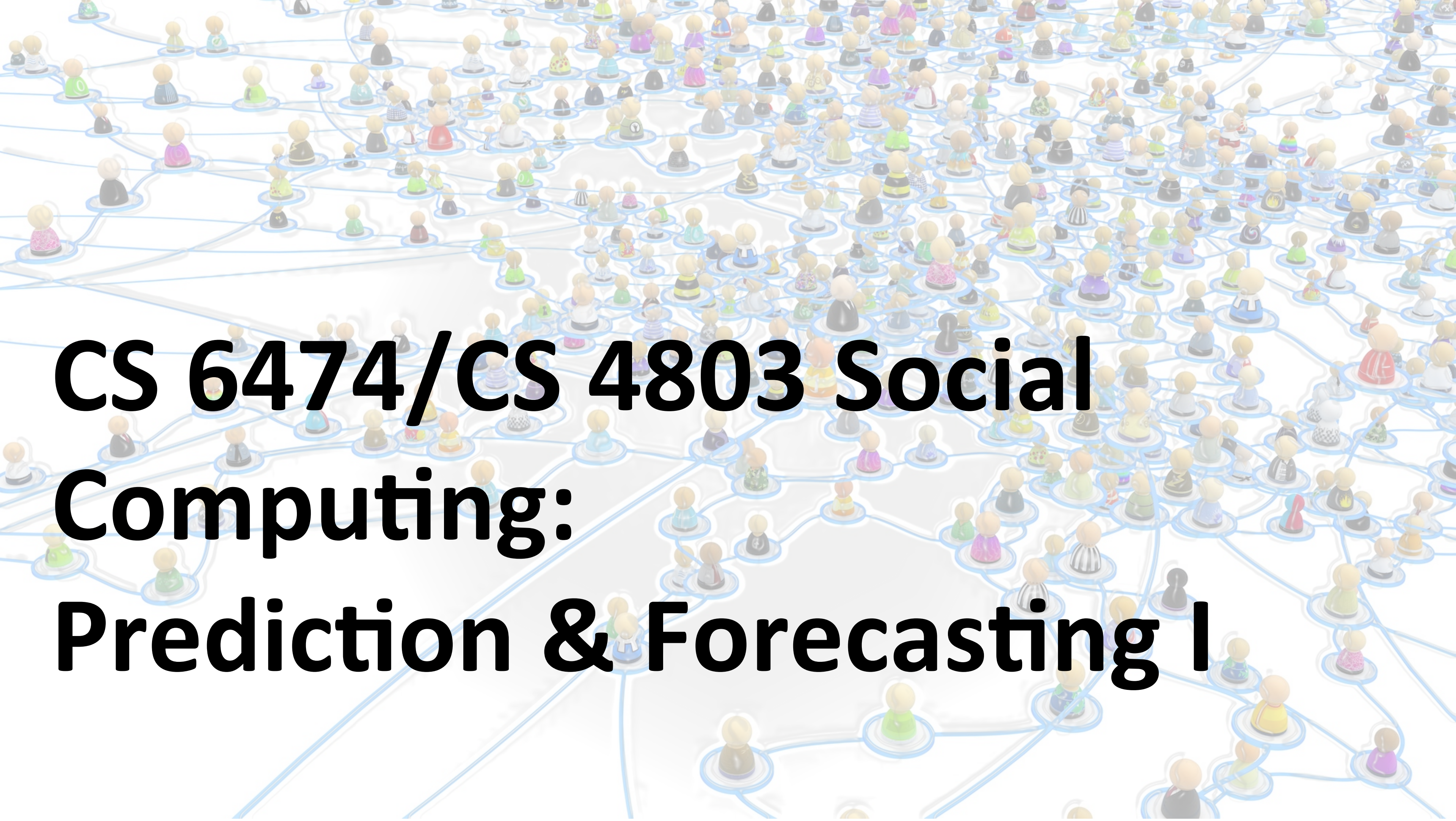
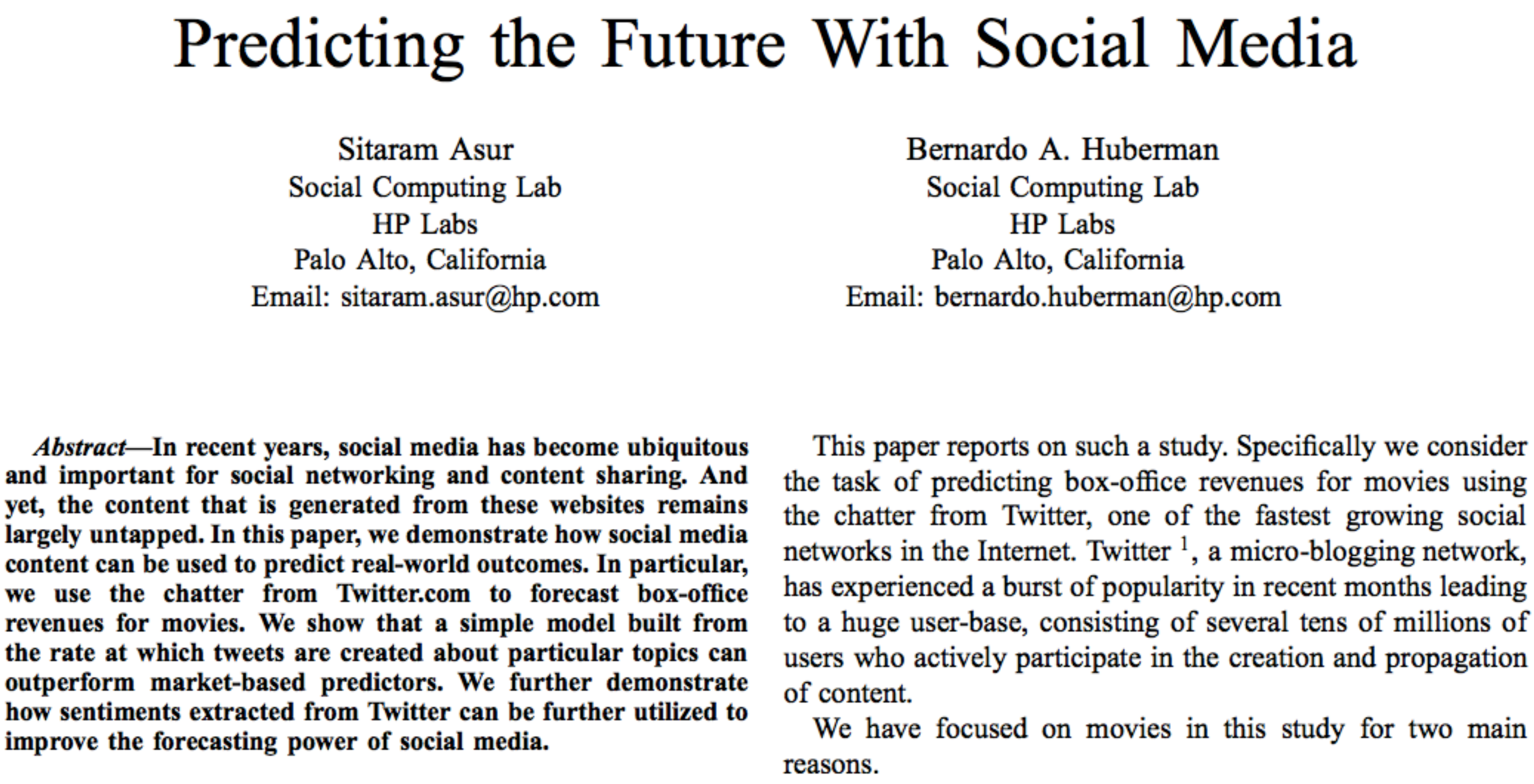


CS 6474/CS 4803 Social Computing: Prediction & Forecasting I

Munmun De Choudhury

munmund@gatech.edu

Week 13 | November 15, 2017



Predicting the Future With Social Media

Sitaram Asur
Social Computing Lab
HP Labs
Palo Alto, California
Email: sitaram.asur@hp.com

Bernardo A. Huberman
Social Computing Lab
HP Labs
Palo Alto, California
Email: bernardo.huberman@hp.com

***Abstract*—In recent years, social media has become ubiquitous and important for social networking and content sharing. And yet, the content that is generated from these websites remains largely untapped. In this paper, we demonstrate how social media content can be used to predict real-world outcomes. In particular, we use the chatter from Twitter.com to forecast box-office revenues for movies. We show that a simple model built from the rate at which tweets are created about particular topics can outperform market-based predictors. We further demonstrate how sentiments extracted from Twitter can be further utilized to improve the forecasting power of social media.**

This paper reports on such a study. Specifically we consider the task of predicting box-office revenues for movies using the chatter from Twitter, one of the fastest growing social networks in the Internet. Twitter [1], a micro-blogging network, has experienced a burst of popularity in recent months leading to a huge user-base, consisting of several tens of millions of users who actively participate in the creation and propagation of content.

