DRAFT — a final version will be posted shortly

CSE 8803 EPI: Data Science for Epidemiology, Fall 2022

Lecturer: B. Aditya Prakash — October 13, 2022
Scribe: Aishwarya Vijaykumar Sheelvant — Lecture 14 : Surveillance-II

1 Surveillence

Surveillance can be performed on populations in different stages of illness progression, as shown in Figure 1 [12].

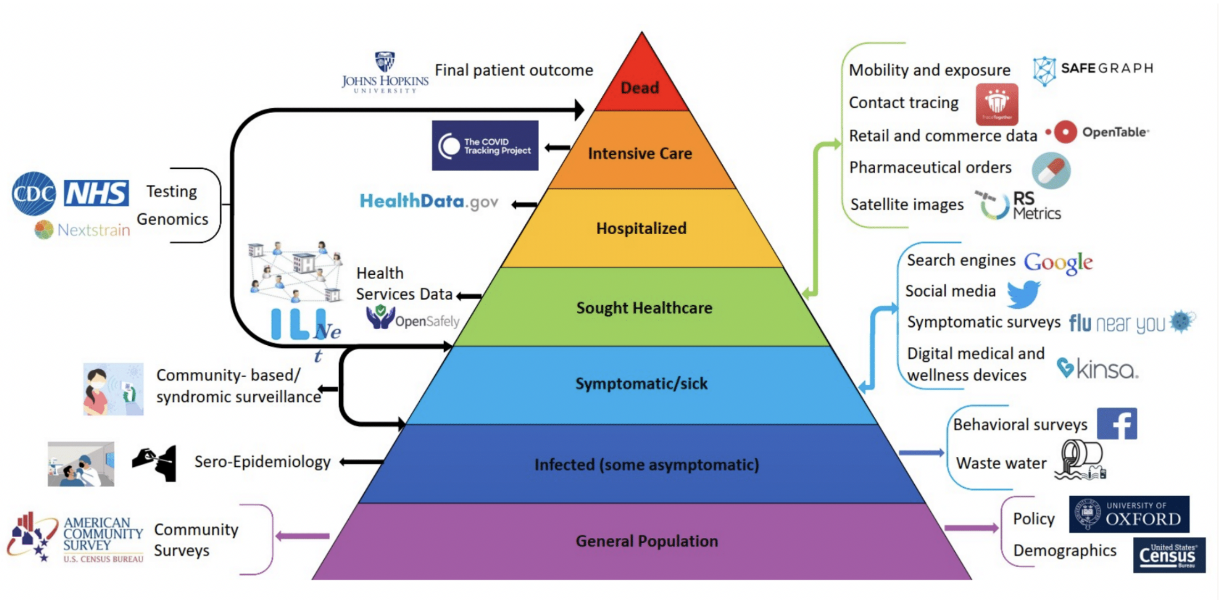


Figure 1: **Surveillance data source hierarchy:** The pyramid represents stages of illness progression, with size proportional to the population count. Each level is linked to predictive modeling datasets: left side shows direct disease tracking, while the right highlights indirect epidemiological indicators [12]

In the realm of data handling, diverse methodologies are essential. For the surveillance of platforms such as search engines, social media, symptomatic surveys, and digital medical and wellness devices, we employ Machine Learning (ML) techniques. These surrogate data sources present a spectrum of attributes: advantageous, detrimental, and some with potential pitfalls. It's crucial to understand that shifts in distribution profoundly influence ML approaches. Such distributional variations, especially in public health contexts, can substantially change search query outcomes. It's noteworthy that many of these data sources are very sensitive to distributional changes.

Proposals for influenza surveillance:

- **Search Queries**: Example: "Miley Cyrus cancels Charlotte concert due to flu." It's important to note that search queries are influenced by public health-related distributional shifts.

- **Over-The-Counter (OTC) Medication Sales**: Sales metrics can be skewed by various factors, including discount sales, hoarding, lack of patient-specific data.
- **Wikipedia**: There is a lack of specificity about visitor locations.
- **Digital social media data (like Tweets)**: There is a lot of evidenced challenges of utilizing digital data. A principal concern is that attention-grabbing data may not necessarily stem from reliable and validated instruments.

No solution has yet been identified. Without human interaction, there is no dataset that can be used that will produce the desired outcomes. Caution has be taken while implementing such approaches.

2 Nowcasting

Nowcasting, as introduced by Choi and Varian in 2012 [4], aims to predict the present, a concept contrasting with traditional forecasting. Government agencies periodically release indicators; however, these releases often come with a reporting lag of several weeks and are typically revised a few months later. Despite the digital age, accessing real-time data, such as the current number of active COVID cases, remains challenging. Thus, both health-related data and even social media data have inherent reporting lags. Recognizing these challenges, there has been an increased reliance on private sector companies that provide real-time economic activity data to fill these gaps.

Incorporating data science into epidemiology, nowcasting employs both statistical and machine learning models to make these real-time predictions. The intuition behind nowcasting is to choose the optimal function from a family of functions that best approximates the forecast target based on input data. Specifically, the goal is to minimize $\min_{f \in \mathcal{H}} \sum_{i=1}^{T} \mathcal{L}(f(x_i) - y_i)$, where $\mathcal{L}$ is the loss function, f is the chosen function from family $\mathcal{H}$, and T is the total number of data points. This approach closely mirrors forecasting. The main distinction is that while forecasting predicts future values using current data, nowcasting is rooted in the present, leveraging surrogate data sources to approximate the now. Despite similarities in their machine learning models, certain techniques, like regression, are more prevalent in nowcasting. Moreover, nowcasting places a stronger emphasis on diverse data sources and indicators compared to traditional forecasting.

