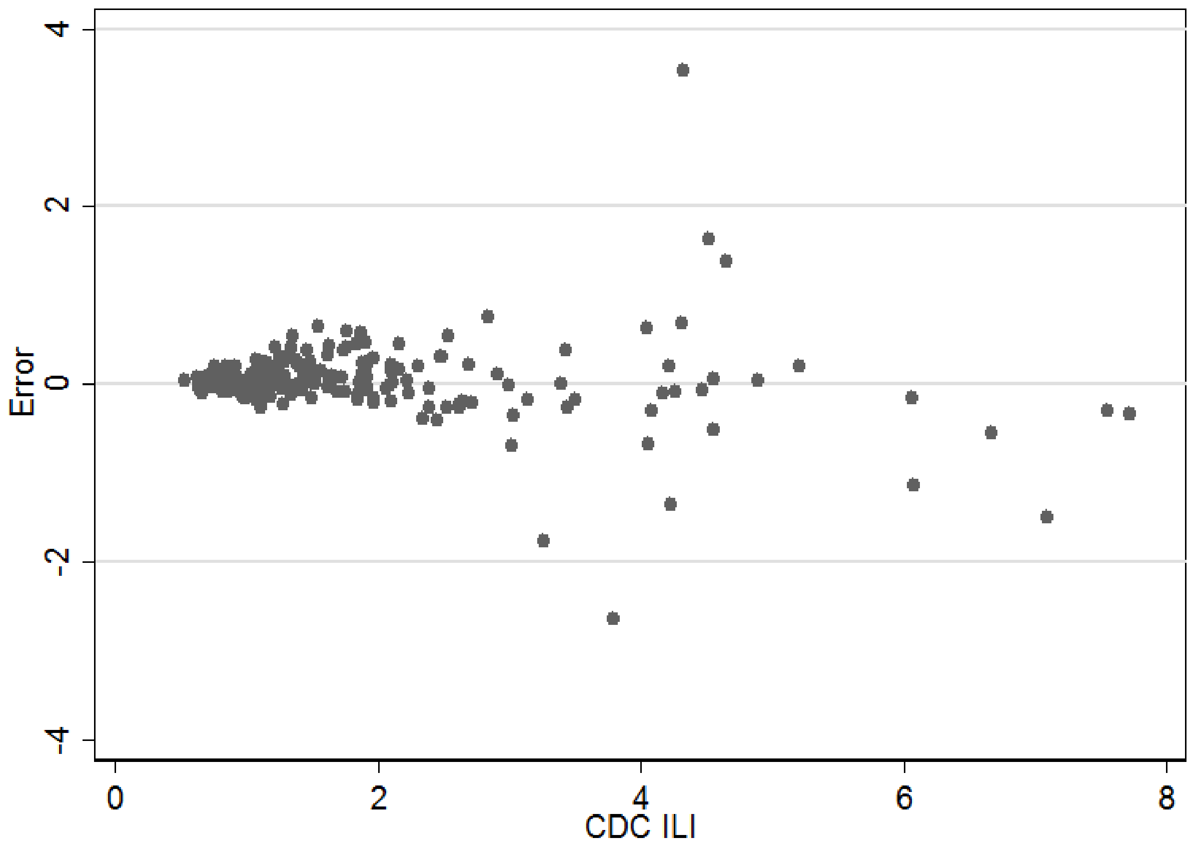
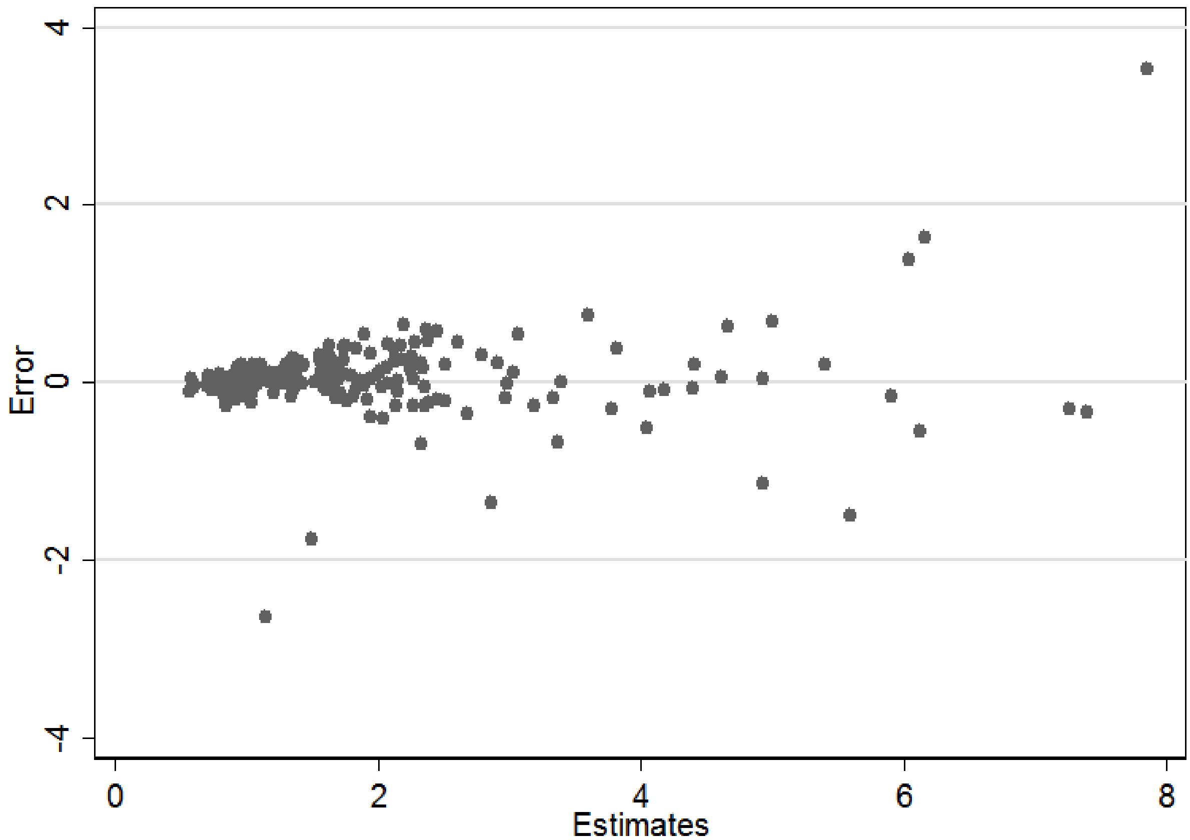