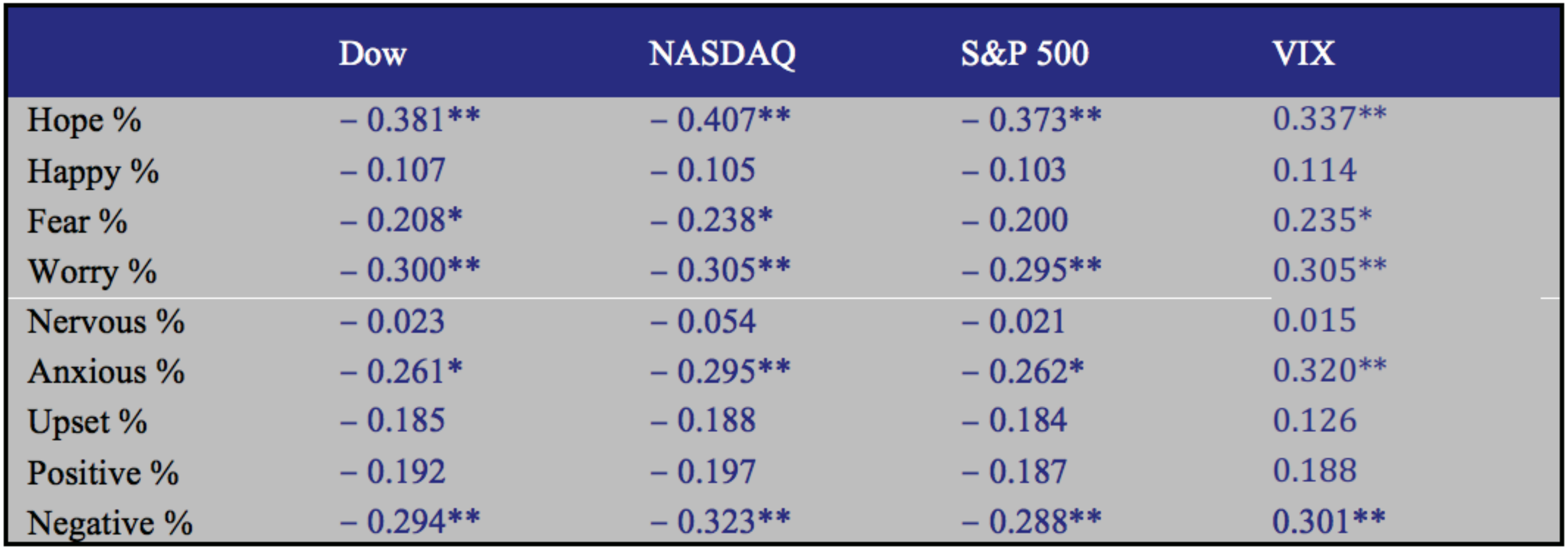
We have focused on movies in this study for two main reasons.

Social media and the decision to participate in political protest: Observations from Tahrir Square

Summary

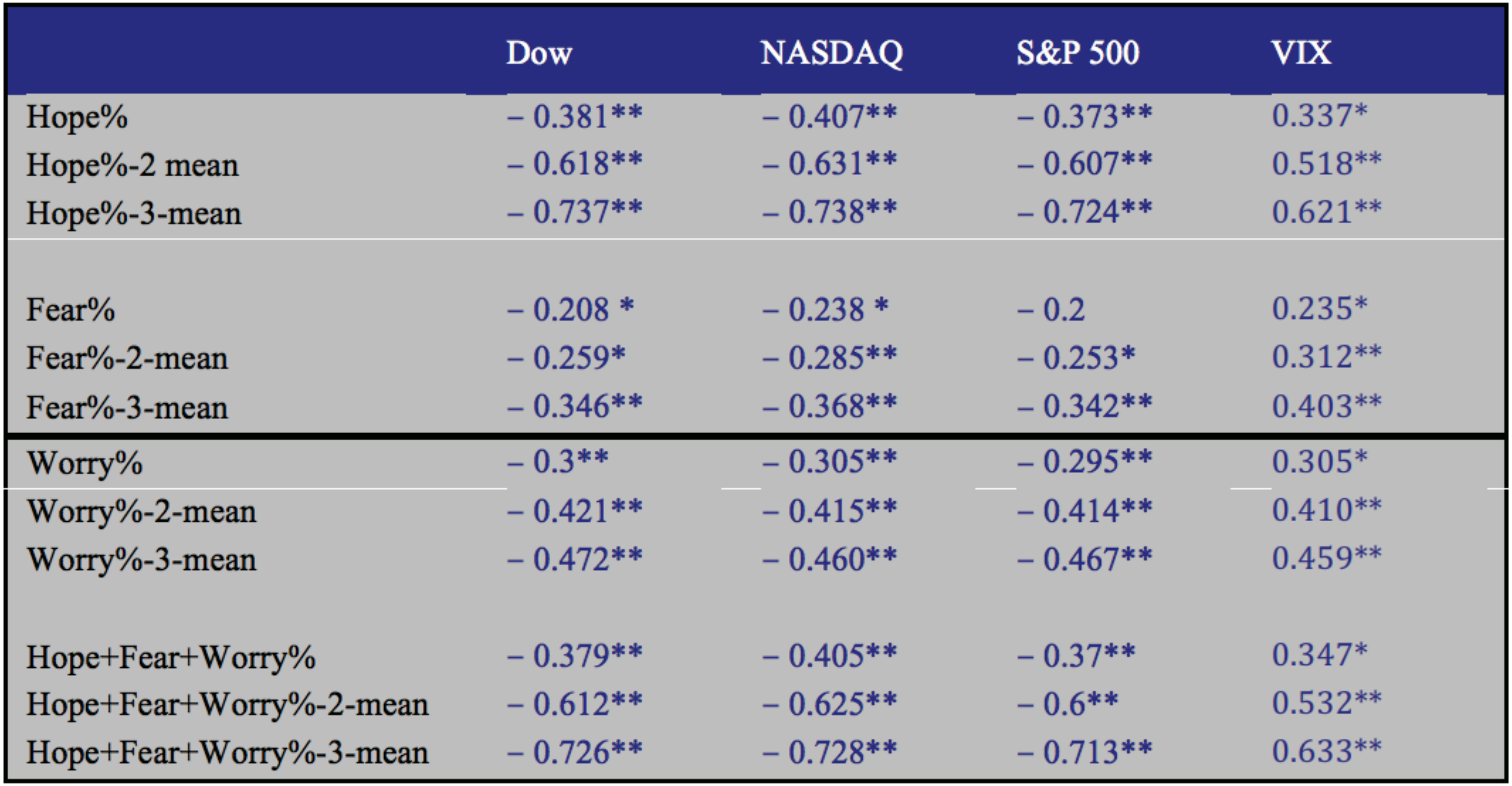
- The paper presents a study of predicting stock market indicators such as Dow Jones, NASDAQ and S&P 500 by analyzing Twitter posts
- Twitter feed of 6 months was used
- The authors measured collective hope and fear on each day and analyzed the correlation between these indices and the stock market indicators

Summary



	Dow	NASDAQ	S&P 500	VIX
Hope %	– 0.381**	– 0.407**	– 0.373**	0.337**
Happy %	– 0.107	– 0.105	– 0.103	0.114
Fear %	– 0.208*	– 0.238*	– 0.200	0.235*
Worry %	– 0.300**	– 0.305**	– 0.295**	0.305**
Nervous %	– 0.023	– 0.054	– 0.021	0.015
Anxious %	– 0.261*	– 0.295**	– 0.262*	0.320**
Upset %	– 0.185	– 0.188	– 0.184	0.126
Positive %	– 0.192	– 0.197	– 0.187	0.188
Negative %	– 0.294**	– 0.323**	– 0.288**	0.301**

