Analytics in Practice Assignment 1
Failed Analytics Project-Real World Example
Topic: Failure of Google Flu Trends

Submitted by- Riba Khan
Email: rkhan10@kent.edu

Introduction

Analytics failure is becoming increasingly widespread as firms have access to more and more data. Large analytics projects are frequently constructed on data that has been neglected, whether owing to incompleteness, poor sanitation, or a lack of early-stage management. Because complicated initiatives require clean and accessible data to be informative and effective, this frequently results in lost time and resources. In this assignment, I'll look at one such Analytics failure called "Google Flu Trends" to see which stage of Business Analytics the project failed at, what faults it made, and what recommendations I can make.

When people talk about "big data," a proposed public health tool called Google Flu Trends is frequently mentioned. Although it become really famous when first introduced to the public, it may not be as effective as many believed. Google Flu Trends (GFT) made headlines in February 2013 (five years after its introduction), but not for the reasons that Google executives or the flu tracking system's inventors had hoped. The Centre's for Disease Control and Prevention (CDC), which bases its projections on monitoring reports from laboratories across the United States, predicted more than double the proportion of doctor visits for influenza-like illness (ILI) than GFT. Despite the fact that GFT was designed to forecast CDC reports, this happened.

Google Flu Trends	
Category: Web Service (Health)	
Started: 2008	Closed: 2015

What is Google Flu Trends?

Google Flu is a web application that was initially released by Google.org in 2008. It is a Google web service that was created to anticipate up-to-date projected influenza activity. Estimates were provided for over 25 different countries.

How does Google predict Flu Trends?

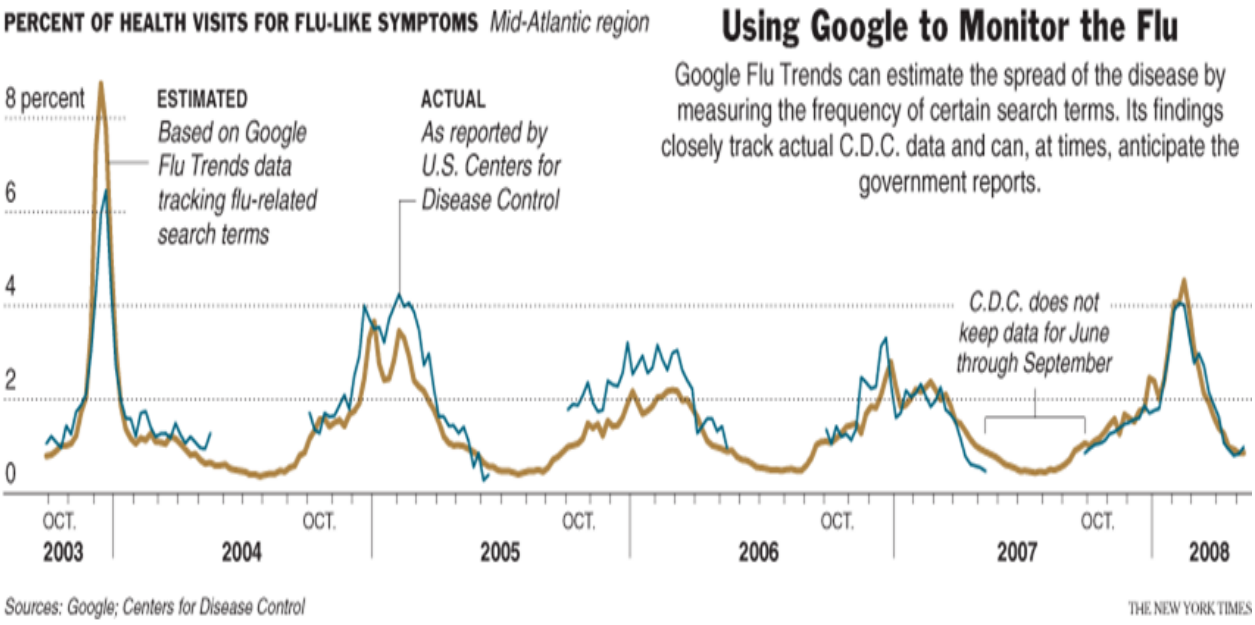
Rather than depending on disease surveillance methods utilized by the US Centers for Disease Control and Prevention (CDC), such as doctor visits and lab testing, Google strategists proposed that epidemics may be predicted using Google Searches.

> **“Google forecasted flu trends by accumulating many search queries and analyzing them to indicate the existence of flu-like sickness in millions of users' health tracking behavior.”**

Google employed a search log that included the region's IP address, which could be used to trace back to the location. Google uses computer programs to obtain and calculate data; no humans were involved in the process. When it was first introduced, Google Flu Trends was able to predict regional flu outbreaks up to ten days ahead of time.

Between 2003 and 2006, the Google team collected over 50 million potential search phrases – not just "flu" – and compared the frequency with which people looked for these terms to the number of communicated influenza-like cases. This data revealed that 45 phrases out of millions provided the best fit to the observed data. The researchers then tested their theory by comparing sickness reports from the subsequent 2007 pandemic. In terms of disease levels in real life, the forecasts appeared to be rather accurate. Because it was able to predict an increase in cases before the CDC, Flu Trends was hailed as the start of the big data era.
Google’s Flu tracker functioned admirably. Even before the outbreak was notified to the CDC, Google was able to forecast it. Due to the time, it took for reports from health care providers and medical labs to be submitted and analyzed, the CDC received information later. Google Flu Trends can be named as an example of an information-sharing tool that may be used to spot trends and make predictions. Not only this, but Google was also able to find out the geographical locations of where these searches came from.

Google was also able to identify geographically where these searches came from. . The graphic below shows that Google's ability to predict outbreaks has been on par with CDC reports from 2003 to 2008.



Sources: Google; Centers for Disease Control

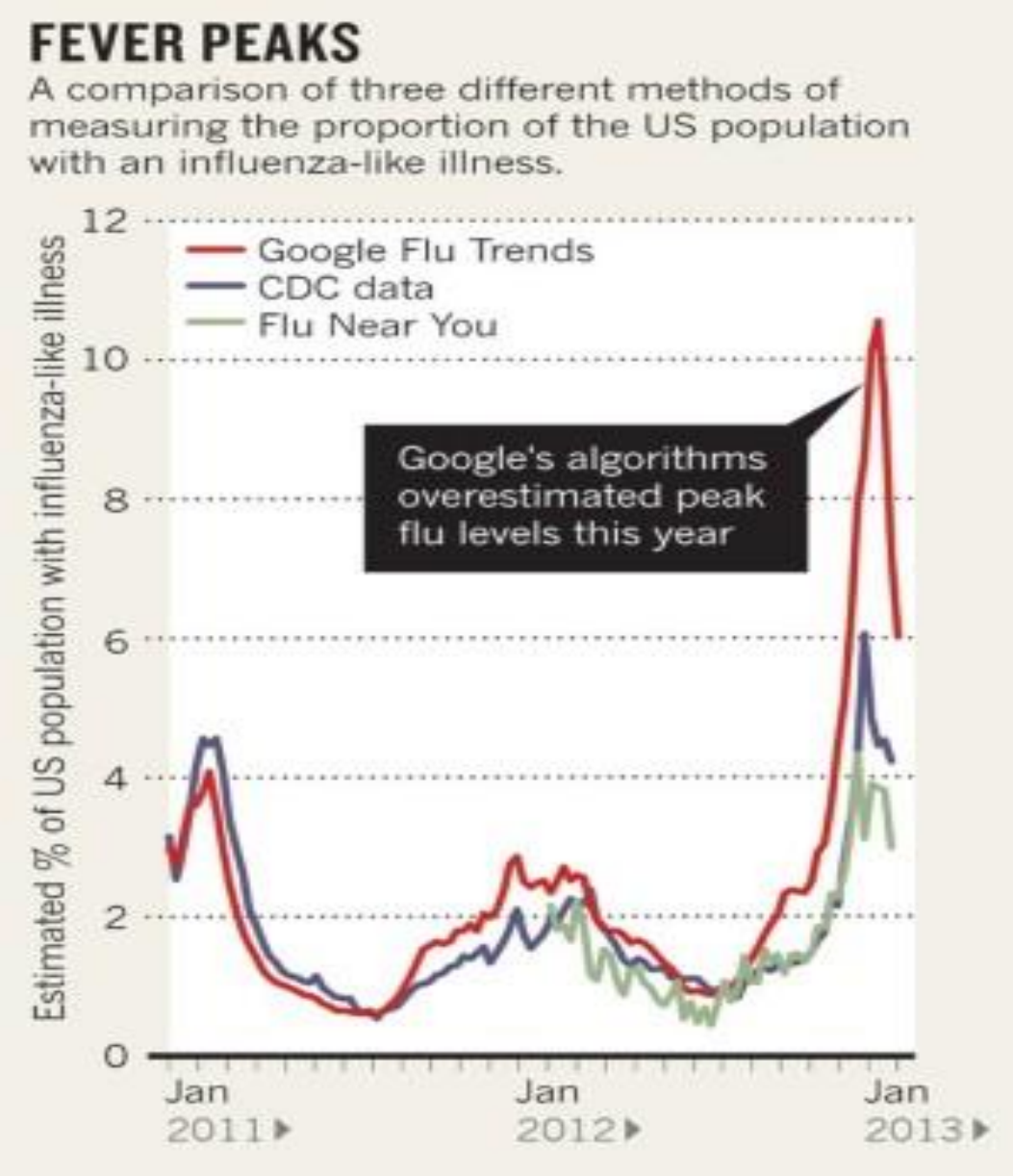
When we first look at the graph, it's tempting to believe that GFT has done a decent job and that it may be used as a quick indicator of Flu trends. But why did it come to a halt?