Fig. S23. Plot of Errors from Combined GFT and CDC Model Against Predicted Values



These figures would seem to indicate that there is potential nonlinearity in the system, but the sparsity of these outliers and their location in the out-of-sample data mean that a nonlinear estimator may not produce any significant improvement.

To test this, we used a simple feed-forward neural network model with a single hidden-layer architecture. This choice of model architecture is only one of many possible options. It does, however, provide a good first-cut estimation of the gains to be made from relaxing the linearity assumptions of the models in the Policy Forum.

The greater flexibility of the neural network model requires us to specify an additional tuning parameter, namely, the number of nodes in the hidden layer. To set this parameter in a manner that avoids overfitting, we used leave-one-out cross-validation on the training data. We tested across a range of possible sizes for the hidden layer. Preference is given to the simplest architecture that cannot be significantly improved with an additional node.

Another issue with neural network models (as is common in more flexible models) is that initial values, which are drawn using a quasi-random number generator, can subtly influence the results, especially when the amount of data available for training is not especially large. We thus evaluate all models over 50 runs to ensure that our results are not due to the behavior of random number selection (*56*).

As in the Policy Forum, the in-sample period is before the fielding of the 2009 GFT model, with the out-of-sample period estimated on a rolling basis—mimicking the deployment of GFT.