Summary



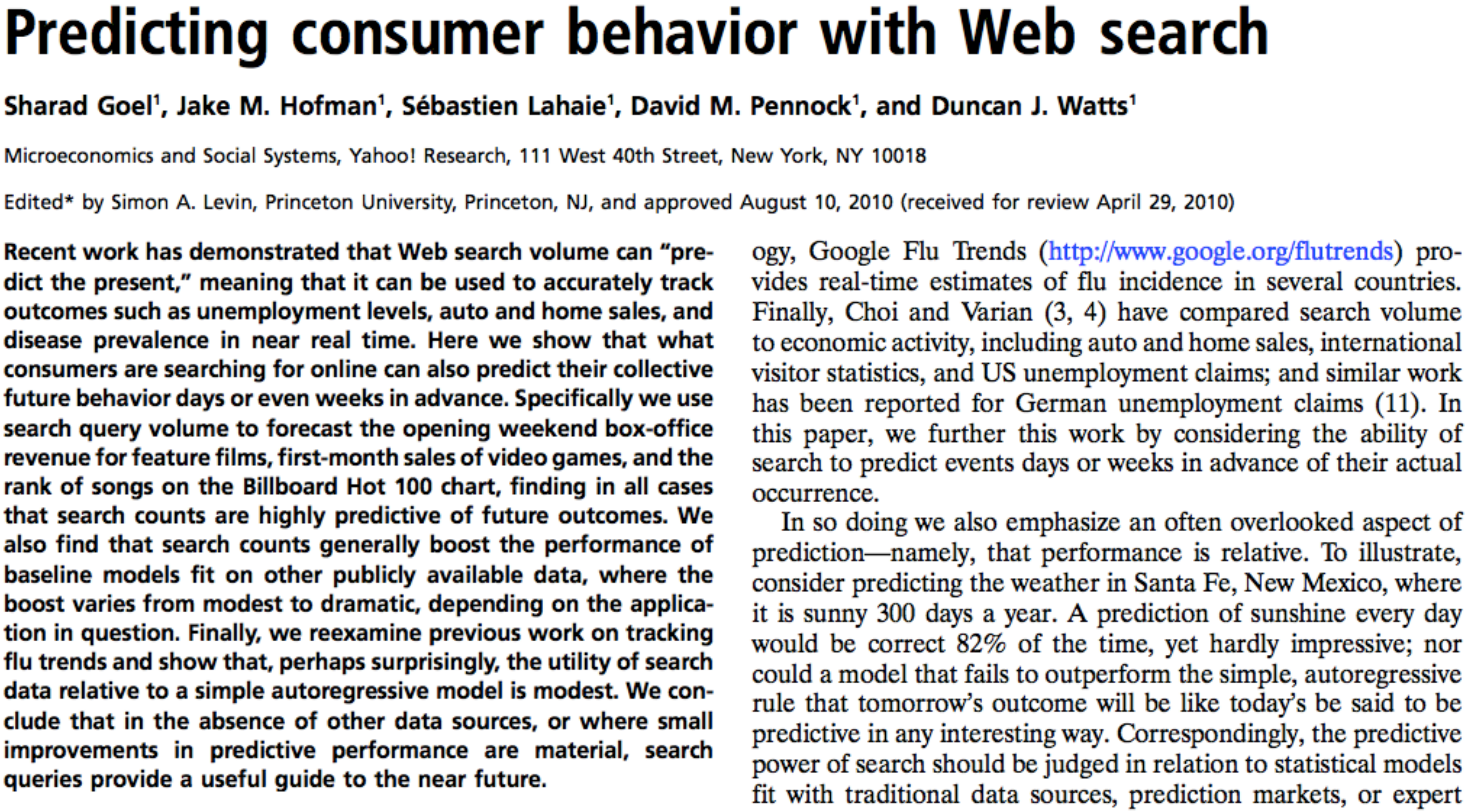
	Dow	NASDAQ	S&P 500	VIX
Hope%	– 0.381**	– 0.407**	– 0.373**	0.337*
Hope%-2 mean	– 0.618**	– 0.631**	– 0.607**	0.518**
Hope%-3-mean	– 0.737**	– 0.738**	– 0.724**	0.621**
Fear%	– 0.208 *	– 0.238 *	– 0.2	0.235*
Fear%-2-mean	– 0.259*	– 0.285**	– 0.253*	0.312**
Fear%-3-mean	– 0.346**	– 0.368**	– 0.342**	0.403**
Worry%	– 0.3**	– 0.305**	– 0.295**	0.305*
Worry%-2-mean	– 0.421**	– 0.415**	– 0.414**	0.410**
Worry%-3-mean	– 0.472**	– 0.460**	– 0.467**	0.459**
Hope+Fear+Worry%	– 0.379**	– 0.405**	– 0.37**	0.347*
Hope+Fear+Worry%-2-mean	– 0.612**	– 0.625**	– 0.6**	0.532**
Hope+Fear+Worry%-3-mean	– 0.726**	– 0.728**	– 0.713**	0.633**

Class Exercise I

Discuss how can social media based stock market predictors could be helpful. Who will it help?

All the things that social media can predict!

- Micro and macro economic indicators – stock markets, (un)employment, consumer sentiment
- Elections – who will win the next election?
- Politics – political phenomena
- Marketing – will a product be successful? Will it sell well?
- Box office – how well will this movie do at the box office?
- Flu – which cities and counties will be infected faster than others?



Predicting consumer behavior with Web search

Sharad Goel[1], Jake M. Hofman[1], Sébastien Lahaie[1], David M. Pennock[1], and Duncan J. Watts[1]

Microeconomics and Social Systems, Yahoo! Research, 111 West 40th Street, New York, NY 10018

Edited* by Simon A. Levin, Princeton University, Princeton, NJ, and approved August 10, 2010 (received for review April 29, 2010)

Recent work has demonstrated that Web search volume can "predict the present," meaning that it can be used to accurately track outcomes such as unemployment levels, auto and home sales, and disease prevalence in near real time. Here we show that what consumers are searching for online can also predict their collective future behavior days or even weeks in advance. Specifically we use search query volume to forecast the opening weekend box-office revenue for feature films, first-month sales of video games, and the rank of songs on the Billboard Hot 100 chart, finding in all cases that search counts are highly predictive of future outcomes. We also find that search counts generally boost the performance of baseline models fit on other publicly available data, where the boost varies from modest to dramatic, depending on the application in question. Finally, we reexamine previous work on tracking flu trends and show that, perhaps surprisingly, the utility of search data relative to a simple autoregressive model is modest. We conclude that in the absence of other data sources, or where small improvements in predictive performance are material, search queries provide a useful guide to the near future.

ogy, Google Flu Trends (http://www.google.org/flutrends) provides real-time estimates of flu incidence in several countries. Finally, Choi and Varian (3, 4) have compared search volume to economic activity, including auto and home sales, international visitor statistics, and US unemployment claims; and similar work has been reported for German unemployment claims (11). In this paper, we further this work by considering the ability of search to predict events days or weeks in advance of their actual occurrence.

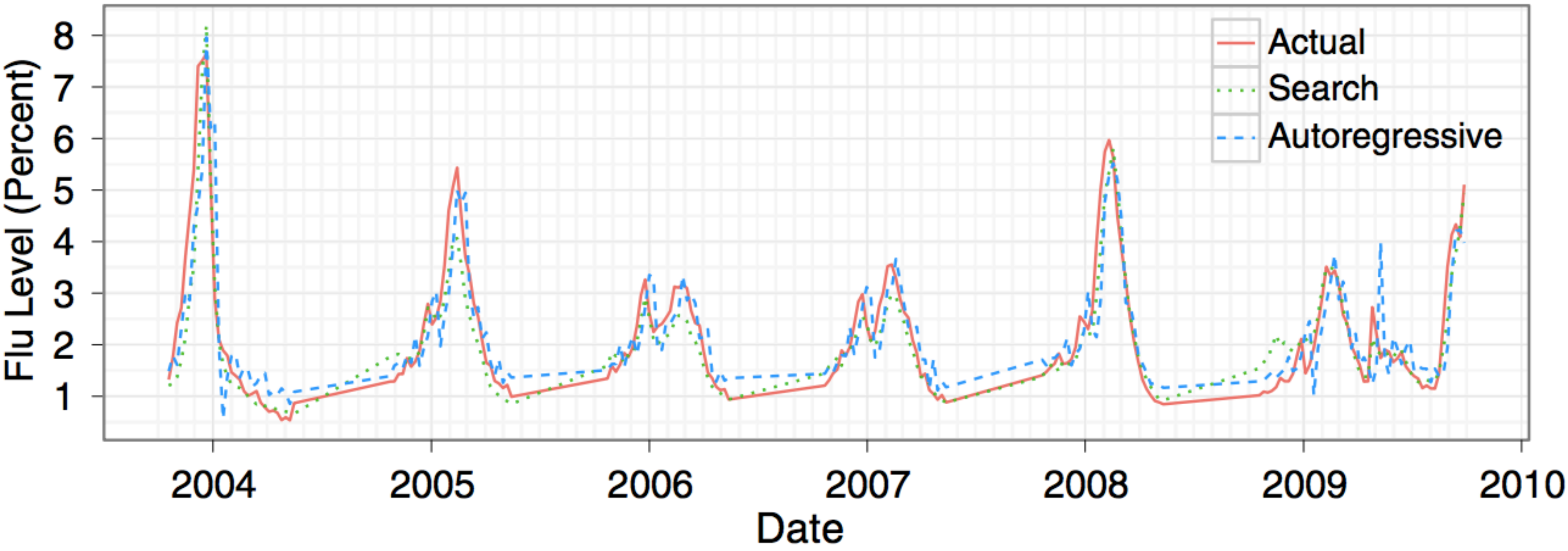
In so doing we also emphasize an often overlooked aspect of prediction—namely, that performance is relative. To illustrate, consider predicting the weather in Santa Fe, New Mexico, where it is sunny 300 days a year. A prediction of sunshine every day would be correct 82% of the time, yet hardly impressive; nor could a model that fails to outperform the simple, autoregressive rule that tomorrow's outcome will be like today's be said to be predictive in any interesting way. Correspondingly, the predictive power of search should be judged in relation to statistical models fit with traditional data sources, prediction markets, or expert

Predicting consumer behavior with Web search

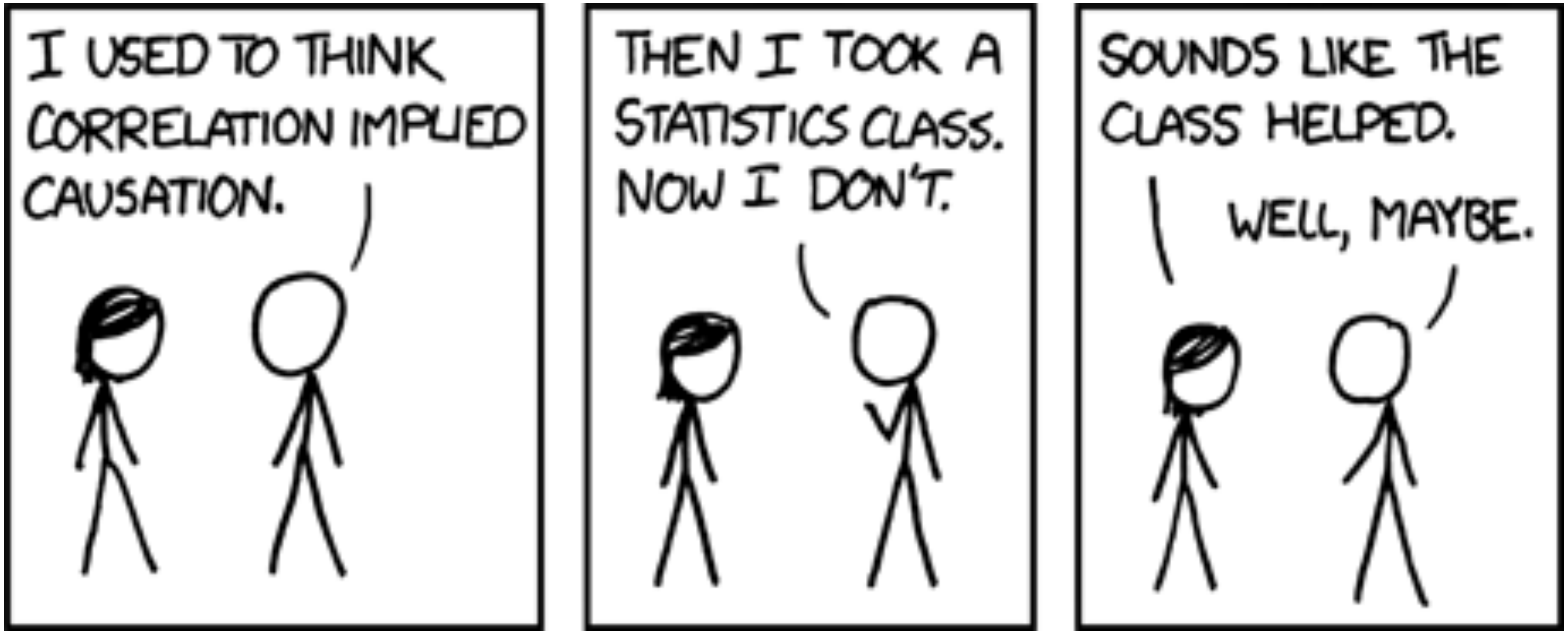
Sharad Goel[1], Jake M. Hofman[1], Sébastien Lahaie[1], David M. Pennock[1], and Duncan J. Watts[1]