3 Google Flu Trends

Introduced in 2009 by Ginsberg et. al. [6], Google Flu Trends (GFT) was a pioneering system that leveraged health-seeking behavior monitoring through Google queries. This nowcasting system started with 50 million candidate queries, which were meticulously refined to a set of 45 that most accurately mirrored the CDC ILI data in the US. Not merely a statistical venture, the development of GFT was a blend of both automated and manual efforts, as queries that correlated with the flu season were hand pruned. Relative query volumes (with respect to weekly search volume per location) were used as independent variables. This process of refining began with those 45 queries, further categorizing them into distinct classes. This idea was previously explored with Yahoo queries by Polgreen et. al. [11].

The model in GFT was a simple linear model for nowcasting ILI. It utilized the search logits of query fractions as features, defined by the equation:

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

where P represents ILI (physician visits) and Q denotes the fraction of search queries that are ILI-related. While GFT made use of an automated, correlation-based selection system, it strictly did not reveak specific queries as a key design element, to avoid the introduction of bias to the public but also due to the sensitive nature of some queries. Despite this, Google provided a transparent dashboard, making the trends accessible to all, thus ensuring the model's efficacy.

GFT vs. Traditional Surveillance

Google Flu Trends (GFT) was initially created as a system that aimed to detect flu trends faster than traditional surveillance methods. The method utilized the analysis of search query data to provide near real-time estimates on flu activity. In early stages, GFT showed promising results, as evidenced by Ortiz et. al. [10]. It was compared to the US Influenza Virologic Surveillance data and CDC ILI surveillance data. The correlation with CDC ILI data, in particular, was as high as 0.94 up to the 2009 H1N1 pandemic.

Shortcomings during the H1N1 pandemic and the H3N2 epidemic

However, the system was not without its flaws. During the 2009 H1N1 pandemic, GFT failed to capture changing trends in keyword correlates and did not handle data drift effectively, as evidenced by Olsen et. al. [9]. It completely missed the first wave of the H1N1 pandemic flu. Additionally, when GFT was evaluated at different geographic scales - national (US), regional (mid-Atlantic), and local (New York city) - it showed misleading correlations. GFT displayed a particular limitation when trying to extrapolate data from these densely populated areas to other regions, suggesting that GFT's prediction model might not be as robust across various geographic scales, as exemplified in Figure 2 for Latin American countries.

Attempts were made to fix the shortcomings of GFT post the H1N1 pandemic. However, during the H3N2 epidemic in 2012 and 2013, GFT faced another significant challenge. Despite the improvements, the system overestimated the intensity of the H3N2 epidemic. This error demonstrated that while the prior issues might have been addressed, new ones emerged, highlighting the challenges of relying solely on digital surveillance methods like GFT.

One of the major challenges faced by GFT was its susceptibility to external influences such as news articles. A surge in flu-related news could lead to an increase in related search queries, leading to false alarms since the actual flu activity wasn't spiking. Moreover, Google's search algorithm was not static; it evolved over time. This dynamic nature of the search algorithm meant that the health-seeking behavior of the population was also constantly changing. Such shifts further complicated GFT's predictive accuracy. Furthermore, there was an issue of transparency with the search terms GFT utilized. The exact terms and their weightage within the algorithm were not made public. This lack of transparency meant that when things went wrong, researchers and public health experts couldn not precisely identify the problematic search queries or terms. This showcased the need for blending digital surveillance data with traditional epidemiological methods to get a more accurate picture.