Now the question arises that what went wrong?
Between 2003 and 2008, flu epidemics in the United States were largely seasonal, with outbreaks occurring every winter. However, the first cases (as reported by the CDC) came around Easter time in 2009. Flu Trends had already made its predictions when the CDC data was posted, but it turned out that the Google model didn't match reality. It had greatly underestimated the scope of the original outbreak.

“In January 2013, Google's algorithms overestimated the number of flu cases by roughly 2 times what the CDC reported.”

Although the Google flu-tracking algorithm has been quite successful in predicting outbreaks, there are other factors that influence the number of Google searches conducted.

It can be seen in the graph below that in January 2013, Google's algorithms overestimated the number of flu cases by roughly 2 times what the CDC reported.

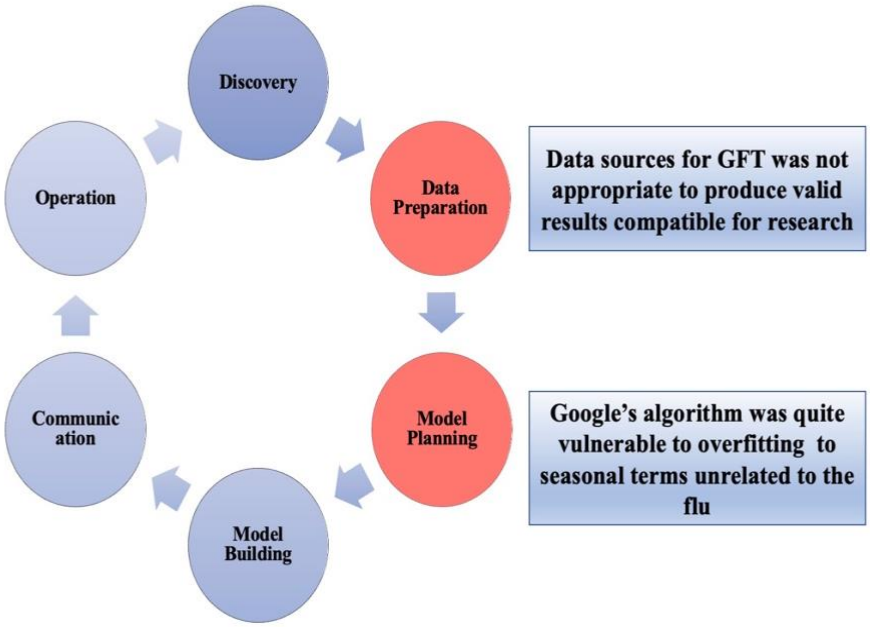


Sources: Google Flu Trends (www.google.org/flutrends); CDC; Flu Near You

Key Takeaways from Graph 1 and Graph 2

I. Google Flu Trends missed massive outbreaks such as the Swine Flu in 2009.
II. Furthermore, predictions from 2011 to 2013 were almost entirely mispredicted.
III. Predictions were not accurate enough. It was evident that CDC showed higher accuracy. Some researchers even argue that GFT was able to predict the past better than the future.

Where did GFT fail?
As we can see below the picture that there are six stages of the business analytics life cycle, and every stage is crucial for the successful implementation and running of a project.



As the above picture shows that Business Analytics Life Cycle has six stages. Google Flu Trends(GFT) failed because of the mistakes done in stages "Data Preparation" and "Model Planning".

The issue with Using Big Data

As I mentioned that when people talk about Big Data, Google Flu Trends is one tool that always comes to mind. The concept behind big data is basically using a large volume of data which can assist us to do things those smaller volumes of data cannot. Google also followed the same approach and believed that having a huge amount of data will make the predictions of flu easier. But that might not be the case in every business analytics project. This is not a proven fact that more data better the results. Sometimes at the second stage of the Business analytics life cycle, which is Termed as Data Preparation, it is crucial to clean the data first and then apply relevant transformation if required. The Google Data collection Team collected more than 50 million probable search terms – all kinds of phrases and not limited to the word "flu" and compared the frequency with which population searches for these words with the amount of reported influenza-like cases between 2003 and 2006. This information revealed that out of millions of phrases, 45 represents the best fit to the actual observations.

> **“The problem was that Flu Trends could only measure what people search for; it didn’t analyse why they were searching for those words. By removing human labour and relying solely on raw data, the model was forced to generate predictions based on only a few years' worth of search queries.”**

Many researchers even argued that not all the information that was used by Google to make predictions that might be called the “input data” in machine learning was not relevant. It was also stated again and again that Google was able to predict the past better than the future.