Microeconomics and Social Systems, Yahoo! Research, 111 West 40th Street, New York, NY 10018

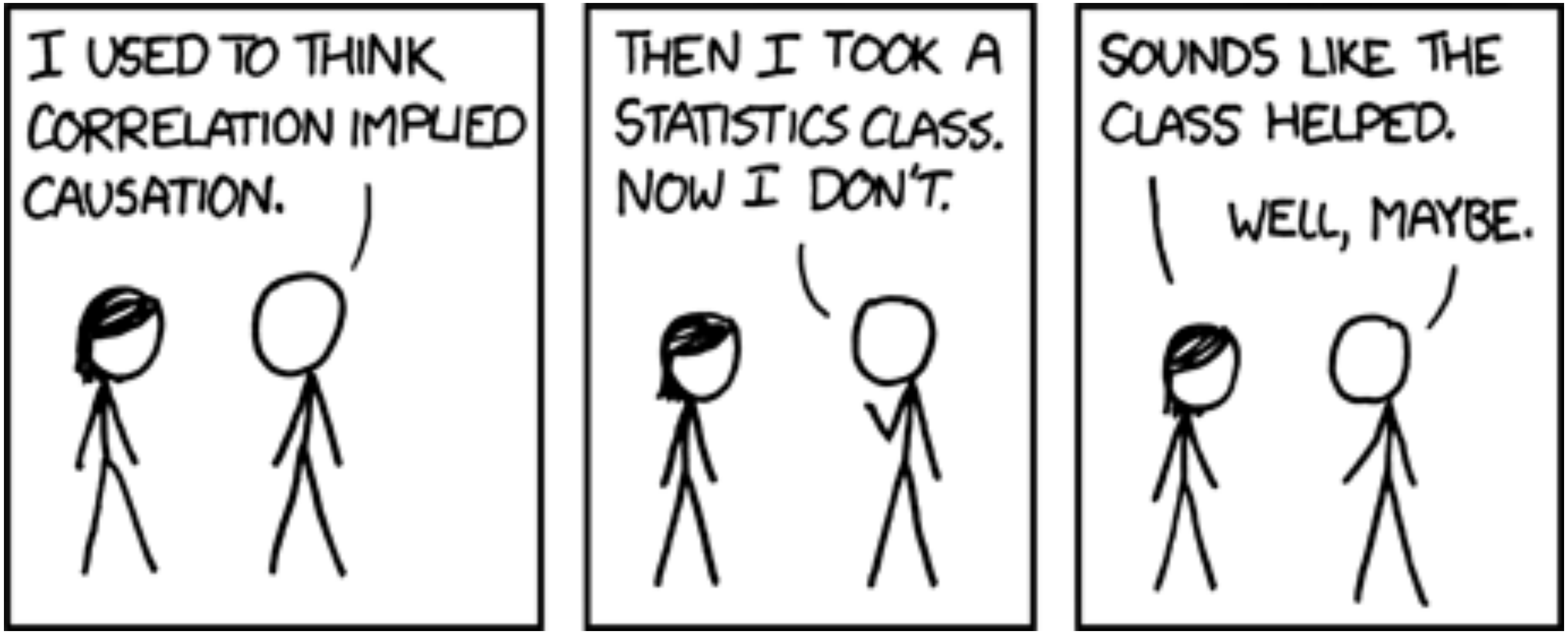
Edited* by Simon A. Levin, Princeton University, Princeton, NJ, and approved August 10, 2010 (received for review April 29, 2010)



If that is true, then when are predictions using social media (and similar data sources) helpful?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Correlation and causation

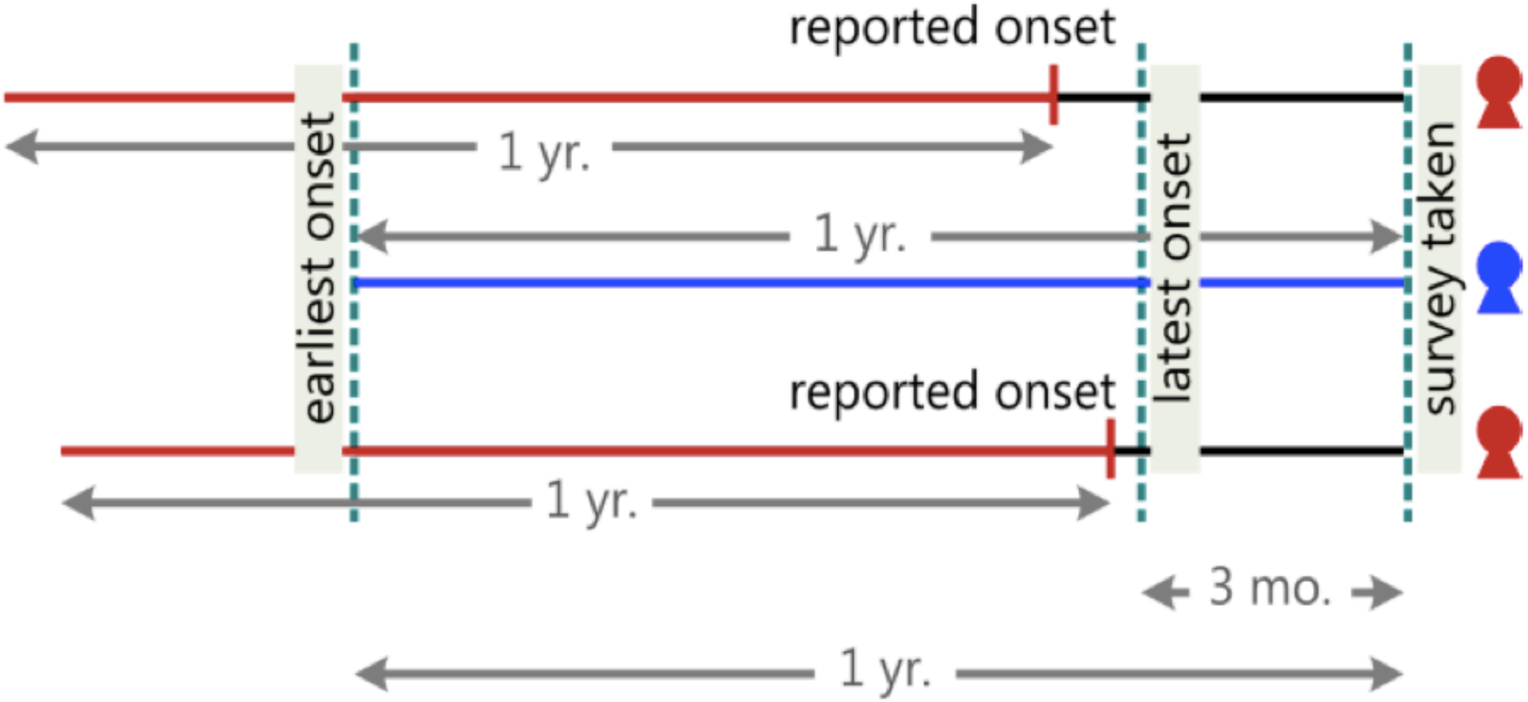

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Do you think the moods observed on Twitter are causal in predicting stock market indices? Can Twitter predict future indices?

