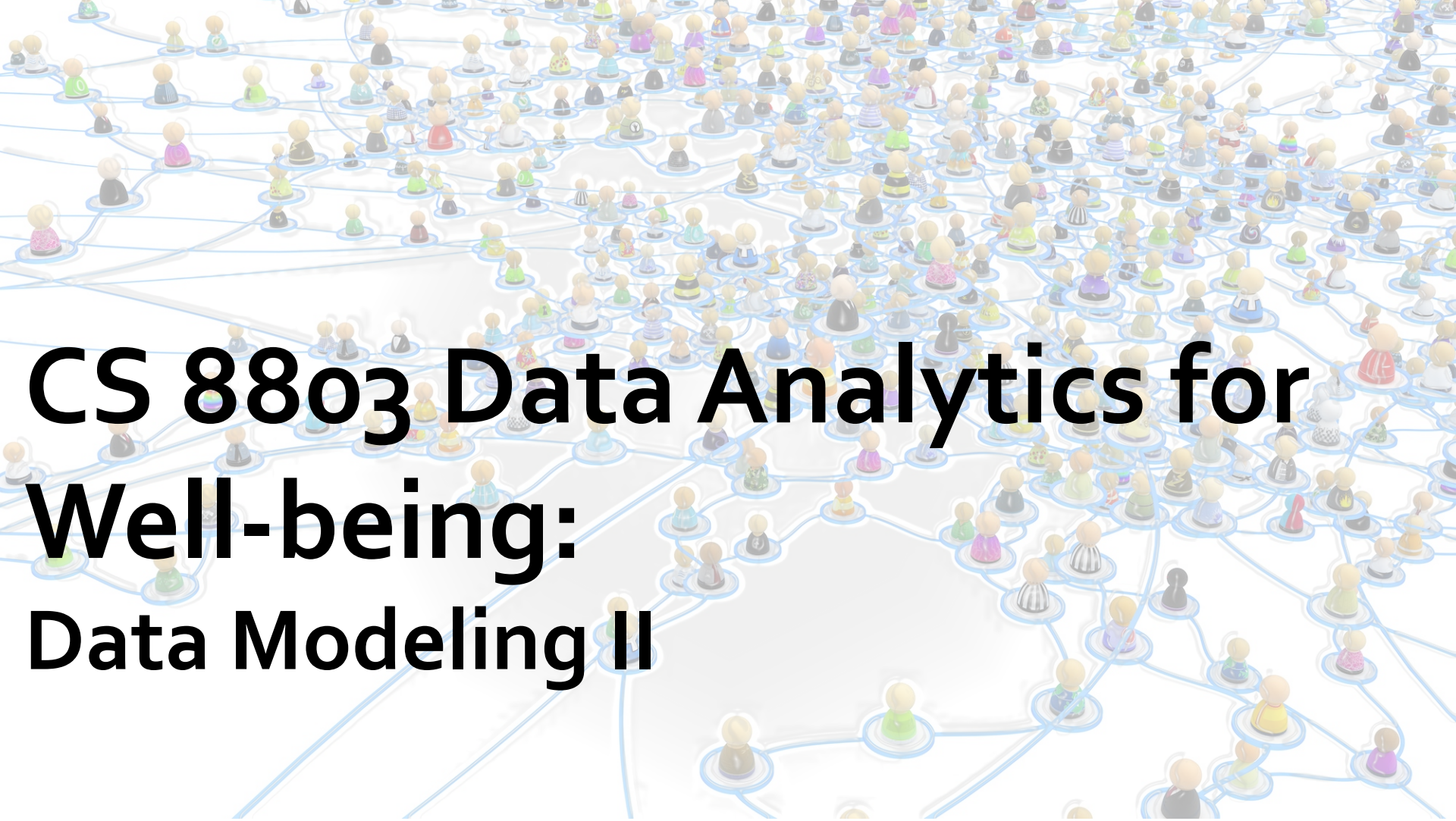


CS 8803 Data Analytics for Well-being:
Data Modeling II

Munmun De Choudhury

munmund@gatech.edu

Week 4 | February 3, 2016

Replacement Class

- At CSCW 2016 on Feb 29 and Mar 2.
- One replacement class: GVU Brown Bag talk tomorrow – distinguished speaker of the semester.
  - TSRB Ballroom 11:30 – lunch available; 12 noon-1pm – talk; 1-1:30pm – student Q&A
  - Talk title: *Online Social Support: Advances in Measuring Support and Understanding Its Effects*