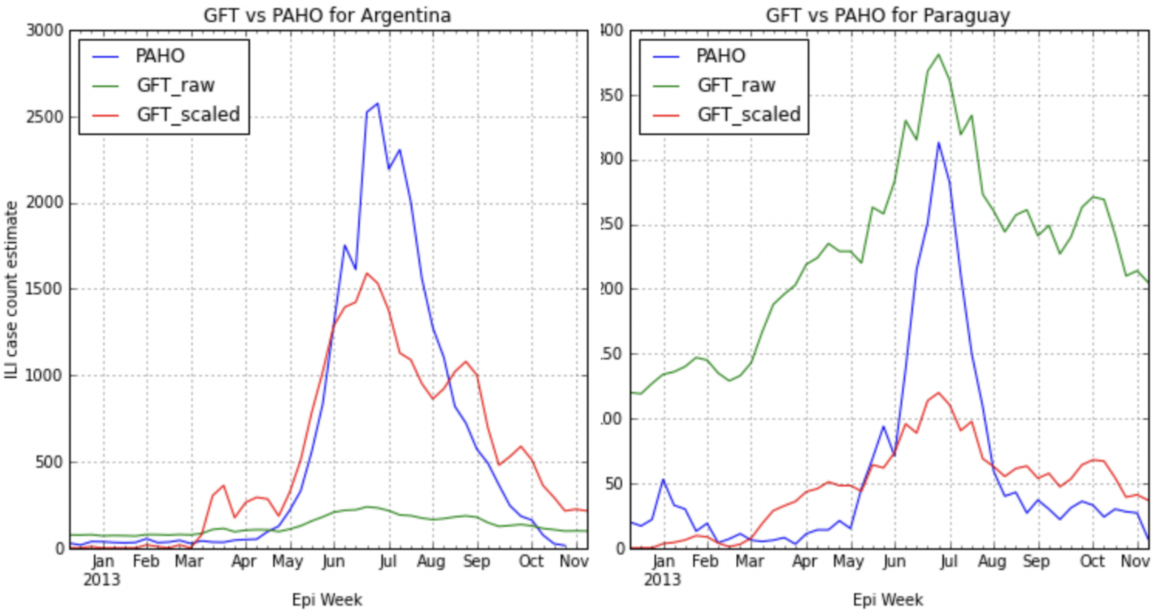


Figure 2: **Google Flu Trends for Latin American Countries**: This graph depicts the trends of flu in Latin American countries. The comparison between raw GFT data and scaled GFT data against the PAHO (Pan American Health Organization) data indicates the relatively poor accuracy and adaptability of the GFT model in different regions.

The final straw: GFT's Demise and Subsequent Efforts

For the two years ending Sep 2013, GFT's estimates were high in 100 out of 108 weeks. After the October 2013 update, discrepancies in GFT's predictions reached concerning levels, with estimates deviating by as much as 30% during the 2013-2014 flu season [8]. Such significant inaccuracies led to waning trust in the tool, and it was soon realized that GFT was no longer a reliable source for flu trend predictions. Consequently, Google decided to shut down GFT. In a subsequent effort to refine the model and address its shortcomings, Google reached out to independent research groups, inviting them to collaborate and work on improving the system.

Future improvements to GFT

The main improvements made to Google Flu Trends (GFT) were aimed at refining its accuracy and adaptability, including:

1. **Inorganic Query Filtering**: GFT was updated to ignore inorganic queries that arose from heightened media coverage. Such queries, often unrelated to actual flu incidence but more about public curiosity or panic, could skew the results. Techniques like long-term

and short-term spike detectors were introduced to identify these aberrations.

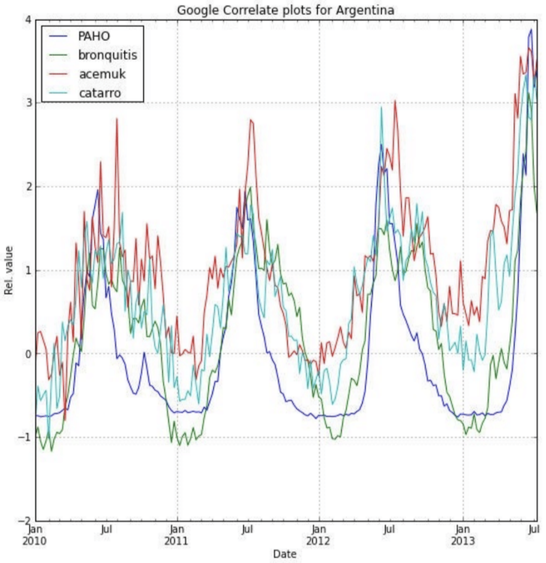
2. **Model Drift Handling**: Another challenge was the model drift over time. To combat this, GFT was regularly updated. The system was retrained every flu season, and regularized parameters or regularizers were introduced to prevent overfitting and stabilize the predictions.

Although the dictionary used for GFT was effective for English-speaking countries, it lacked resources for other languages. This posed a challenge, particularly when trying to predict flu trends in non-English speaking regions. To address this, a new dictionary was designed, which was backed by:

- **Pseudo Query Expansion Methods**: These methods help in defining custom queries. It began with gathering queries from flu-related datasets sourced from health ministry websites, tweets, and news articles.

- **Google Correlate**: This tool was used to align the volume of keyword search queries with the PAHO (Pan American Health Organization) time series data. It offered insights into how well the search terms corresponded with actual flu trends, as shown below in Figure 3.

After these refinements, GFT showed better compatibility with multiple languages and was better aligned with query expansion techniques.



Symptomatic words: “bronquitis”, “catarro”, “tos seca” (whooping cough)

Medicinal words: “acemuk”, “claritromicina” (clarithromycin)

Interesting words: ginger (“jengibre”), leave letter (“letra de deja”)

Figure 3: **Automatically Discovered Words in GFT with new dictionary**: This chart showcases the correlation between certain symptomatic, medicinal, and other interesting search queries in relation to flu trends in Argentina. The inclusion of diverse terms like ”bronquitis” and ”claritromicina” highlights the expanded lexicon of the dictionary.

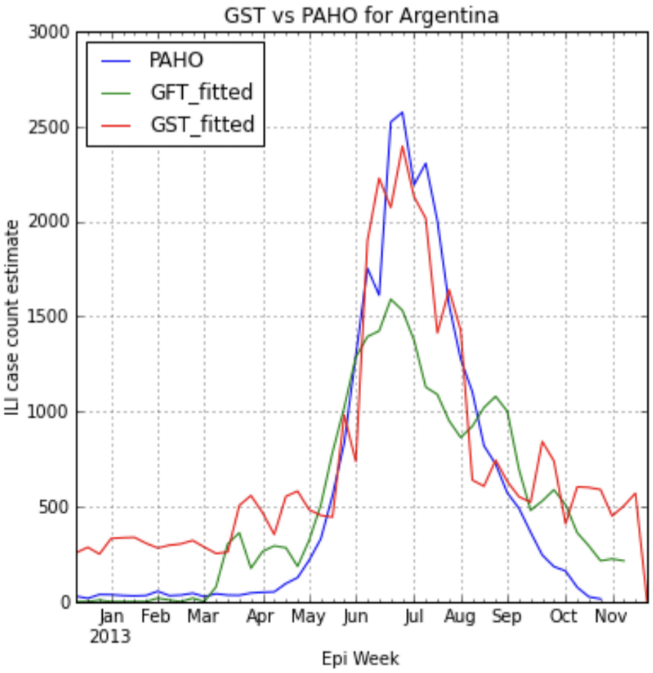


Figure 4: **Performance Comparison of Refined GFT Model**: This graph provides a comparison with Figure 2 of the GFT's performance post dictionary enhancement. The alignment between PAHO data and the fitted GFT model indicates a significant improvement in prediction accuracy for Argentina's flu trends.

4 Food borne illness detection

There is a novel approach, implemented by Sadelik et. al. [13] that uses ”machine-learned epidemiology” for real-time detection of foodborne illnesses at scale. The method applies ML models to analyze Google search queries and location logs to determine which restaurants might have significant food safety violations like poor sanitation, leading to foodborne illness outbreaks. The data used for this analysis is aggregated and anonymous, sourced from users who have opted to share their location data.

Approach for Detection:

The system identifies search queries that suggest a foodborne illness (e.g., symptoms like food poisoning). It then cross-references these queries with the history of previous location data to pinpoint restaurants the users in aggregate might have visited before showing signs of illness through their search queries. The system calculates, for each restaurant, the proportion of users who visited and later showed symptoms, which might suggest a potential outbreak. However, a user's location is deduced only if they performed a search or posted a message from that location.

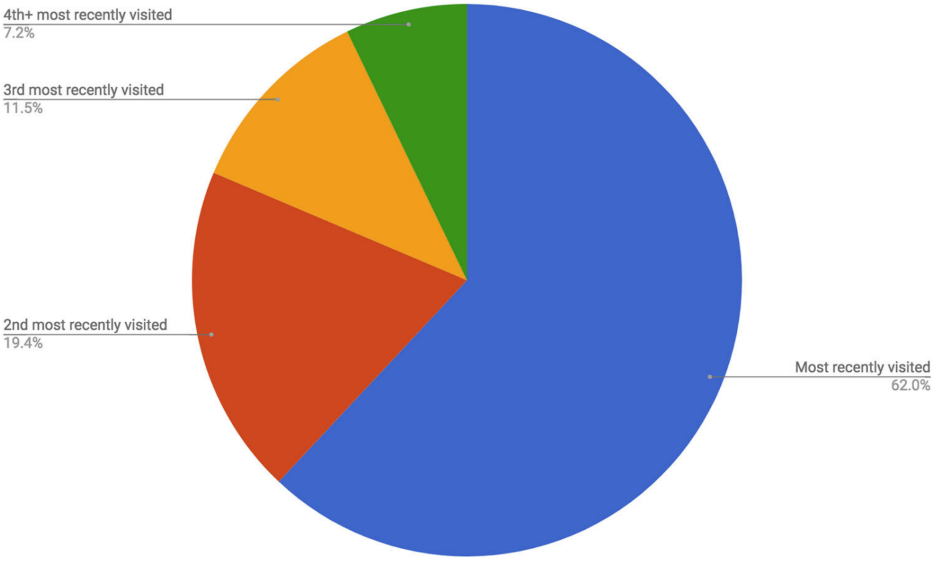
Challenges in the Data:

One of the primary challenges in utilizing this method is the inherent noise and ambiguity present in individual search queries. Users frequently visited several restaurants, making it difficult to ascertain which specific establishment might be the source of a potential illness. Moreover, searches for symptoms akin to food poisoning could arise from various causes. To address these complexities, an algorithm was developed to narrow down the potential sources. This algorithm, resembling a voting or message passing system, calculated the proportion of users who visited a particular restaurant and subsequently exhibited signs of illness.

A significant aspect of this study was its reliance on passive data, such as location information from users. The collection of passive data imposes a heightened sense of responsibility due to privacy concerns. To ensure user privacy, a supervised, privacy-preserving machine-learned classifier was implemented. This classifier evaluated multiple signals beyond just the search query, including:

- The search results displayed in response to the query.
- Aggregated data on which search results users clicked.
- The content of web pages accessed from those search results.

This study was implemented in Las Vegas and Chicago. This method has a more targeted approach to testing the restaurants which caused a significant reduction in the amount of work needed. The results of this study are shown in the pie chart below in Figure 5. As observed, a significant 62.0% of the illnesses can be attributed to the most recently visited restaurant, indicating a strong correlation between the most recent eatery visited and the onset of symptoms. The subsequent decrease in percentages for the 2nd, 3rd, and 4th or later visits suggests that the likelihood of a restaurant being the source of foodborne illness diminishes as we move further back in a user's visit history.



Frequency with which illness can be attributed to recently visited restaurants, among FINDER restaurants. $N = 132$

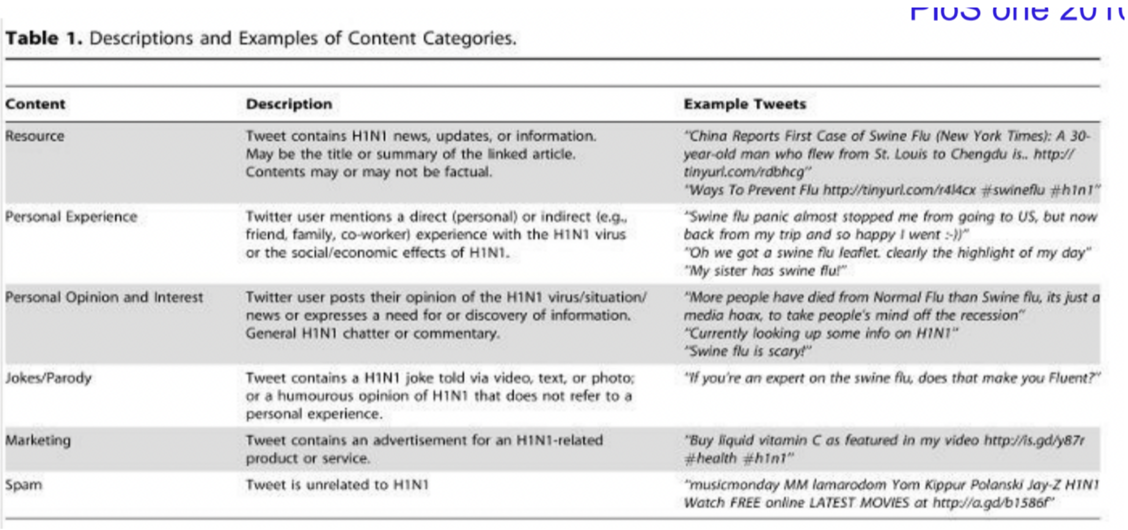
Figure 5: **Results of food-borne illness detection and attribution:** This pie chart showcases the distribution of the frequency with which illnesses were attributed to the most recent restaurant visits.

5 Nowcasting with twitter

The advent of digital data collection has paved the way for numerous approaches to data analysis, with social media platforms becoming particularly significant. Building on the success of prior methodologies, researchers have increasingly leveraged data from social media platforms, notably Twitter, for nowcasting purposes. One significant investigation by [5] delved into techniques reminiscent of Google Flu Trends (GFT), aiming to predict Influenza-Like Illness (ILI) case counts using Twitter data. Their methodology integrated geolocation to concentrate on specific regions and employed document filtering to discern ILI-relevant tweets. Subsequent regression analyses revealed that employing multiple keyword independent variables resulted in enhanced performance compared to the simple linear regression as was adopted in GFT. Specifically, a Lasso-based linear regression model incorporating n-grams as features proved effective in case number predictions.

Furthermore, the research presented by [14] underscored the utility of Twitter data during the H1N1 pandemic. Tweets with geolocation, tagged based on US domestic locations and encompassing flu-related keywords, were collected. This raw data underwent refinement by eliminating stopwords and applying stemming processes. Utilizing the derived keywords, a new lexicon was established. Support Vector Regression was subsequently employed to correlate these dictionaries with CDC ILI rates. This model was trained on data from nine out of the ten CDC US regions and assessed on the tenth.

In the wake of challenges associated with Google Flu Trends, particularly concerning data gripes and shifts in distribution, subsequent investigations like the one by [3] pivoted towards a more content-oriented analytical paradigm. This led to the application of coding rules to classify tweets. Some illustrative categories of such content categorizations are shown as examples in Figure 6 below. The hypothesis was that only a subset of tweets were actually useful whereas a smaller subset of tweets were spam. It was found that 52.6% of the tweets were about news and information and 4.5% were misinformation.



PloS one 201

Table 1. Descriptions and Examples of Content Categories.

Content	Description	Example Tweets
Resource	Tweet contains H1N1 news, updates, or information. May be the title or summary of the linked article. Contents may or may not be factual.	*"China Reports First Case of Swine Flu (New York Times): A 30-year-old man who flew from St. Louis to Chengdu is.. http://tinyurl.com/rdbhcg"* *"Ways To Prevent Flu http://tinyurl.com/r4l4cx #swineflu #h1n1"*
Personal Experience	Twitter user mentions a direct (personal) or indirect (e.g., friend, family, co-worker) experience with the H1N1 virus or the social/economic effects of H1N1.	*"Swine flu panic almost stopped me from going to US, but now back from my trip and so happy I went :-))"* *"Oh we got a swine flu leaflet. clearly the highlight of my day"* *"My sister has swine flu!"*
Personal Opinion and Interest	Twitter user posts their opinion of the H1N1 virus/situation/news or expresses a need for or discovery of information. General H1N1 chatter or commentary.	*"More people have died from Normal Flu than Swine flu, its just a media hoax, to take people's mind off the recession"* *"Currently looking up some info on H1N1"* *"Swine flu is scary!"*
Jokes/Parody	Tweet contains a H1N1 joke told via video, text, or photo; or a humourous opinion of H1N1 that does not refer to a personal experience.	*"If you're an expert on the swine flu, does that make you Fluent?"*
Marketing	Tweet contains an advertisement for an H1N1-related product or service.	*"Buy liquid vitamin C as featured in my video http://is.gd/y87r #health #h1n1"*
Spam	Tweet is unrelated to H1N1	*"musicmonday MM lamarodom Yom Kippur Polanski Jay-Z H1N1 Watch FREE online LATEST MOVIES at http://a.gd/b1586f"*

Figure 6: **Content Categorization of Tweets** [3]: Classification of tweets with respective descriptions and illustrative examples for H1N1 content categorization

Therefore, some recent studies such as [1] have delved into the intricate process of multi-level tweet classification. As depicted in the image on the left side of Figure 7 shown below, tweets undergo a sequential filtration process involving multiple filters to ascertain their relevance or weed out extraneous content. Subsequent to this filtering, a regression was executed based on the infection detection algorithm. The results for this during the 2012-2013 influenza season are presented in the graph on the right side of Figure 7. Observing the graph, it becomes evident that the infection detection algorithm exhibits a trend that closely mirrors the CDC data, underscoring its potential utility in tracking influenza outbreaks.

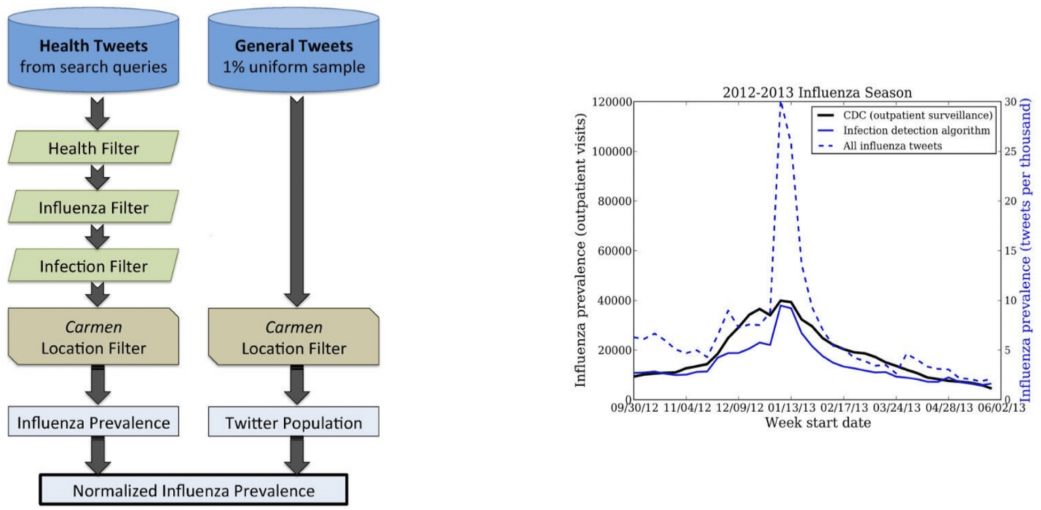
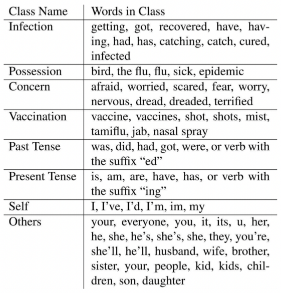


Figure 7: **Multi-level tweet classification** [1]: Classification of tweets after multi-level filtration (left) and comparison of infection detection algorithm with CDC data for the 2012-2013 influenza season (right)

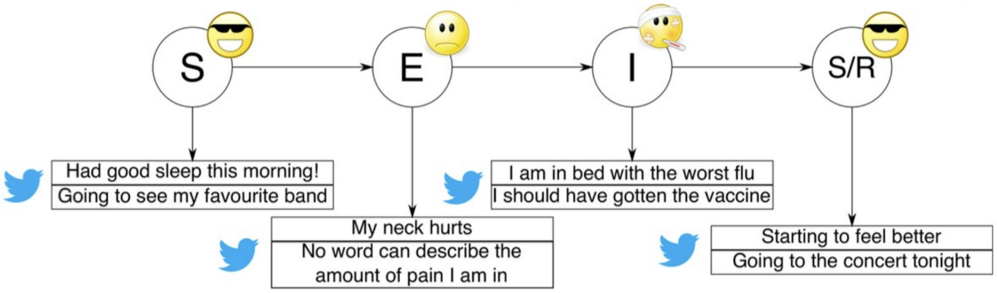
Many other approaches have followed this method of filtering out unrelated data. For example a paper by Lamb in 2013 [7], tries to develop further distinctions by trying to distinguish between infection vs concerned awareness by building meaningful classifiers and building parts of speech templates from world class features.



Class Name	Words in Class
Infection	getting, got, recovered, have, having, had, has, catching, catch, cured, infected
Possession	bird, the flu, flu, sick, epidemic
Concern	afraid, worried, scared, fear, worry, nervous, dread, dreaded, terrified
Vaccination	vaccine, vaccines, shot, shots, mist, tamiflu, jab, nasal spray
Past Tense	was, did, had, got, were, or verb with the suffix “ed”
Present Tense	is, am, are, have, has, or verb with the suffix “ing”
Self	I, I’ve, I’d, I’m, im, my
Others	your, everyone, you, it, its, u, her, he, she, he’s, she’s, she, they, you’re, she’ll, he’ll, husband, wife, brother, sister, your, people, kid, kids, children, son, daughter

Figure 8: **Parts of speech templates** [7]: Word classifications used to break down tweets for processing

The paper ”Flu Gone Viral” by Chen et. al. [2], proposes a temporal topic model for inferring the biological state of the user and an EM algorithm for modelling the hidden epidemiological state of the user (S, E, I, R). The Hidden Flu State from Tweets (HSFTM) model generates the state form the tweet and then the topic from the word. Then EM algorithm is used to infer the topic distributions and state transition probabilities. This method may suffer from large noisy vocabulary and can be improvised by introducing an already curated list of keywords from an expert.