The inputs for our three models are as follows:

$$\{gflu_t, flu_{t-2}, (gflu_{t-2} - flu_{t-2})\} \tag{1}$$

$$\{gflu_t, flu_{t-2}, (gflu_{t-2} - flu_{t-2}), week_{it}\} \tag{2}$$

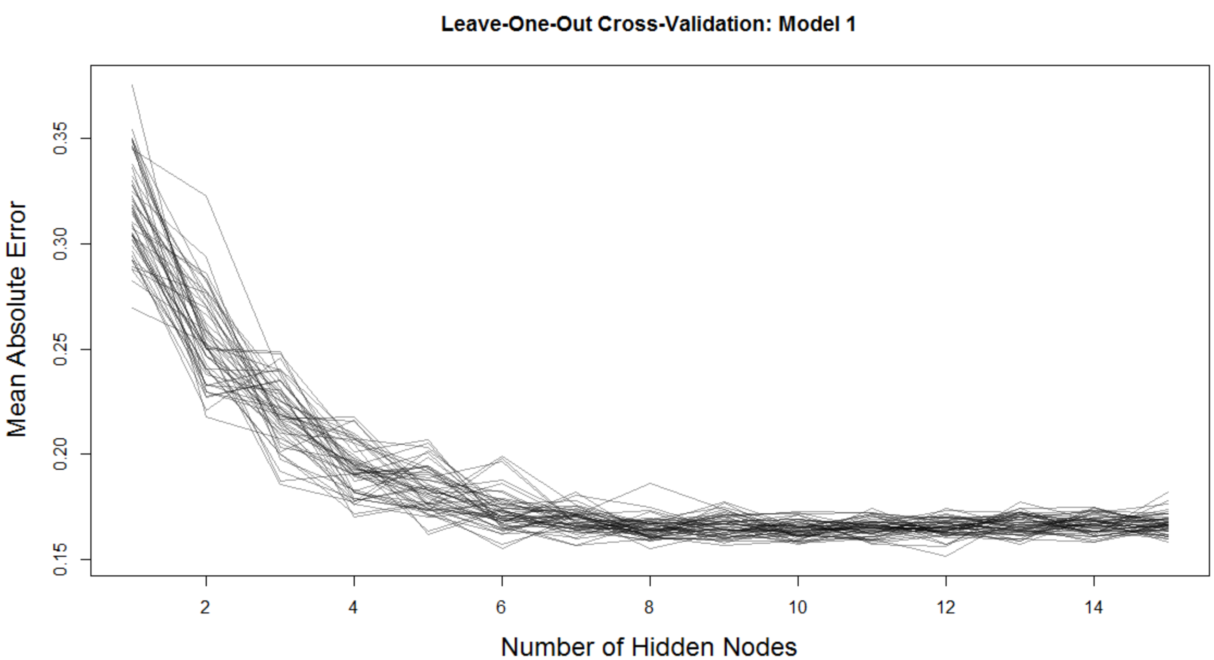
$$\{gflu_t, flu_{t-2}, (gflu_{t-2} - flu_{t-2}), (gflu_{t-3} - flu_{t-3}), week_{it}\} \tag{3}$$

where *flu* is the CDC estimate of ILI and *gflu* is the GFT estimate of ILI.

The variables used in models 2 and 3 are the same as those used in model 8 and 9 in the linear models. The variables for model 1 are a subset of those in model 2 without the seasonality component.

The results of the leave-one-out cross-validation for model 1 are reported in Fig. S24. The lines chart the MAE across each run. Error seems to stabilize at its lowest level with 10 hidden nodes.