Robert E. Kraut
Home Page

Herbert A. Simon Professor of Human-Computer Interaction
Human-Computer Interaction Institute, School of Computer Science
Tepper School of Business
Center for the Future of Work, Heinz School
Carnegie Mellon University
Pittsburgh, PA 15213

Home

Research

Detecting influenza epidemics using search engine query data

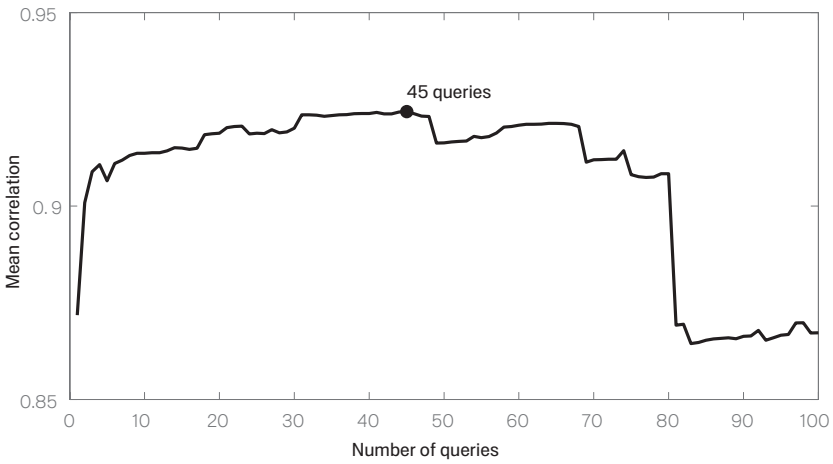
Main Idea

- The academic paper on Google Flu Trends: it relied on trends in search terms, such as headache and chills, to estimate the number of flu cases.
- The search terms were correlated with flu outbreak data collected by the Centers for Disease Control and Prevention (CDC).

Method

- By aggregating historical logs of online web search queries submitted between 2003 and 2008, the authors computed time series of weekly counts for 50 million of the most common search queries in the US.
- The authors built a linear model using the log-odds of an influenza-like illness (ILI) physician visit (CDC data from nine regions) and the log-odds of an ILI-related search query.

$$logit(P) = \beta_0 + \beta_1 \times logit(Q) + \varepsilon$$



Search Query Topic	Top 45 Queries N	Top 45 Queries Weighted	Next 55 Queries N	Next 55 Queries Weighted
Influenza Complication	11	18.15	5	3.40
Cold/Flu Remedy	8	5.05	6	5.03
General Influenza Symptoms	5	2.60	1	0.07
Term for Influenza	4	3.74	6	0.30
Specific Influenza Symptom	4	2.54	6	3.74
Symptoms of an Influenza Complication	4	2.21	2	0.92
Antibiotic Medication	3	6.23	3	3.17
General Influenza Remedies	2	0.18	1	0.32
Symptoms of a Related Disease	2	1.66	2	0.77
Antiviral Medication	1	0.39	1	0.74
Related Disease	1	6.66	3	3.77
Unrelated to Influenza	0	0.00	19	28.37
	45	**49.40**	**55**	**50.60**

Table 1: Topics found in search queries which were found to be most correlated with CDC ILI data. The top 45 queries were used in our final model; the next 55 queries are presented for comparison purposes. The number of queries in each topic is indicated, as well as query volume-weighted counts, reflecting the relative frequency of queries in each topic.