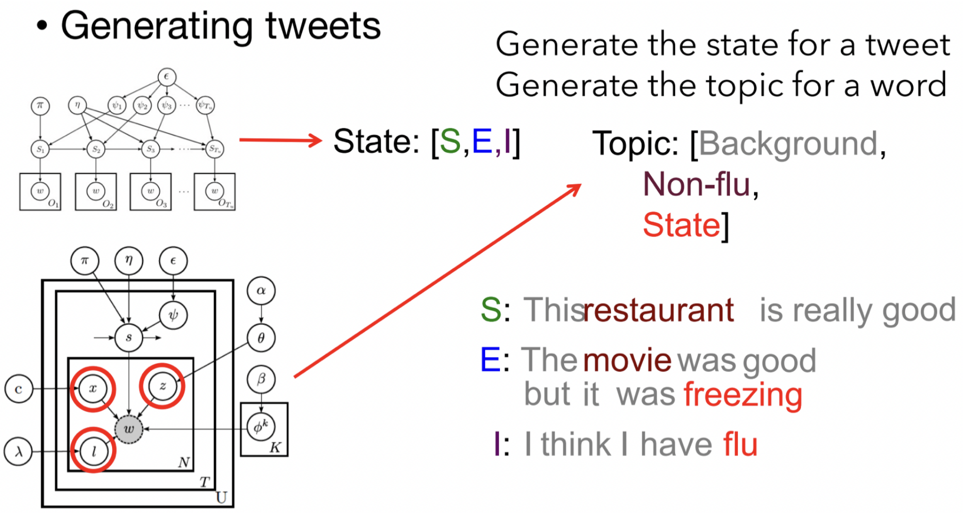
Had good sleep this morning!
Going to see my favourite band

My neck hurts
No word can describe the amount of pain I am in

I am in bed with the worst flu
I should have gotten the vaccine

Starting to feel better
Going to the concert tonight

Figure 9: **SEIR model** [2]: A model depicting how a typical SEIR/SEIS model (with states: Susceptible, Exposed, Infected, Recovered) would overlap with a model that uses tweet classification to detect viral spread in a population, including example tweets of what one could expect at each stage



- Generating tweets

Generate the state for a tweet
Generate the topic for a word

State: [S,E,I]

Topic: [Background,
Non-flu,
State]

S: Thisrestaurant is really good

E: The movie was good but it was freezing

I: I think I have flu

Figure 10: **HFSTM and State Transition**: A model showing how the HFSTM and State Transition models can be used to probabilistically create tweets in different states using words that have been pre-categorized

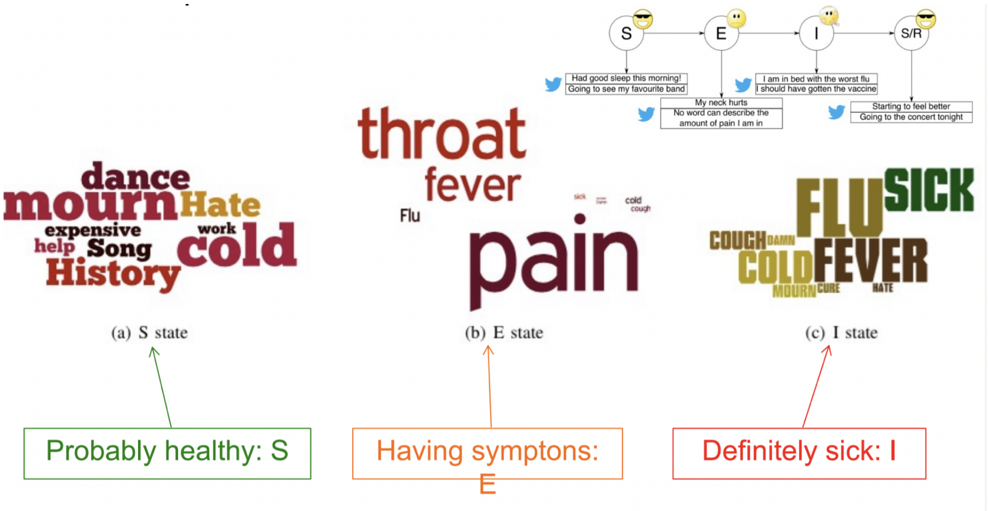


Figure 11: **Learned word distributions**: The most probable words learned in each state

Data obtained from PAHO was considered as ground truth. The number of keywords were counted as as features and the ground truth curve was regressed. Google flu trends data was used to regress the PAHO curve. Using the HSFTM, they distinguished the states of the keywords, and only the keywords in I state were identified the again used to regress to PAHO. The model can learn transition probabilities (S, E, I, R). This model is observed to learn transitions well.

6 Tracking COVID-19 with Online Search

This study uses time series of online search query frequencies to gain insights about the prevalence of COVID 19 in multiple countries. They first built unsupervised modeling techniques based on associated symptom categories identified by United Kingdom's National Health Service and Public Health England and then they created an online search time series. One of the challenges was to clean the data. They tried to minimize an expected bias in these signals caused by public interest. Symptom categories were weighted based on their reported ratio occurrence in cases of COVID-19. They reduced the effect of news via autoregression. Their study confirms the unsupervised approach's insights and demonstrates how early warnings may have been gathered from areas that had already felt the effects of COVID-19. Then they conducted a correlation and regression analysis to uncover potentially useful online search queries that refer to underlying behavioural or symptomatic patterns in relation to confirmed COVID-19 cases.The output of this model provides useful insights including early warnings for potential disease spread, and showcases the effect of physical distancing measures.