Predicting Depression via Social Media

Summary

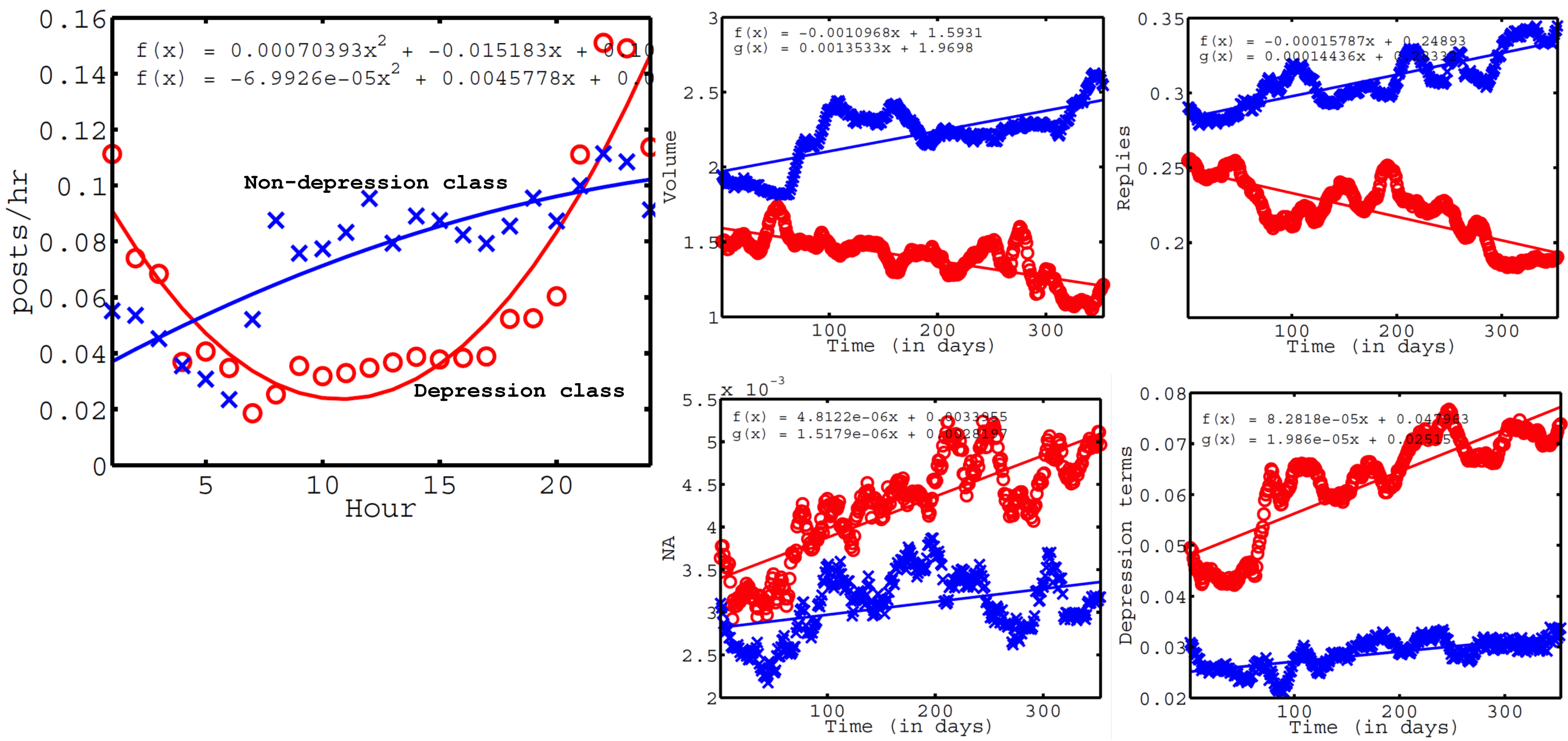
- Can social media activities and connectedness predict risk to major depressive disorder?
- Recruitment of a sample of Twitter users through a survey methodology over Amazon's Mechanical Turk
  - ~40% provided access to Twitter data



Summary

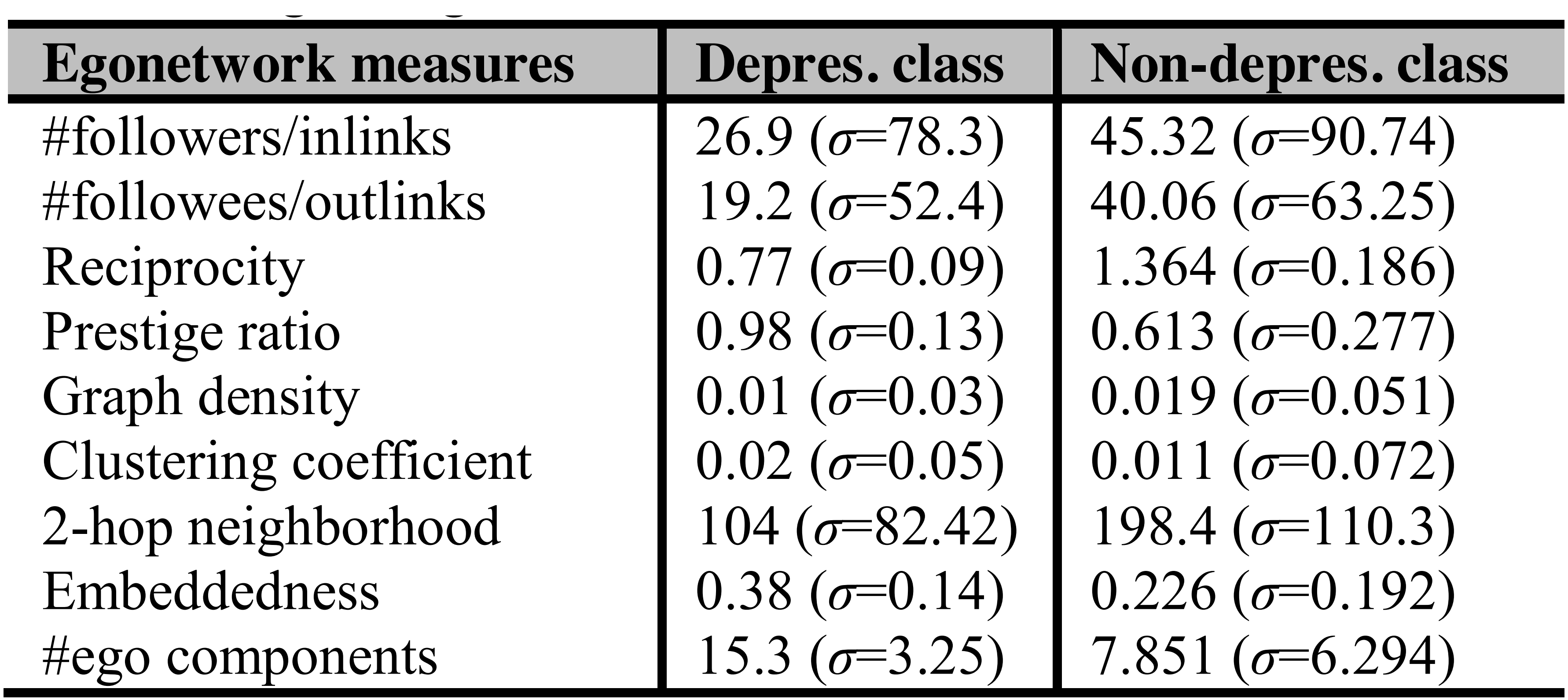
- Social engagement
- "Insomnia index" – mean *z*-score of an individual's volume of Twitter activity per hour
- Ego-centric social graph – nodal properties (*inlinks, outlinks*); dyadic properties (*reciprocity, interpersonal exchange*); neighborhood properties (*density, clustering coefficient, two-hop neighborhood, embeddedness, number of ego components*)
- Language
  - Depression lexicon – top uni- and bigrams compiled from Yahoo! Answers category on mental health
  - Linguistic style

Summary



f(x) = 0.00070393x2 + -0.015183x + 0.10
f(x) = -6.9926e-05x2 + 0.0045778x + 0.0
Non-depression class
Depression class
posts/hr
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
5
10
15
20
Hour
f(x) = -0.0010968x + 1.5931
g(x) = 0.0013533x + 1.9698
Volume
3
2.5
2
1.5
1
100
200
300
Time (in days)
f(x) = -0.00015787x + 0.24893
g(x) = 0.00014436x + 0.28332
Replies
0.35
0.3
0.25
0.2
100
200
300
Time (in days)
x 10-3
f(x) = 4.8122e-06x + 0.0033955
g(x) = 1.5179e-06x + 0.0028197
NA
5.5
5
4.5
4
3.5
3
2.5
2
100
200
300
Time (in days)
f(x) = 8.2818e-05x + 0.047963
g(x) = 1.986e-05x + 0.02515
Depression terms
0.08
0.07
0.06
0.05
0.04
0.03
0.02
100
200
300
Time (in days)

Summary



Egonetwork measures	Depres. class	Non-depres. class
#followers/inlinks	26.9 (σ=78.3)	45.32 (σ=90.74)
#followees/outlinks	19.2 (σ=52.4)	40.06 (σ=63.25)
Reciprocity	0.77 (σ=0.09)	1.364 (σ=0.186)
Prestige ratio	0.98 (σ=0.13)	0.613 (σ=0.277)
Graph density	0.01 (σ=0.03)	0.019 (σ=0.051)
Clustering coefficient	0.02 (σ=0.05)	0.011 (σ=0.072)
2-hop neighborhood	104 (σ=82.42)	198.4 (σ=110.3)
Embeddedness	0.38 (σ=0.14)	0.226 (σ=0.192)
#ego components	15.3 (σ=3.25)	7.851 (σ=6.294)

In the depression prediction paper, the ground truth was obtained from Amazon mechanical turk workers. Anything unique about this population that may have affected the findings? What would be alternative ways of recruiting people?

A consistent challenge in many prediction tasks like the one on predicting emotional or psychological state, is gathering gold standard information (or ground truth).

What could be different ways to get at this problem?