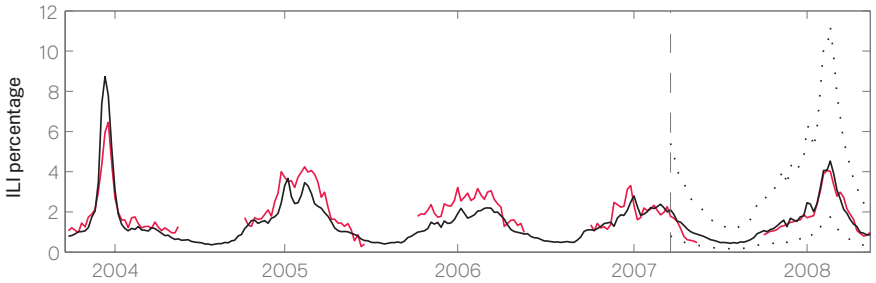
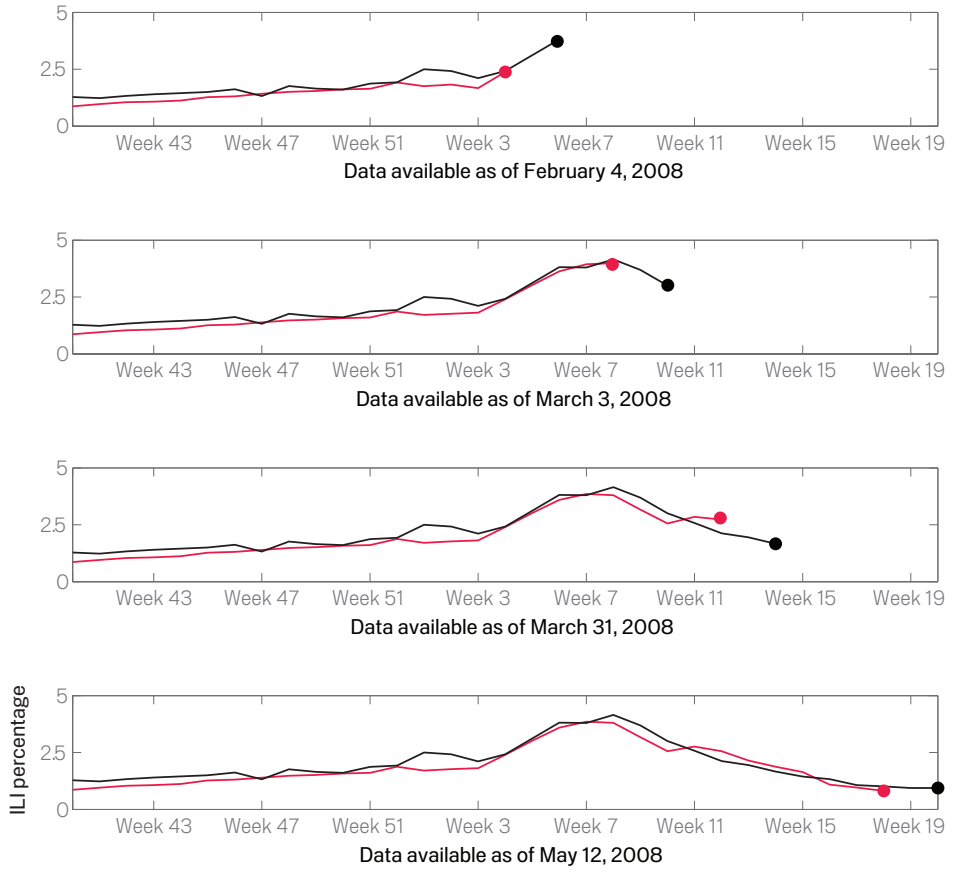


Figure 2: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of 0.96 was obtained over 42 validation points. 95% prediction intervals are indicated.



What happened later…

- Initially Google's tracker appeared to be pretty good, matching CDC data's late-breaking data somewhat closely.
- Two notable stumbles led to its ultimate downfall:
  - an underestimate of the 2009 H1N1 swine flu outbreak
  - an alarming overestimate (almost double real numbers) of the 2012-2013 flu season's cases
- But what happened really?

One of the fatal problems with the model was that it didn't account for shifts in people's search behavior.

For instance, in the 2012-2013 flu season many people were searching for news about the flu season, rather than plugging in flu symptoms for diagnostics.