To reduce the effect of news via autoregression, for the weighted score of symptom-related online searches $g = g_p + g_c$, where g_p is infected users and g_c is concerned users, then there exists a constant $\gamma \in [0, 1]$ such that $g_p = \gamma g$ and $g_c = (1 - \gamma)g$. On any given day the proportion of news articles about the COVID-19 pandemic is $m \in [0, 1]$, then $AR(g, m)$ is an autoregressive function on g and m defined as:
$arg\ min_{w,v,b_2} \frac{1}{N} \sum_{t=1}^{N} (g_t - w + 1g_{t-1} - w_2 g_{t-2} - v_1 m_t - v_2 m_{t-1} - v_3 m_{t-2} - b_2)^2$

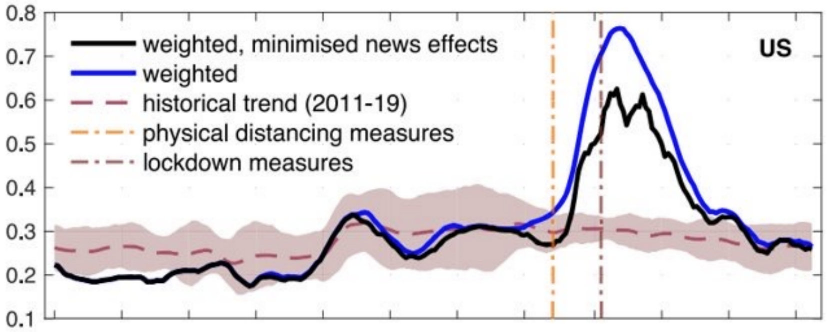


Figure 12: **Early warning signal** - Online searches precede reported confirmed cases by 16.7 and deaths by 22.1 days: A model showing the confirmed COVID-19 cases, the symptom-related search frequency weighted for news effect, and the confirmed COVID-19 cases shifted to where they best line up with the weighted searches.

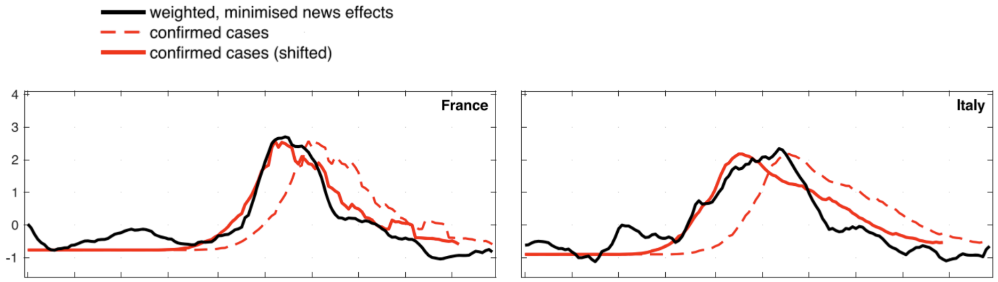


Figure 13: **Results of the model**: A graph showing the historical trend of symptom-related search, that years' trend of searches both weighted and weighted with minimised news effects, and when that year physical distancing measures and lockdown measures were put in to place

References

[1] D. A. Broniatowski, M. J. Paul, and M. Dredze. National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, 8(12):e83672, 2013.

[2] L. Chen, K. S. M. Tozammel Hossain, P. Butler, N. Ramakrishnan, and B. A. Prakash. Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models. In *2014 IEEE International Conference on Data Mining*, pages 755–760, 2014.

[3] C. Chew and G. Eysenbach. Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118, 2010.

[4] H. Choi and H. Varian. Predicting the present with google trends. *Economic record*, 88:2–9, 2012.

[5] A. Culotta. Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, page 115–122, New York, NY, USA, 2010. Association for Computing Machinery.

[6] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[7] A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[8] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of google flu: traps in big data analysis. *science*, 343(6176):1203–1205, 2014.

[9] D. R. Olson, K. J. Konty, M. Paladini, C. Viboud, and L. Simonsen. Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS computational biology*, 9(10):e1003256, 2013.

[10] J. R. Ortiz, H. Zhou, D. K. Shay, K. M. Neuzil, A. L. Fowlkes, and C. H. Goss. Monitoring influenza activity in the united states: a comparison of traditional surveillance systems with google flu trends. *PloS one*, 6(4):e18687, 2011.

[11] P. M. Polgreen, Y. Chen, D. M. Pennock, F. D. Nelson, and R. A. Weinstein. Using internet searches for influenza surveillance. *Clinical infectious diseases*, 47(11):1443–1448, 2008.

[12] A. Rodríguez, H. Kamarthi, P. Agarwal, J. Ho, M. Patel, S. Sapre, and B. A. Prakash. Data-centric epidemic forecasting: A survey, 2022.

[13] A. Sadilek, S. Caty, L. DiPrete, R. Mansour, T. Schenk Jr, M. Bergtholdt, A. Jha, P. Ramaswami, and E. Gabrilovich. Machine-learned epidemiology: real-time detection of foodborne illness at scale. *NPJ digital medicine*, 1(1):36, 2018.

[14] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.