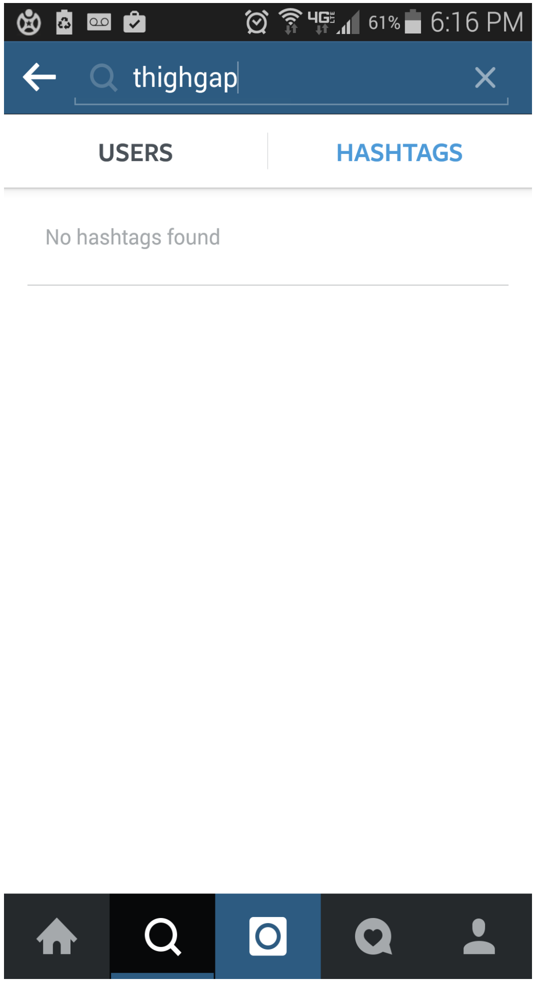
Depression is not an online condition, but one that spans both the online and the offline life. The paper does not take offline attributes into their models.

Is there a way to that into account? What would be the most significant offline attributes to consider?

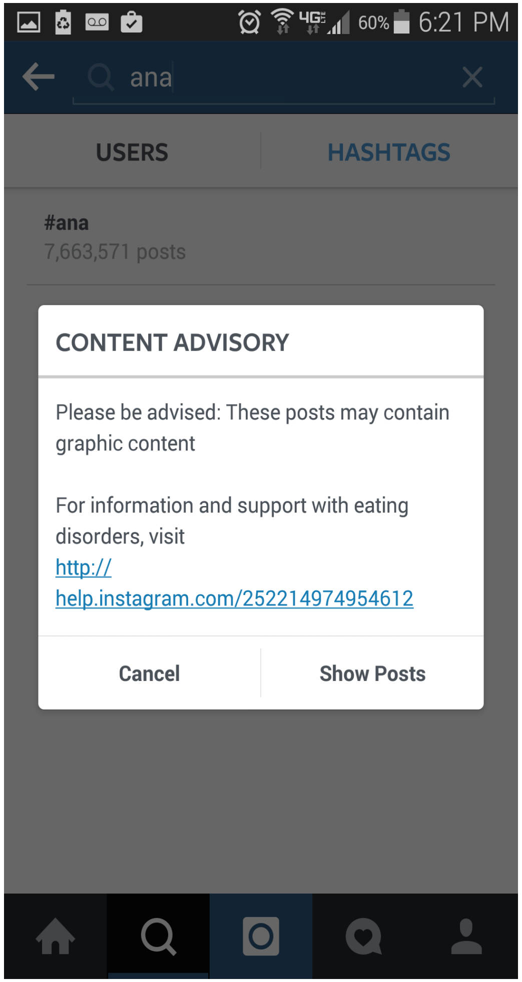
Class Exercise II

Discuss how can social media based depression (or other mental health condition) predictors could be helpful to people.

Improving “Blanket” Interventions



No hashtags found

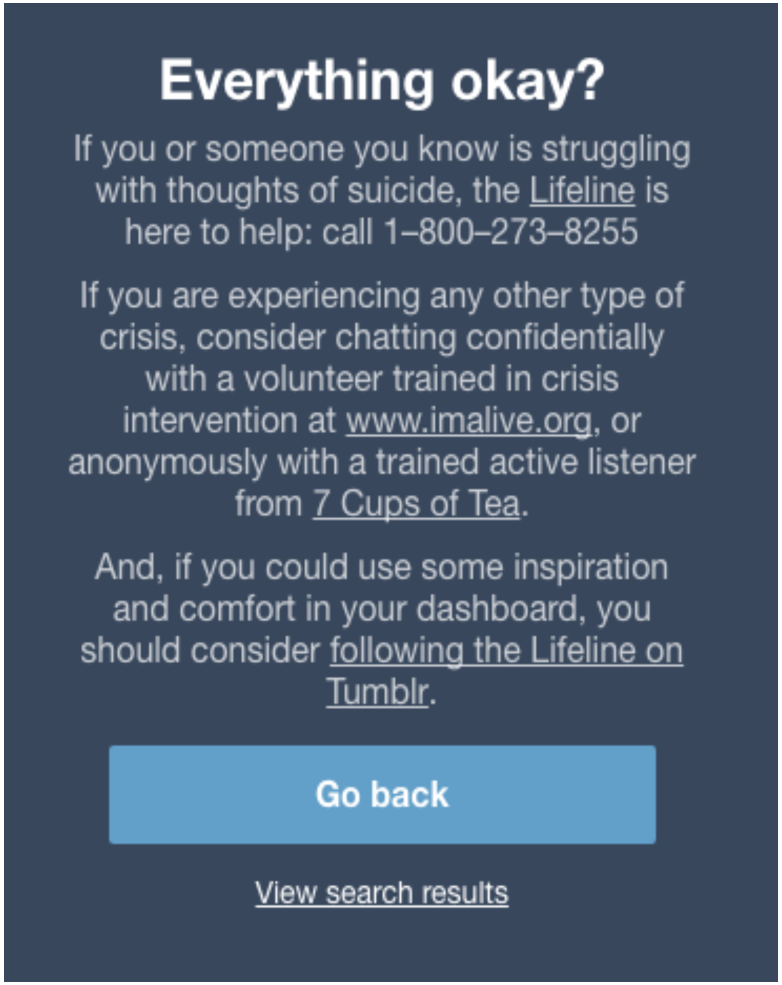


#ana
7,663,571 posts

CONTENT ADVISORY

Please be advised: These posts may contain graphic content

For information and support with eating disorders, visit http://help.instagram.com/252214974954612

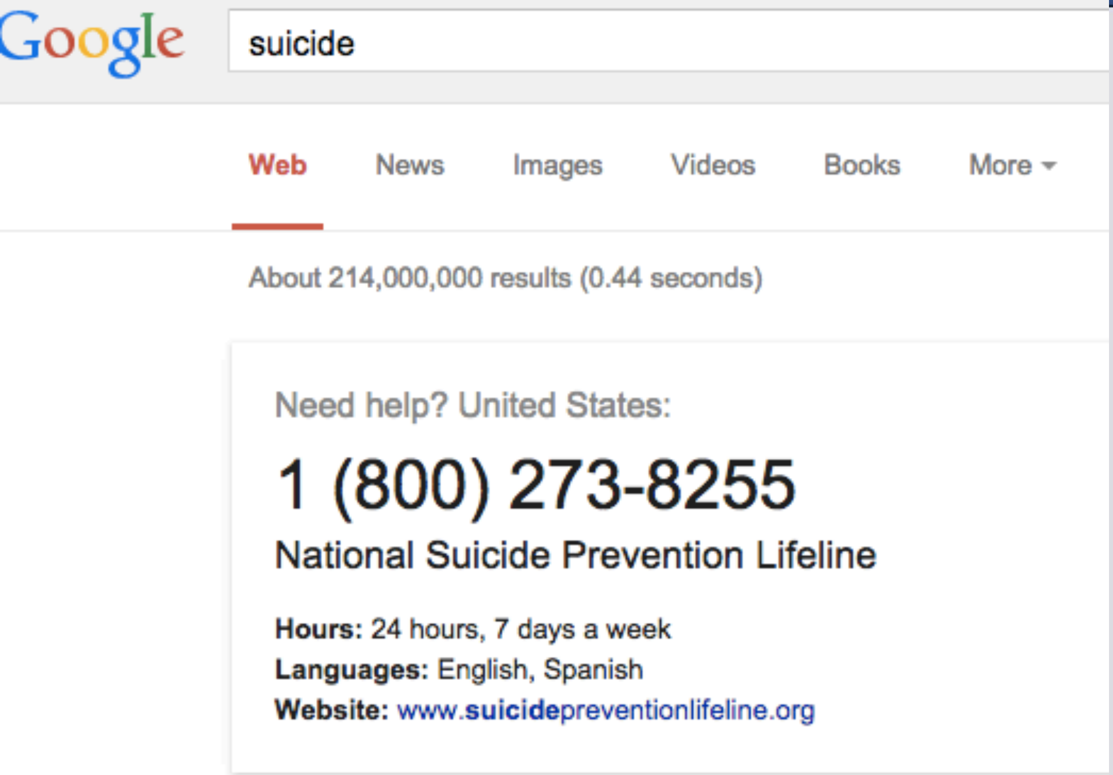


Everything okay?

If you or someone you know is struggling with thoughts of suicide, the Lifeline is here to help: call 1–800–273–8255

If you are experiencing any other type of crisis, consider chatting confidentially with a volunteer trained in crisis intervention at www.imalive.org, or anonymously with a trained active listener from 7 Cups of Tea.

And, if you could use some inspiration and comfort in your dashboard, you should consider following the Lifeline on Tumblr.