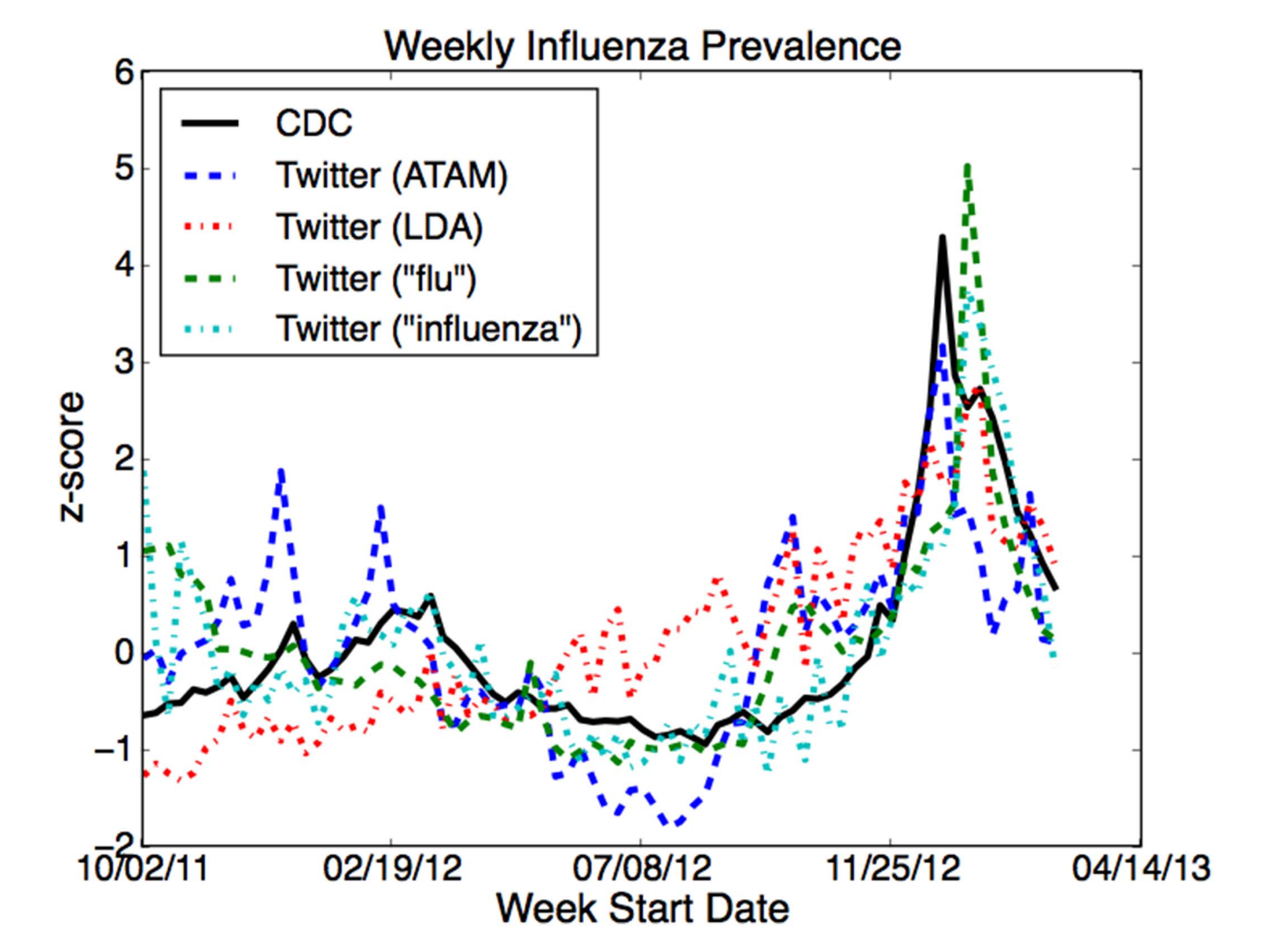
Over time, people also use different terms to search for the same things.