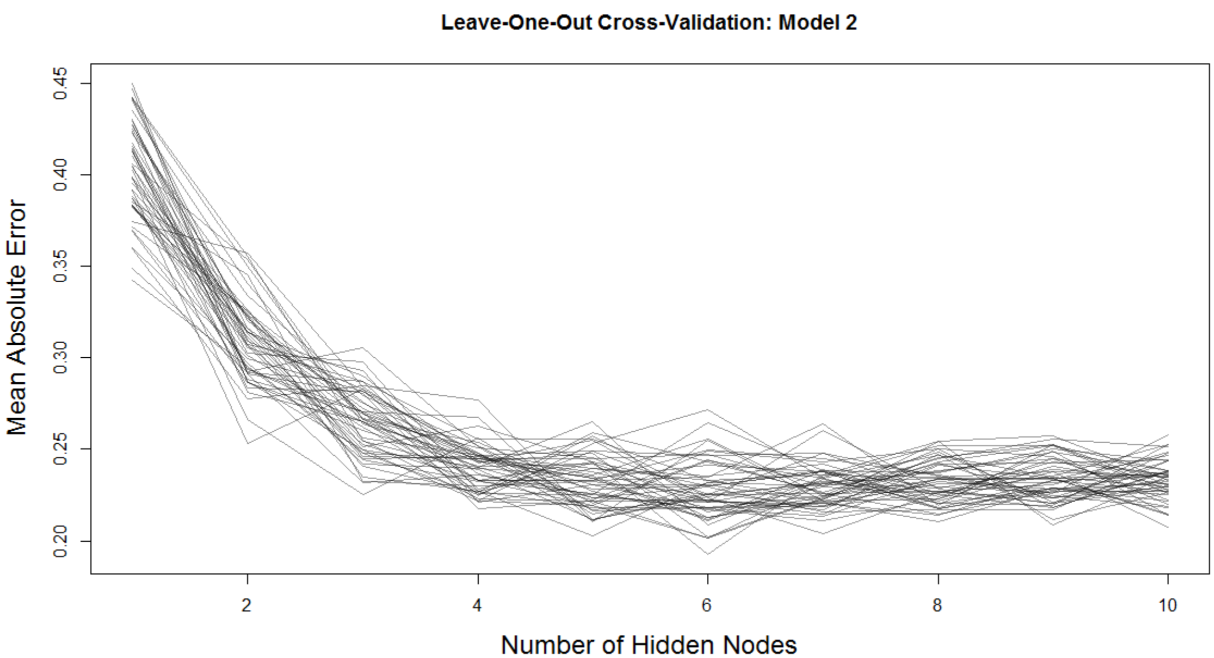
Fig. S24. Leave-One-Out Cross-Validation for Model 1



After 50 runs of the neural network model with 10 nodes in the hidden layer, we evaluated the MAE of the model. MAE for this neural networks model is 0.249, with a range between 0.215 and 0.318. This is not a substantial improvement over the linear models presented in the Policy Forum.

Model 2 adds in the seasonality component. Here again, we start with the leave-one-out cross-validation to establish the number of hidden nodes. Fig. S25 seems to suggest stabilization around five nodes in the hidden layer.

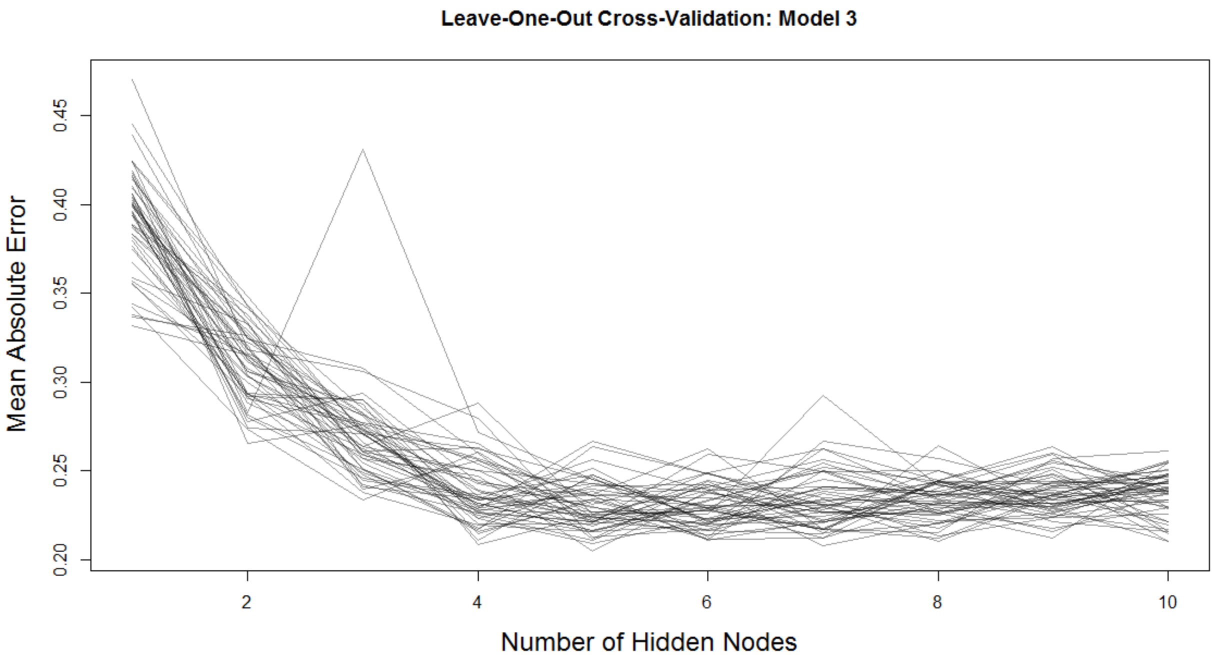
Fig. S25. Leave-One-Out Cross-Validation for Model 2