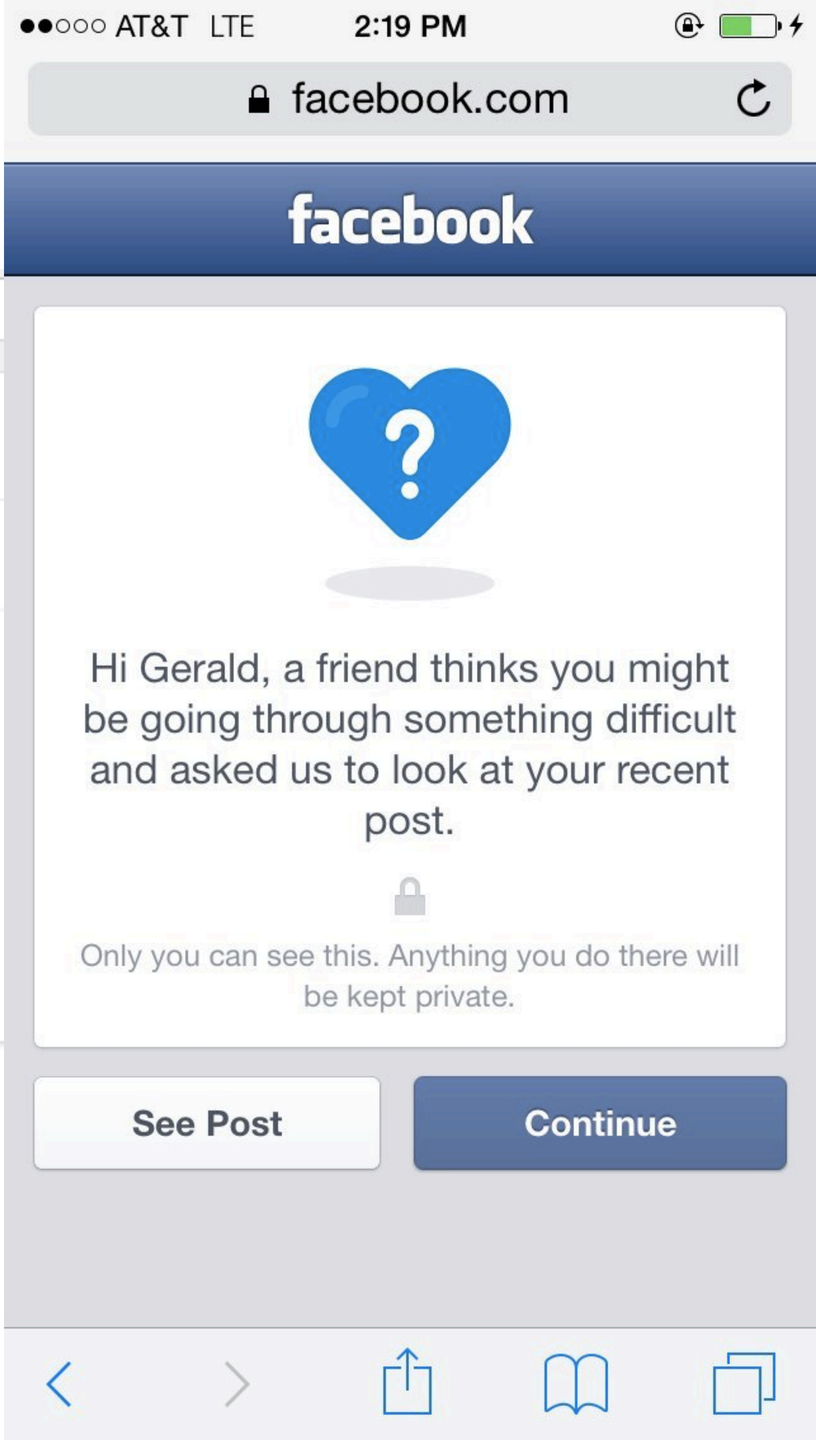
About 214,000,000 results (0.44 seconds)

Need help? United States:

1 (800) 273-8255

National Suicide Prevention Lifeline

Hours: 24 hours, 7 days a week
Languages: English, Spanish
Website: www.suicidepreventionlifeline.org



Hi Gerald, a friend thinks you might be going through something difficult and asked us to look at your recent post.

Only you can see this. Anything you do there will be kept private.

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

Class Exercise III

Would the results of the papers generalize to non-Twitter social media? Why or why not?

Predicting flu from search query (Google) data

	Top 45 Queries		Next 55 Queries	
Search Query Topic	N	Weighted	N	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.

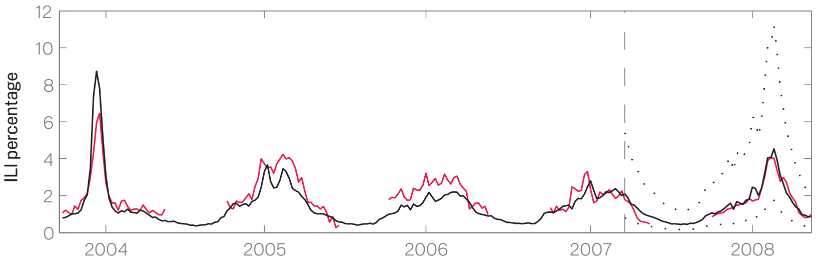
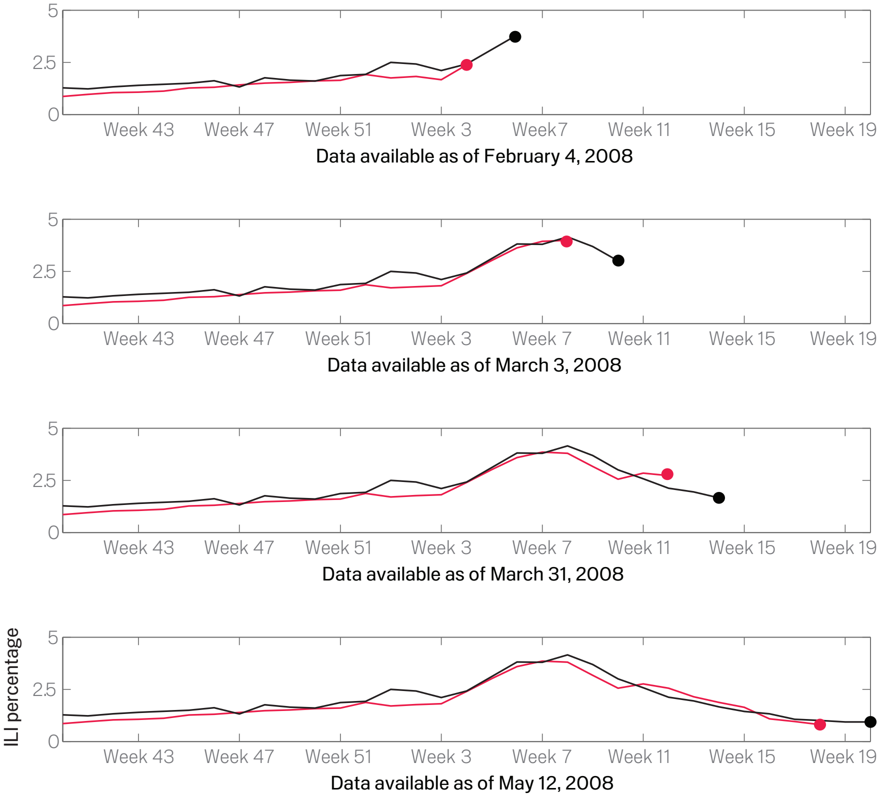


Figure 2: A comparison of model estimates for the Mid-Atlantic Region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, while a correlation of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