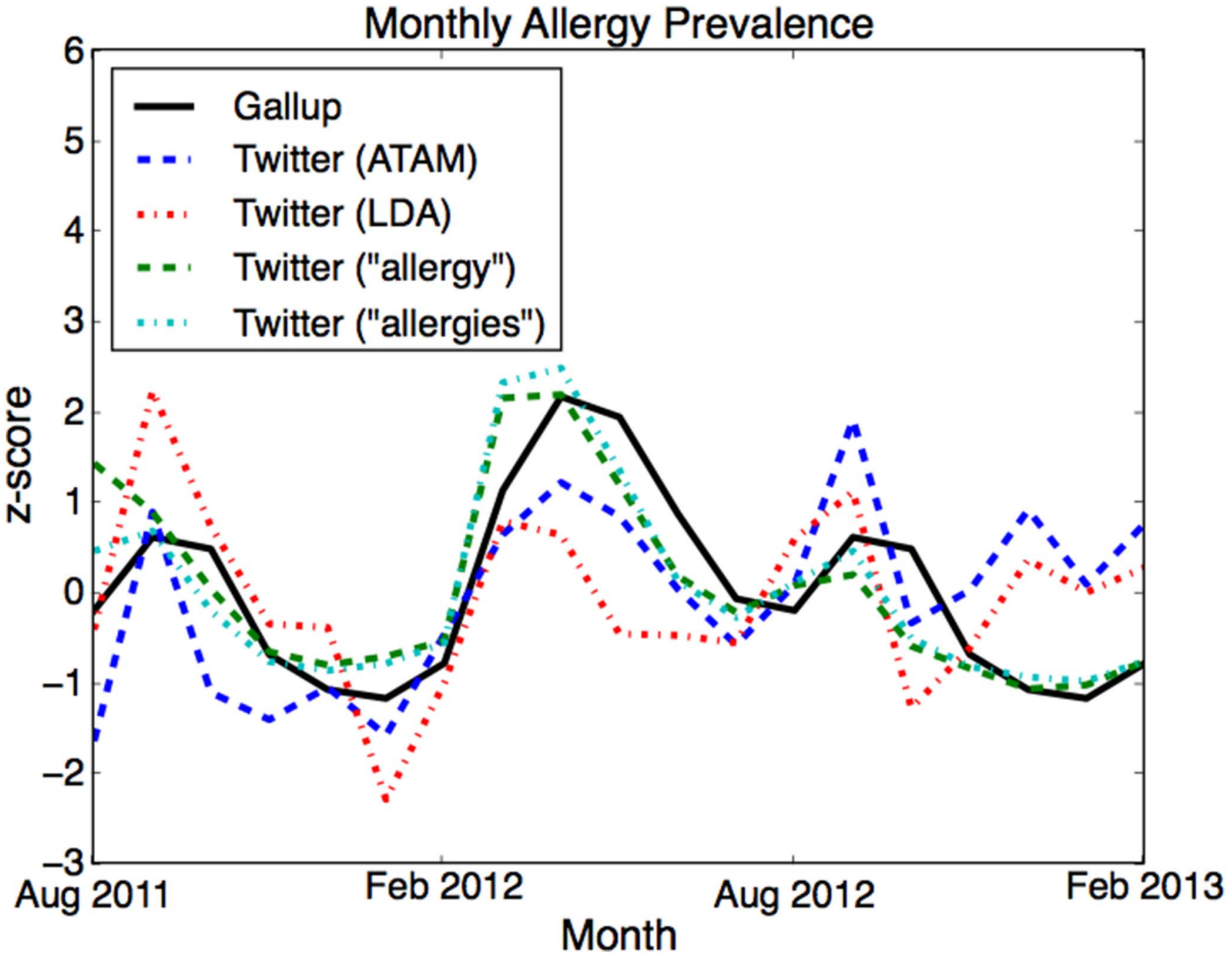
Discovering Health Topics in Social Media Using Topic Models

Main Idea

- The article proposes an unsupervised machine learning methodology to discover ailment topics in Twitter content
- The authors specifically present a statistical topic model, the Ailment Topic Aspect Model (ATAM) for the purpose – it combines a background model, a topic model and an ailment model with word (WebMD) and topic (ailment) priors

- Results demonstrate that it is possible to automatically discover topics that attain statistically significant correlations with ground truth data, despite using minimal human supervision and no historical data to train the model


Weekly Influenza Prevalence
CDC
Twitter (ATAM)
Twitter (LDA)
Twitter ("flu")
Twitter ("influenza")
z-score
6
5
4
3
2
1
0
−1
−2
10/02/11
02/19/12
07/08/12
11/25/12
04/14/13
Week Start Date


Monthly Allergy Prevalence
Gallup
Twitter (ATAM)
Twitter (LDA)
Twitter ("allergy")
Twitter ("allergies")
z-score
6
5
4
3
2
1
0
−1
−2
−3
Aug 2011
Feb 2012
Aug 2012
Feb 2013
Month

Twitter is used by millions, but could it also have bias in such measurements?

The first paper focuses on flu and the second on a variety of chronic or seasonal illnesses. Are there conditions that will be hard to capture through search logs or social media?

Could alternative social media platforms help us understand some of these other less socially sharable illnesses?

The topic modeling approach improves on simple techniques like those using dictionaries. What are its limitations?

Could there be latent factors that drive sharing of ailment related topics on social media?

People use social media for all kinds of reasons and purposes. Would that affect the moods they express?

Would “self-presentation”, “social comparison” or identity impact the kinds of illnesses shared?

The ATAM model paper does not show that the illness trends precede the CDC or the Gallup measures. How can one leverage these social media based measurements?