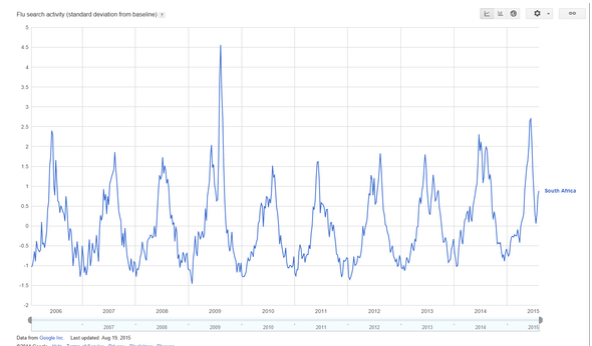# Google Flu Trends

Google Flu Trends (**GFT**) was a web service operated by Google. It provided estimates of influenza activity for more than 25 countries. By aggregating Google Search queries, it attempted to make accurate predictions about flu activity. This project was first launched in 2008 by Google.org to help predict outbreaks of flu.[1]

Google Flu Trends stopped publishing current estimates on 9 August 2015. Historical estimates are still available for download, and current data are offered for declared research purposes.[2]

Google Flu Trends data, South Africa

Contents

History

The idea behind Google Flu Trends was that, by monitoring millions of users' health tracking behaviors online, the large number of Google search queries gathered can be analyzed to reveal if there is the presence of flu-like illness in a population. Google Flu Trends compared these findings to a historic baseline level of influenza activity for its corresponding region and then reports the activity level as either minimal, low, moderate, high, or intense. These estimates have been generally consistent with conventional surveillance data collected by health agencies, both nationally and regionally.

Roni Zeiger helped develop Google Flu Trends.[3]

Methods

Google Flu Trends was described as using the following method to gather information about flu trends.[4][5]

First, a time series is computed for about 50 million common queries entered weekly within the United States from 2003 to 2008. A query's time series is computed separately for each state and normalized into a fraction by dividing the number of each query by the number of all queries in that state. By identifying the IP address associated with each search, the state in which this query was entered can be determined.

A linear model is used to compute the log-odds of Influenza-like illness (ILI) physician visit and the log-odds of ILI-related search query:

$$\text{logit}(P) = \beta_0 + \beta_1 \times \text{logit}(Q) + \epsilon$$

P is the percentage of ILI physician visit and Q is the ILI-related query fraction computed in previous steps. β_0 is the intercept and β_1 is the coefficient, while ϵ is the error term.

Each of the 50 million queries is tested as Q to see if the result computed from a single query could match the actual history ILI data obtained from the U.S. Centers for Disease Control and Prevention (CDC). This process produces a list of top queries which gives the most accurate predictions of CDC ILI data when using the linear model. Then the top 45 queries are chosen because, when aggregated together, these queries fit the history data the most accurately. Using the sum of top 45 ILI-related queries, the linear model is fitted to the weekly ILI data between 2003 and 2007 so that the coefficient can be gained. Finally, the trained model is used to predict flu outbreak across all regions in the United States.

This algorithm has been subsequently revised by Google, partially in response to concerns about accuracy, and attempts to replicate its results have suggested that the algorithm developers "felt an unarticulated need to cloak the actual search terms identified".[6]

Privacy concerns

Google Flu Trends tries to avoid privacy violations by only aggregating millions of anonymous search queries, without identifying individuals that performed the search.[1][7] Their search log contains the IP address of the user, which could be used to trace back to the region where the search query is originally submitted. Google runs programs on computers to access and calculate the data, so no human is involved in the process. Google also implemented the policy to anonymize IP address in their search logs after 9 months.[8]

However, Google Flu Trends has raised privacy concerns among some privacy groups. Electronic Privacy Information Center and Patient Privacy Rights sent a letter to Eric Schmidt in 2008, then the CEO of Google.[9] They conceded that the use of user-generated data could support public health effort in significant ways, but expressed their worries that "user-specific investigations could be compelled, even over Google's objection, by court order or Presidential authority".

Impact

An initial motivation for GFT was that being able to identify disease activity early and respond quickly could reduce the impact of seasonal and pandemic influenza. One report was that Google Flu Trends was able to predict regional outbreaks of flu up to 10 days before they were reported by the CDC (Centers for Disease Control and Prevention).[10]

In the 2009 flu pandemic Google Flu Trends tracked information about flu in the United States.[11] In February 2010, the CDC identified influenza cases spiking in the mid-Atlantic region of the United States. However, Google's data of search queries about flu symptoms was able to show that same spike two weeks prior to the CDC report being released.

"The earlier the warning, the earlier prevention and control measures can be put in place, and this could prevent cases of influenza," said Dr. Lyn Finelli, lead for surveillance at the influenza division of the CDC. "From 5 to 20 percent of the nation's population contract the flu each year, leading to roughly 36,000 deaths on average." [10]

Google Flu Trends is an example of collective intelligence that can be used to identify trends and calculate predictions. The data amassed by search engines is significantly insightful because the search queries represent people's unfiltered wants and needs. "This seems like a really clever way of using data that is created unintentionally by the users of Google to see patterns in the world that would otherwise be invisible," said Thomas W. Malone, a professor at the Sloan School of Management at MIT. "I think we are just scratching the surface of what's possible with collective intelligence." [10]

Accuracy

The initial Google paper stated that the Google Flu Trends predictions were 97% accurate comparing with CDC data.[4] However subsequent reports asserted that Google Flu Trends' predictions have sometimes been very inaccurate—especially over the interval 2011–2013, when it consistently overestimated relative flu incidence,[6] and over one interval in the 2012-2013 flu season predicted twice as many doctors' visits as the CDC recorded.[6][12]

One source of problems is that people making flu-related Google searches may know very little about how to diagnose flu; searches for flu or flu symptoms may well be researching disease symptoms that are similar to flu, but are not actually flu.[13] Furthermore, analysis of search terms reportedly tracked by Google, such as "fever" and "cough", as well as effects of changes in their search algorithm over time, have raised concerns about the meaning of its predictions.[6] In fall 2013, Google began attempting to compensate for increases in searches due to prominence of flu in the news, which was found to have previously skewed results.[14] However, one analysis concluded that "by combining GFT and lagged CDC data, as well as dynamically recalibrating GFT, we can substantially improve on the performance of GFT or the CDC alone."[6] A later study also demonstrates that Google search data can indeed be used to improve estimates, reducing the errors seen in a model using CDC data alone by up to 52.7 per cent.[15]

By re-assessing the original GFT model, researchers uncovered that the model was aggregating queries about different health conditions, something that could lead to an over-prediction of ILI rates; in the same work, a series of more advanced linear and nonlinear better-performing approaches to ILI modelling have been proposed.[16]

However, followup work was able to substantially improve the accuracy of GFT through the use of a random forrest regression model trained on both the incidence of influenza-like illness and the output of the original GFT model.[17]

Related systems

Similar projects such as the flu-prediction project[18] by the institute of Cognitive Science Osnabrück carry the basic idea forward, by combining social media data e.g. Twitter with CDC data, and structural models that infer the spatial and temporal spreading [19] of the disease.

References

1. "Google Flu Trends | How" (http://www.google.org/flutrends/about/how.html). Archived (https://web.archive.org/web/20121022154915/http://www.google.org/flutrends/about/how.html) from the original on 22 October 2012. Retrieved 10 November 2012.

2. Fred O'Connor (20 August 2015). "Google Flu Trends calls out sick, indefinitely" (http://www.p cworld.com/article/2974153/websites/google-flu-trends-calls-out-sick-indefinitely.html). *PCWorld*. Archived (https://web.archive.org/web/20150823220808/http://www.pcworld.com/arti cle/2974153/websites/google-flu-trends-calls-out-sick-indefinitely.html) from the original on 23 August 2015. Retrieved 23 August 2015.

3. Zeiger, Roni (6 October 2009). "Google Flu Trends Overview" (https://www.youtube.com/watc h?v=6111nS66Dpk). *youtube.com*. YouTube. Archived (https://web.archive.org/web/20130606 164710/http://www.youtube.com/watch?v=6111nS66Dpk) from the original on 6 June 2013. Retrieved 6 June 2013.

4. Ginsberg, Jeremy. "Detecting influenza epidemics using search engine query data" (http://stati c.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archiv e/papers/detecting-influenza-epidemics.pdf) (PDF). Archived (https://web.archive.org/web/201 20904220105/http://static.googleusercontent.com/external_content/untrusted_dlcp/research.g oogle.com/en/us/archive/papers/detecting-influenza-epidemics.pdf) (PDF) from the original on 4 September 2012. Retrieved 10 November 2012.

5. Ginsberg, Jeremy; Mohebbi, Matthew H.; Patel, Rajan S.; Brammer, Lynnette; Smolinski, Mark S.; Brilliant, Larry (19 February 2009). "Detecting influenza epidemics using search engine query data". *Nature*. **457** (7232): 1012–1014. Bibcode:2009Natur.457.1012G (https://ui.adsab s.harvard.edu/abs/2009Natur.457.1012G). doi:10.1038/nature07634 (https://doi.org/10.1038% 2Fnature07634). PMID 19020500 (https://pubmed.ncbi.nlm.nih.gov/19020500). S2CID 125775 (https://api.semanticscholar.org/CorpusID:125775).

6. Lazer, David; Kennedy, Ryan; King, Gary; Vespignani, Alessandro (14 March 2014). "The Parable of Google Flu: Traps in Big Data Analysis" (https://dash.harvard.edu/bitstream/handle/ 1/12016836/The%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf?sequence= 1) (PDF). *Science*. **343** (6176): 1203–1205. Bibcode:2014Sci...343.1203L (https://ui.adsabs.h arvard.edu/abs/2014Sci...343.1203L). doi:10.1126/science.1248506 (https://doi.org/10.1126% 2Fscience.1248506). PMID 24626916 (https://pubmed.ncbi.nlm.nih.gov/24626916). S2CID 206553739 (https://api.semanticscholar.org/CorpusID:206553739). Archived (https://w eb.archive.org/web/20191119022520/https://dash.harvard.edu/bitstream/handle/1/12016836/T he%20Parable%20of%20Google%20Flu%20%28WP-Final%29.pdf?sequence=1) (PDF) from the original on 19 November 2019. Retrieved 19 December 2019.

7. Helft, Miguel (13 November 2008). "Is There a Privacy Risk in Google Flu Trends?" (http://bits. blogs.nytimes.com/2008/11/13/does-google-flu-trends-raises-new-privacy-risks/). *The New York Times*. Archived (https://web.archive.org/web/20121106055230/http://bits.blogs.nytimes.c om/2008/11/13/does-google-flu-trends-raises-new-privacy-risks/) from the original on 6 November 2012. Retrieved 10 November 2012.

8. "Privacy Policy – Policies & Principles – Google" (http://www.google.com/intl/en-US/policies/pri vacy/). Archived (https://web.archive.org/web/20121103041504/http://www.google.com/intl/en- US/policies/privacy/) from the original on 3 November 2012. Retrieved 10 November 2012.

9. Peel, Deborah. "EPIC's November 12, 2008 Letter to Google Concerning Google Flu Trends" (http://epic.org/privacy/flutrends/EPIC_ltr_FluTrends_11-08.pdf) (PDF). Archived (https://web.a rchive.org/web/20121115055603/http://epic.org/privacy/flutrends/EPIC_ltr_FluTrends_11-08.p df) (PDF) from the original on 15 November 2012. Retrieved 10 November 2012.

10. "Google Uses Searches to Track Flu's Spread" (http://msl1.mit.edu/furdlog/docs/nytimes/2008 -11-11_nytimes_google_influenza.pdf) (PDF). Archived (https://web.archive.org/web/20121115 062114/http://msl1.mit.edu/furdlog/docs/nytimes/2008-11-11_nytimes_google_influenza.pdf) (PDF) from the original on 15 November 2012. Retrieved 10 November 2012.

11. Cook, S.; Conrad, C.; Fowlkes, A. L.; Mohebbi, M. H. (2011). Cowling, Benjamin J (ed.). "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3158788). *PLOS ONE*. **6** (8): e23610. Bibcode:2011PLoSO...623610C (https://ui.adsabs.harvard.edu/abs/2011 PLoSO...623610C). doi:10.1371/journal.pone.0023610 (https://doi.org/10.1371%2Fjournal.po ne.0023610). PMC 3158788 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3158788). PMID 21886802 (https://pubmed.ncbi.nlm.nih.gov/21886802).

12. Butler, Declan (13 February 2013). "When Google got flu wrong" (https://doi.org/10.1038%2F4 94155a). *Nature*. **494** (7436): 155–156. Bibcode:2013Natur.494..155B (https://ui.adsabs.harva rd.edu/abs/2013Natur.494..155B). doi:10.1038/494155a (https://doi.org/10.1038%2F494155a). PMID 23407515 (https://pubmed.ncbi.nlm.nih.gov/2340 7515).

13. "Google Flu Trends: A case of Big Data gone bad?" (http://siliconangle.com/blog/2014/03/24/g oogle-flu-trends-a-case-of-big-data-gone-bad.). *SiliconANGLE*. 24 March 2014.

14. Richard Harris (2014-03-13). "Google's Flu Tracker Suffers From Sniffles" (https://www.npr.or g/templates/transcript/transcript.php?storyId=289802934). NPR. Archived (https://web.archive. org/web/20190319224139/https://www.npr.org/templates/transcript/transcript.php?storyId=289 802934) from the original on 2019-03-19. Retrieved 2019-12-19.

15. Preis, Tobias; Moat, Helen Susannah (29 October 2014). "Adaptive nowcasting of influenza outbreaks using Google searches" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4448892). *Royal Society Open Science*. **1** (2): 140095. Bibcode:2014RSOS....140095P (https://ui.adsab s.harvard.edu/abs/2014RSOS....140095P). doi:10.1098/rsos.140095 (https://doi.org/10.109 8%2Frsos.140095). PMC 4448892 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4448892). PMID 26064532 (https://pubmed.ncbi.nlm.nih.gov/26064532).

16. Lampos, Vasileios; Miller, Andrew C.; Crossan, Steve; Stefansen, Christian (3 Aug 2015). "Advances in nowcasting influenza-like illness rates using search query logs" (https://www.ncb i.nlm.nih.gov/pmc/articles/PMC4522652). *Scientific Reports*. **5** (12760): 12760. Bibcode:2015NatSR...512760L (https://ui.adsabs.harvard.edu/abs/2015NatSR...512760L). doi:10.1038/srep12760 (https://doi.org/10.1038%2Fsrep12760). PMC 4522652 (https://www.n cbi.nlm.nih.gov/pmc/articles/PMC4522652). PMID 26234783 (https://pubmed.ncbi.nlm.nih.go v/26234783).

17. Kandula, Sasikiran; Shaman, Jeffrey (2019-08-02). "Reappraising the utility of Google Flu Trends". *PLOS Computational Biology*. doi:10.1371/journal.pcbi.1007258 (https://doi.org/10.13 71%2Fjournal.pcbi.1007258).

18. "Flu prediction project by the University Osnabrück and IBM WATSON" (http://www.flu-predicti on.com). Archived (https://web.archive.org/web/20190527185256/http://www.flu-prediction.co m/) from the original on 2019-05-27. Retrieved 2019-12-19.

19. Schumacher Johannes (2015). "A statistical framework to infer delay and direction of information flow from measurements of complex systems" (http://www.mitpressjournals.org/do i/full/10.1162/NECO_a_00756#.Vu5zr0eAY4A). *Neural Computation*. **27** (8): 1555–1608. doi:10.1162/NECO_a_00756 (https://doi.org/10.1162%2FNECO_a_00756). PMID 26079751 (https://pubmed.ncbi.nlm.nih.gov/26079751). S2CID 25156368 (https://api.semanticscholar.or g/CorpusID:25156368). Archived (https://web.archive.org/web/20190212230413/https://www. mitpressjournals.org/doi/full/10.1162/NECO_a_00756#.Vu5zr0eAY4A) from the original on 2019-02-12. Retrieved 2016-03-20.

External links

- Official website (https://www.google.org/flutrends/about/) ✏