Here again, however, this model does not, on average, outperform the simple linear model in the Policy Forum. The MAE for this model is 0.240 with a range from 0.217 to 0.297. It is clearly an improvement on the model without seasonality, but it is still not an improvement on the linear model.

The final model uses all of the variables from the combined CDC and GFT model in the Policy Forum. Fig. S26 shows the results of the cross-validation procedure. As with model 2, it appears to stabilize around five hidden nodes.

Fig. S26. Leave-One-Out Cross-Validation for Model 3



Although this model comes close to replicating the error rate of the linear model, it is very difficult to distinguish from the linear model (the average difference is measured in 1/1000ths of a point). The average MAE across 50 runs of the model is 0.236, with a range from 0.198 to 0.273.

Clearly, the results presented here cannot reject the possibility that a nonlinear model will perform better (i.e., perhaps a more complex neural network architecture would improve the results), but it does suggest that the linear models presented in the Policy Forum are very difficult to outperform through simple relaxation of the linearity assumption. We should also note that linear models likely perform well in our situation because we are only predicting a couple of steps into the future (technically, we are "nowcasting"—trying to figure out current levels when measurement lags the event). In studies that have attempt to predict influenza levels months in advance, like those cited in the Policy Forum, nonlinearity looms much larger. It is also possible that, since there were only a few large outliers that occurred primarily in the out-of-sample data, nonlinear prediction will become more pertinent over time.

References and Notes

1. D. Butler, When Google got flu wrong. *Nature* **494**, 155–156 (2013). doi:10.1038/494155a Medline

2. D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, L. Simonsen, Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLOS Comput. Biol.* **9**, e1003256 (2013). doi:10.1371/journal.pcbi.1003256 Medline

3. A. McAfee, E. Brynjolfsson, Big data: the management revolution. *Harv. Bus. Rev.* **90**, 60–66, 68, 128 (2012). Medline

4. S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, D. J. Watts, Predicting consumer behavior with Web search. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17486–17490 (2010). doi:10.1073/pnas.1005962107 Medline

5. A. Tumasjan *et al.*, in Proceedings of the 4th International AAAI Conference on Weblogs and Social Media, Atlanta, Georgia, 11 to 15 July 2010 (Association for Advancement of Artificial Intelligence, 2010), pp. 178–185.

6. J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market. *J. Comput. Sci.* **2**, 1–8 (2011). doi:10.1016/j.jocs.2010.12.007

7. F. Ciulla, D. Mocanu, A. Baronchelli, B. Gonçalves, N. Perra, A. Vespignani, Beating the news using social media: the case study of American Idol. *EPJ Data Sci.* **1**, 8 (2012). doi:10.1140/epjds8

8. P. T. Metaxas *et al.*, in Proceedings of PASSAT—IEEE Third International Conference on Social Computing, Boston, MA, 9 to 11 October 2011 (IEEE, 2011), pp. 165–1171. doi:10.1109/PASSAT/SocialCom.2011.98

9. D. Lazer, A. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, M. Van Alstyne, Social science. Computational social science. *Science* **323**, 721–723 (2009). doi:10.1126/science.1167742 Medline

10. A. Vespignani, Predicting the behavior of techno-social systems. *Science* **325**, 425–428 (2009). doi:10.1126/science.1171990 Medline

11. G. King, Ensuring the data-rich future of the social sciences. *Science* **331**, 719–721 (2011). doi:10.1126/science.1197872 Medline

12. D. Boyd, K. Crawford, Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Inform. Commun. Soc.* **15**, 662 (2012). doi:10.1080/1369118X.2012.678878

13. J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, L. Brilliant, Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009). doi:10.1038/nature07634 Medline

14. S. Cook, C. Conrad, A. L. Fowlkes, M. H. Mohebbi, Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. *PLoS ONE* **6**, e23610 (2011). doi:10.1371/journal.pone.0023610 Medline

15. P. Copeland *et al*., *Int. Soc. Negl. Trop. Dis.* **2013**, 3 (2013).

16. C. Viboud, P. Y. Boëlle, F. Carrat, A. J. Valleron, A. Flahault, Prediction of the spread of influenza epidemics by the method of analogues. *Am. J. Epidemiol.* **158**, 996–1006 (2003). doi:10.1093/aje/kwg239 Medline

17. W. W. Thompson, L. Comanor, D. K. Shay, Epidemiology of seasonal influenza: use of surveillance data and statistical models to estimate the burden of disease. *J. Infect. Dis.* **194**, (Suppl 2), S82–S91 (2006). doi:10.1086/507558 Medline

18. I. M. Hall, R. Gani, H. E. Hughes, S. Leach, Real-time epidemic forecasting for pandemic influenza. *Epidemiol. Infect.* **135**, 372–385 (2007). doi:10.1017/S0950268806007084 Medline

19. J. B. S. Ong, M. I. Chen, A. R. Cook, H. C. Lee, V. J. Lee, R. T. Lin, P. A. Tambyah, L. G. Goh, Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. *PLoS ONE* **5**, e10036 (2010). doi:10.1371/journal.pone.0010036 Medline

20. J. R. Ortiz, H. Zhou, D. K. Shay, K. M. Neuzil, A. L. Fowlkes, C. H. Goss, Monitoring influenza activity in the United States: a comparison of traditional surveillance systems with Google Flu Trends. *PLoS ONE* **6**, e18687 (2011). doi:10.1371/journal.pone.0018687 Medline

21. Organizing lists of related searches, Google; http://insidesearch.blogspot.com/2011/06/organizing-lists-of-related-searches_16.html

22. Improving health searches, because your health matters, Google; http://insidesearch.blogspot.com/2012/02/improving-health-searches-because-your.html

23. E. Mustafaraj, P. Metaxas, in *Proceedings of the WebSci10*, Raleigh, NC, 26 and 27 April 2010 (Web Science Trust, 2010); http://journal.webscience.org/317/.

24. J. Ratkiewicz *et al*., in *Proceedings of 5th International AAAI Conference on Weblogs and Social Media*, San Francisco, CA, 7 to 11 August 2011 (AAAI, 2011), pp. 297–304.

25. G. King, *PS Polit. Sci. Polit.* **28**, 443 (1995).

26. P. Voosen, *Chronicle of Higher Education*, 13 September 2013; http://chronicle.com/article/Researchers-Struggle-to-Secure/141591/.

27. R. Lazarus, K. P. Kleinman, I. Dashevsky, A. DeMaria, R. Platt, Using automated medical records for rapid identification of illness syndromes (syndromic surveillance): the example of lower respiratory infection. *BMC Public Health* **1**, 9 (2001). doi:10.1186/1471-2458-1-9 Medline

28. R. Chunara, S. Aman, M. Smolinski, J. S. Brownstein, Flu Near You: An Online Self-reported Influenza Surveillance System in the USA. *Online J. Public Health Inform.* **5**, e133 (2013). doi:10.5210/ojphi.v5i1.4456

29. D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, A. Vespignani, Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21484–21489 (2009). doi:10.1073/pnas.0906910106 Medline

30. D. L. Chao, M. E. Halloran, V. J. Obenchain, I. M. Longini, Jr., FluTE, a publicly available stochastic influenza epidemic simulation model. *PLOS Comput. Biol.* **6**, e1000656 (2010). doi:10.1371/journal.pcbi.1000656 Medline

31. J. Shaman, A. Karspeck, Forecasting seasonal outbreaks of influenza. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 20425–20430 (2012). doi:10.1073/pnas.1208772109 Medline

32. J. Shaman, A. Karspeck, W. Yang, J. Tamerius, M. Lipsitch, Real-time influenza forecasts during the 2012-2013 season. *Nat. Commun.* **4**, 2837 (2013). doi:10.1038/ncomms3837 Medline

33. E. O. Nsoesie, R. J. Beckman, S. Shashaani, K. S. Nagaraj, M. V. Marathe, A Simulation Optimization Approach to Epidemic Forecasting. *PLoS ONE* **8**, e67164 (2013). doi:10.1371/journal.pone.0067164 Medline

34. A. Hannak *et al.*, in *Proceedings of 22nd International World Wide Web Conference*, Rio de Janeiro, 13 to 17 May 2013 (Association for Computing Machinery, New York, 2013), pp. 527–538.

35. A. J. Berinsky, G. A. Huber, G. S. Lenz, Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk. *Polit. Anal.* **20**, 351–368 (2012). doi:10.1093/pan/mpr057

36. D. Lazer, R. Kennedy, G. King, A. Vespignani, "Replication data for: The Parable of Google Flu: Traps in Big Data Analysis" (2014); http://dx.doi.org/10.7910/DVN/24823 UNF:5:BJh9WzZQNEeSEpV3EWs+xg== IQSS Dataverse Network [Distributor].

37. Fifteen years on—and were just getting started, Google; http://insidesearch.blogspot.com/2013/09/fifteen-years-onand-were-just-getting.html.

38. Health-related searches; Google: http://googleblog.blogspot.com/2009/05/understanding-health-related-searches.html.

39. Google Trends, www.google.com/trends/.

40. Note, About Google Trends; https://support.google.com/trends/?hl=en&rd=2#topic=4365599.

41. L. Rossignol, C. Pelat, B. Lambert, A. Flahault, E. Chartier-Kastler, T. Hanslik, A method to assess seasonality of urinary tract infections based on medication sales and google trends. *PLoS ONE* **8**, e76020 (2013). doi:10.1371/journal.pone.0076020 Medline

42. Our Blog, Epiphany; http://www.epiphanysearch.co.uk/blog/2011/07/warning-google-makes-insights-for-search-data-useless/.

43. Inside Search, Google; http://insidesearch.blogspot.com/2012/11/look-up-medications-more-quickly-and.html.

44. Inside Search, Google; http://insidesearch.blogspot.com/2011/11/some-thoughts-on-personalization.html.

45. Inside Search, Google; http://insidesearch.blogspot.com/2013/09/new-ways-to-explore-whats-trending-on.html.

46. Google Correlate, https://www.google.com/trends/correlate

478. Correlated with ILI (CDC), Google Correlate; https://www.google.com/trends/correlate/search?e=id:20xKcnNqHrk&t=weekly#.

489. Download data, Internet Archive: Wayback Machine; http://web.archive.org/web/20090303211715/http://www.google.org/about/flutrends/download.html.

490. F. X. Diebold, R. S. Mariano, Comparing predictive accuracy. *J. Bus. Econ. Stat.* **13**, 253–265 (1995).

501. R. Ashley, C. W. J. Granger, R. Schmalensee, Advertising and aggregate consumption: An analysis of causality. *Econometrica* **59**, 817–858 (1980).

512. R. Ashley, Inflation and the distribution of price changes across markets. *Econ. Inq.* **19**, 650–660 (1981). doi:10.1111/j.1465-7295.1981.tb00344.x

523. R. Ashley, Statistically significant forecasting improvements: How much out-of-sample data is likely necessary? *Int. J. Forecast.* **19**, 229–239 (2003). doi:10.1016/S0169-2070(01)00139-X

534. A. S. Weigend, N. A. Gershenfeld, Eds., *Time Series Prediction: Forecasting the Future and Understanding the Past* (Santa Fe Institute Series, Addison-Wesley, Reading, MA,1993).

545. B. Ripley, Package "nnet," The Comprehensive R Archive Network (CRAN), http://cran.r-project.org/web/packages/nnet/nnet.pdf (R, Vienna, 2013).

556. L.A. Smith, Does a meeting in Santa Fe imply chaos? in (*523*), pp. 323–343.

567. In some applications this is handled by setting a seed for the random number generator. We eschew this approach here because it is important to know about the stability of the model's predictions in order to avoid a favorable/unfavorable seed influencing the interpretation of the results.