A bigger problem with Google Flu, though, is that most people who think they have "the flu" do not. The vast majority of doctors' office visits for flu-like symptoms turn out to be other viruses. CDC tracks these visits under "influenza-like illness" because so many turn out to be something else. To illustrate, the CDC reports that in the most recent week for which data is available, *only 8.8%* of specimens tested positive for influenza.

You can't expect internet searches to be a reliable source of information when 80-90 per cent of individuals visiting the doctor for "flu" don't have it.

Google Model didn’t match reality

In 2007, the Google team tested their model against disease reports from the epidemic that took place in the same year. The results that were produced by Google appeared to be much like real-life disease levels. This is the reason many people believed that this is the beginning of the big data age because Flu trends were able to predict the rise in cases before CDC. Flu epidemics in the United States were significantly seasonal between 2003 and 2008, with outbreaks occurring every winter. In 2009, however, the first instances (as reported by the CDC) began around Easter. When the CDC data was released, Flu Trends had already made its forecasts, but it turned out that the Google algorithm didn't match reality. It had grossly misjudged the original outbreak's extent.
Six months after the epidemic began, Google modified their model to reflect the 2009 CDC data, now that they had the information at hand. Despite these improvements, the upgraded version of Flu Trends had problems last winter when it overstated the size of the influenza pandemic in New York State. The 2009 and 2012 instances raised the question of how effective Flu Trends is at predicting future epidemics as opposed to merely identifying patterns in historical data.

Key takeaways from Google Flu Trends Failure:

1) There was a huge collection of misinformation. Google never disclosed what 45 search terms it used to predict the Flu trends.
2) The problem appears to be with both the data it worked with and the way it was analyzed.
3) Google runs programs on computers to access and calculate the data so “no human” was involved in the process.
4) Flu trends did not constantly update in its algorithm what keywords people used and how they used them in their search. Predictive models can’t be just left alone there are changes around them over a year that needs to be updated which Google overlooked.
5) GFT's first version had a particularly difficult mix of massive and little data. Quantity of data does not imply that core challenges like measurements, construct validity and reliability, and data dependencies can be overlooked. The main problem is that most big data that has gotten public attention isn't the result of equipment that are meant to create reliable and valid data for scientific study.

From the above paragraphs, it is quite evident that Google Flu Trends failed in collecting the right amount of data for its prediction and developing an algorithm that can keep up with the trend. Since we have talked a lot about what went wrong, let us now talk about what mistakes should be avoided in the future that would make a model like GFT successful.

Suggestions

- Future tools like GFT should, for example, keep updating the data's fit to flu prevalence; otherwise, the data stream's value will gradually deteriorate.
- Despite the criticism GFT faced, many researchers believed that monitoring internet search queries could still be useful if they are combined with other surveillance and better prediction tools.
- Additional scholars have indicated that other digital data sources, such as Twitter feeds and mobile phone GPS could be beneficial in epidemic research. Such tools could be used to study human movement and the distribution of public health information, in addition to helping academics analyze outbreaks (or misinformation).
- Although web-based technologies have received a lot of attention, another sort of big data is already having a significant impact on illness research. Researchers are using genome sequencing to put together how diseases spread and where they originate. Even the existence of a novel disease variation can be discovered using sequence data.

It's worth noting that GFT taught us that data size isn't necessarily the most important factor. Big data opens a world of possibilities for new insights, a comprehensive understanding of the human system, and finding relationships and non-linearities between

variables. Traditional data, on the other hand, frequently provides information not found in big data, and the same factors that have enabled big data are also enabling more traditional data collection. Standard surveys, trials, and health reporting have all improved because of the internet. Instead of continuously focusing on the "big data revolution," perhaps it's time to pay more attention to the "all data revolution."

Conclusion: **Google no longer publishes flu statistics, but it has demonstrated that big data research may be useful in transforming how we respond to health crises. Future updates should, for example, keep updating the data's fit to flu prevalence; otherwise, the data stream's value will fast deteriorate. Big data can be beneficial, but not when the data is inaccurate.**

Bibliography

Salzbeg, S. (2014, March). *Why Google Flu Is A Failure*. Retrieved from Forbes : https://www.forbes.com/sites/stevensalzberg/2014/03/23/why-google-flu-is-a-failure/?sh=32b775b35535

Kucharski, A. (2014, October). *Google'S flu fail shows the problem with big data* . Retrieved from The CONVERSATION: https://theconversation.com/googles-flu-fail-shows-the-problem-with-big-data-19363

Lazer, D. R. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science* , 1203-1205.

Hodson, H. (2014, March 13). Google Flu Trends get it wrong three years running.

Rana, S. (2019, November Monday). "Big Data Hubris"– belief that huge volume of data always leads to better results . *TechNews*.

Kulgler, L. (2016). What happens when big data blunders? *Communication of the ACM*, 15-16.