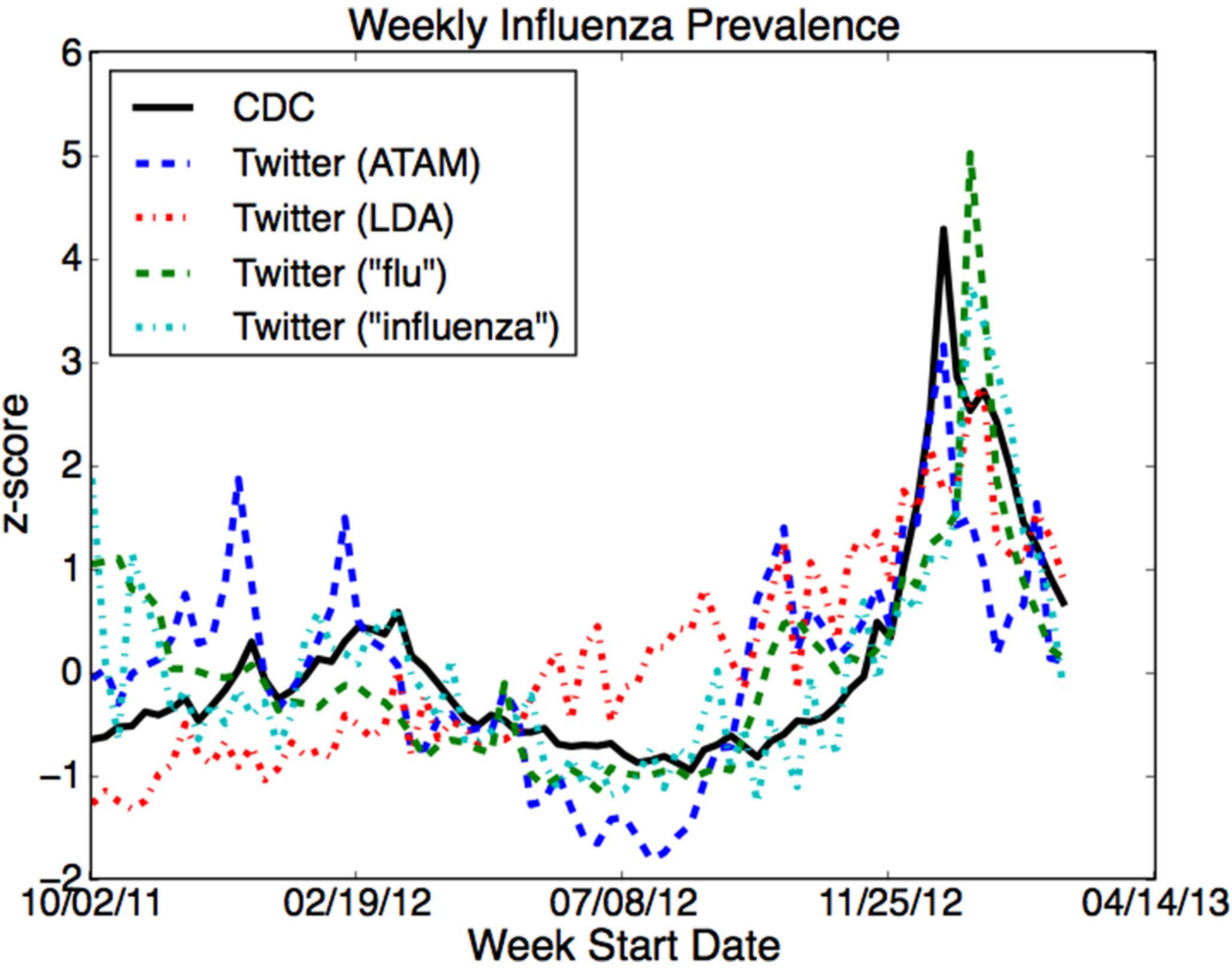


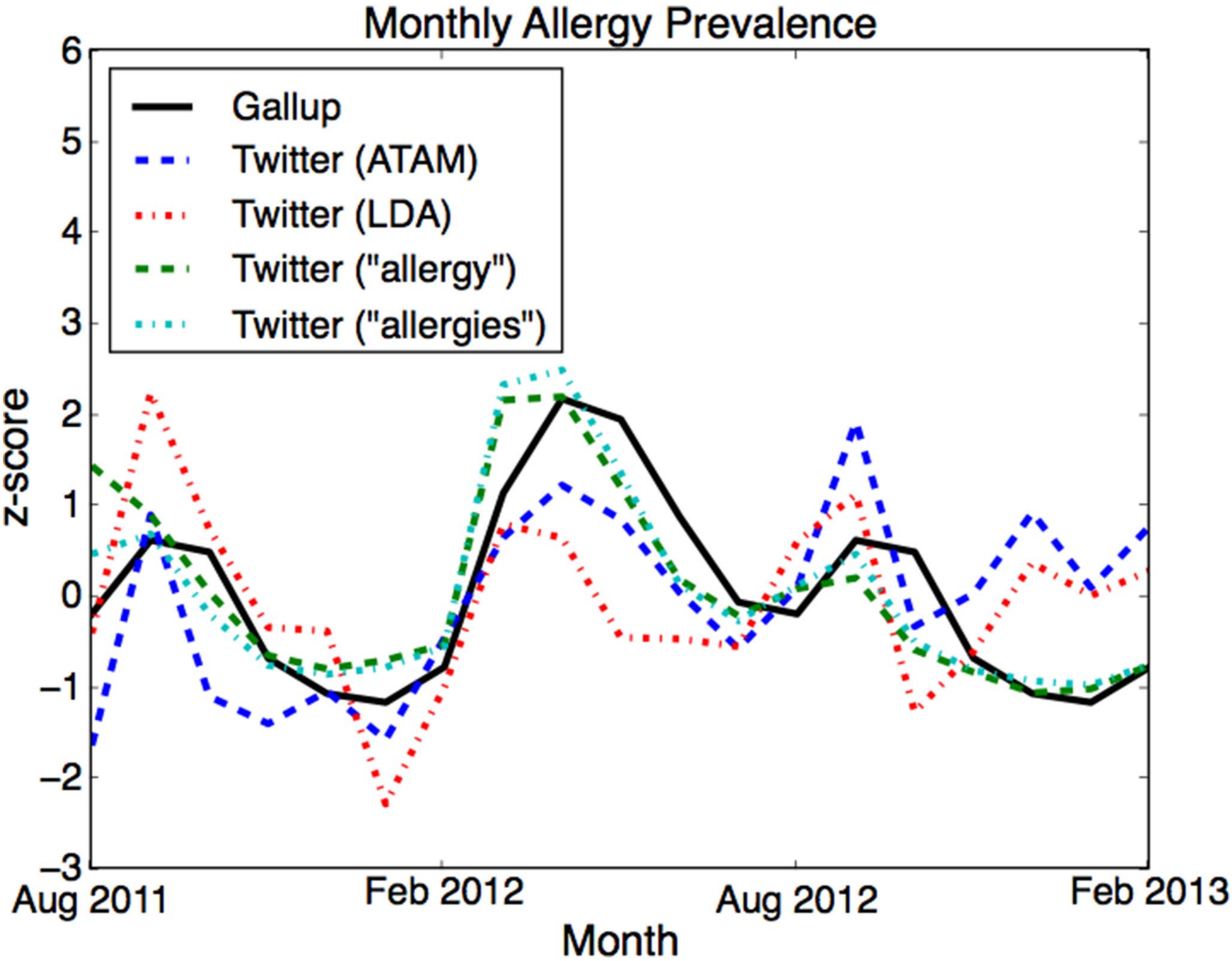
Data available as of February 4, 2008

Data available as of March 3, 2008

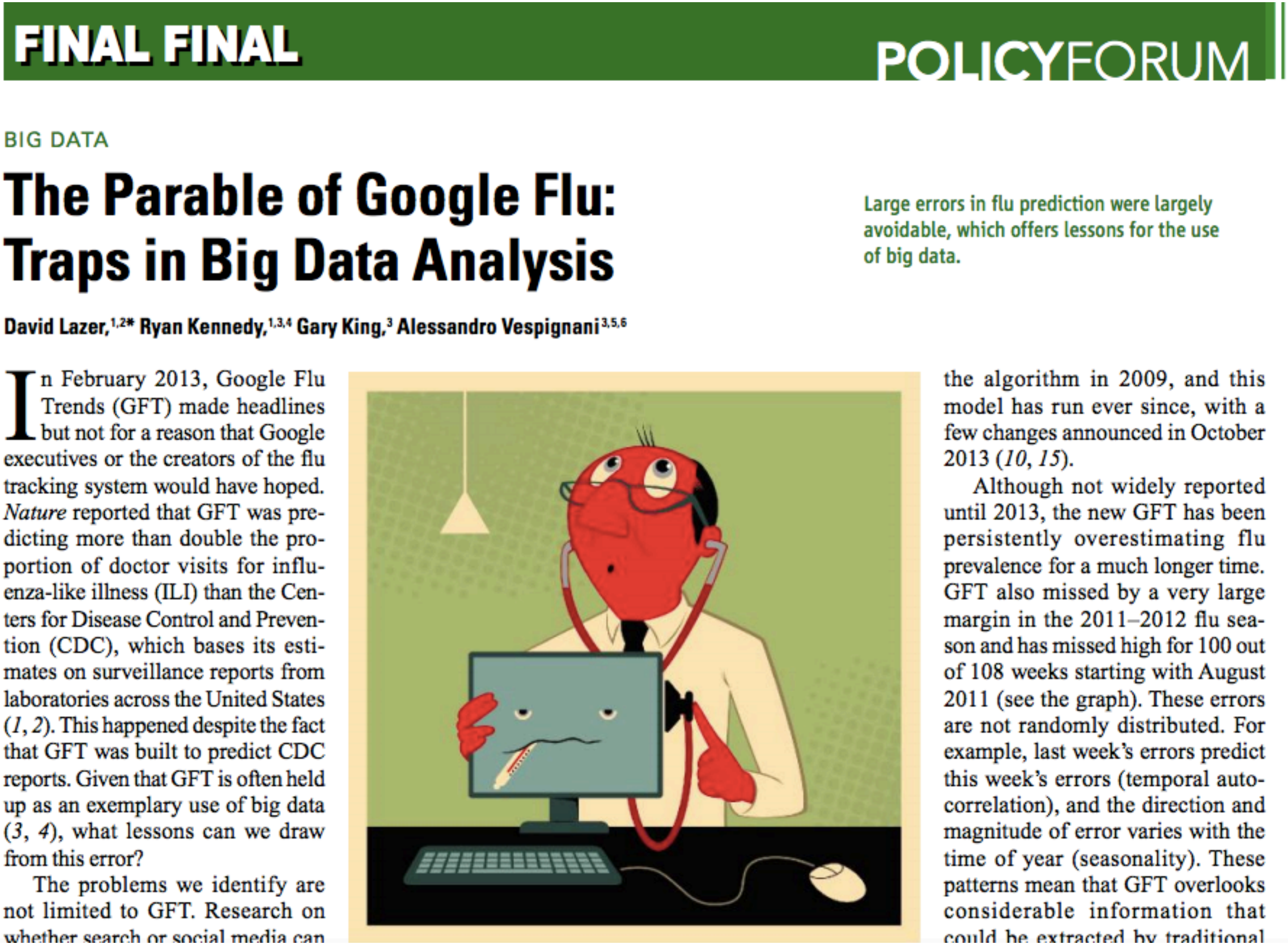
Data available as of March 31, 2008

Data available as of May 12, 2008


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

What are the limits of prediction? Can they fail?



FINAL FINAL

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[3,5,6]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can

the algorithm in 2009, and this model has run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional