CS 8803 Social Computing: Trends and Forecasting

Munmun De Choudhury

munmund@gatech.edu

Week 13 | November 10, 2014

Wednesday Nov 19's class

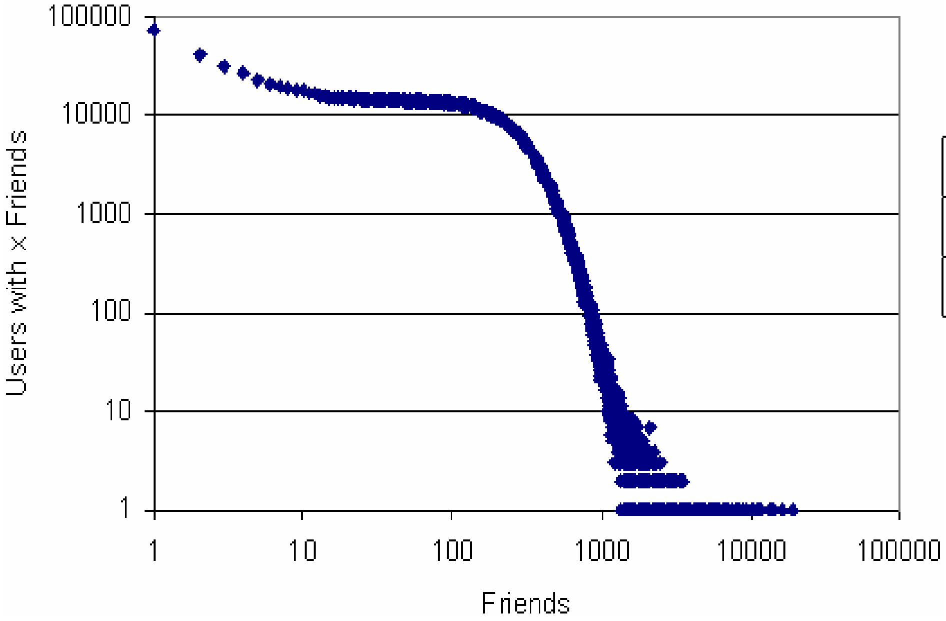
- Atlanta Computational Social Science Workshop
- Schedule: http://css-workshop.gatech.edu/schedule.html
- Attend Prof. Noah Smith's (CMU) talk – 10am to 11am, Friday Nov 21
  - Talk title: "Machine Learning About People From Their Language"
- *OR*
- Attend Prof. Arthur Spirling's (Harvard) talk – 1:30-2:30pm, Friday Nov 21

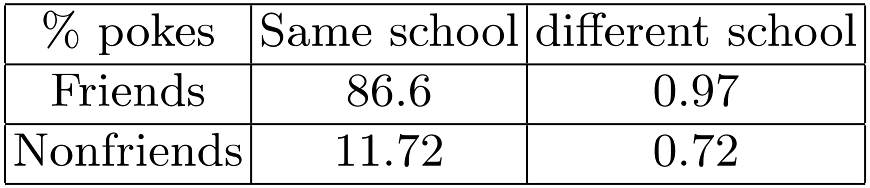
- Location: TSRB Ballroom
- *Either do Nov 19's reading reflections, or attend one of the above talks. If later, attendance will count for the grade toward Nov 19's reading reflections*

Rhythms of Social Interaction: Messaging Within a Massive Online Network

Summary

- The paper analyzed anonymized headers of 362 million messages shared by 4.2 million college students
- Paper set in college-era Facebook – 2004-06
- Main finding – many temporal regularities in communication; Facebook is clustered by colleges





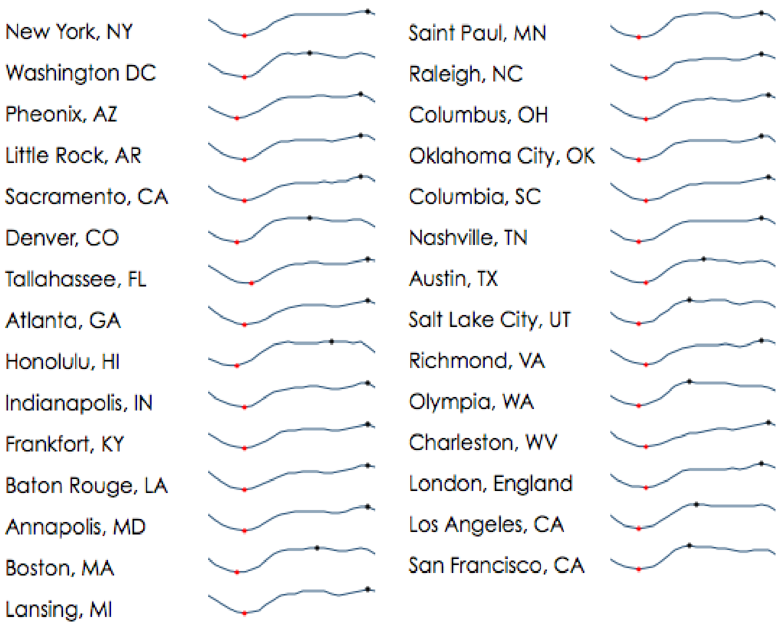
% pokes	Same school	different school
Friends	86.6	0.97
Nonfriends	11.72	0.72

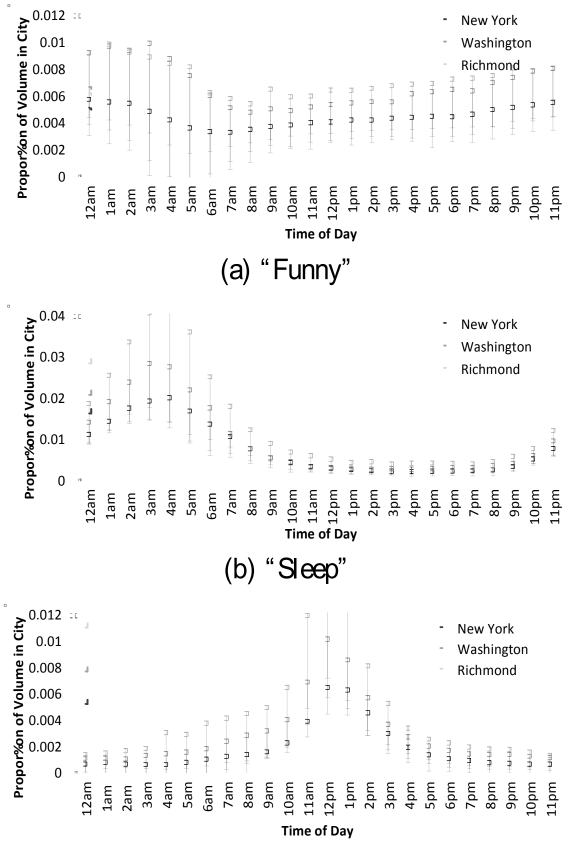
On the Study of Diurnal Urban Routines on Twitter

Summary

- The article analyses temporal patterns of activity in different geographic areas – primarily urban settings
  - examine within-day variability and across-day variability of diurnal keyword patterns for different locations
- Findings:
  - Only a few cities currently provide the magnitude of content needed to support across- day variability analysis for more than a few keywords
  - However within-day diurnal variability can help in comparing activities and finding similarities between cities

Summary

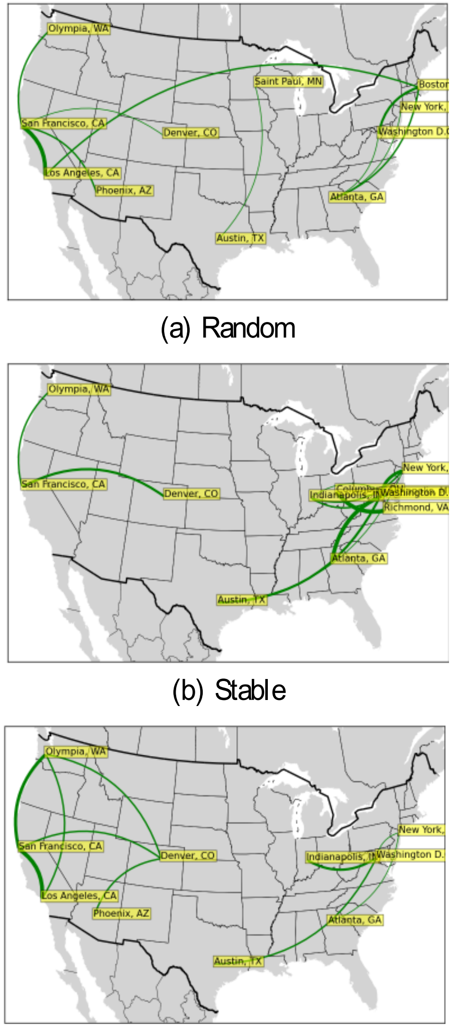




(a) "Funny"

(b) "Sleep"

(c) "Lunch"



(a) Random

(b) Stable

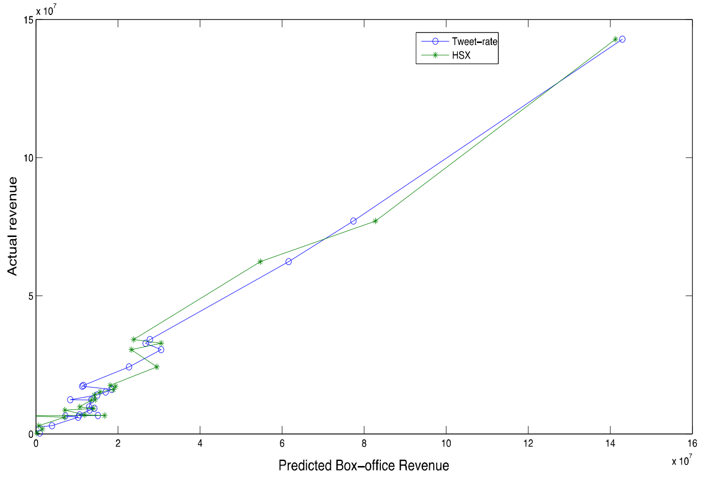
(c) Significant

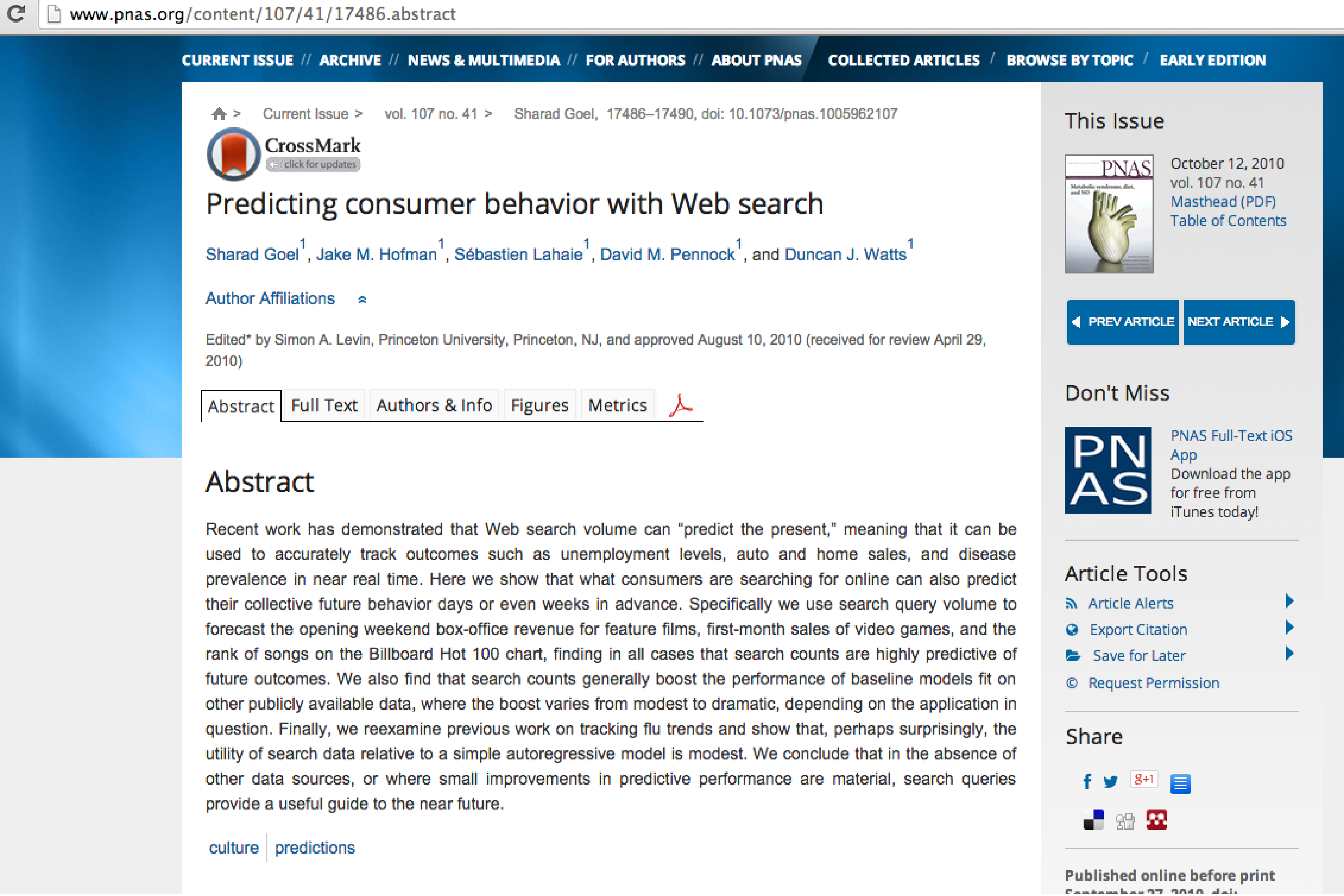
What is the value of studying diurnal urban keyword patterns in cities?

Predicting the Future With Social Media

Summary

- The article examines if social media can be used to predict real-world outcomes
- One of the earliest “predicting the present” papers using social media
- Method – simple tweet volume model outperformed market predictors of box office revenues
  - Compared with an HDX model – Hollywood Stock Exchange
- Potential shown how Twitter sentiment may be tapped




www.pnas.org/content/107/41/17486.abstract
CURRENT ISSUE // ARCHIVE // NEWS & MULTIMEDIA // FOR AUTHORS // ABOUT PNAS
COLLECTED ARTICLES / BROWSE BY TOPIC / EARLY EDITION
> Current Issue > vol. 107 no. 41 > Sharad Goel, 17486–17490, doi: 10.1073/pnas.1005962107
CrossMark
click for updates

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]

Author Affiliations

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

Abstract
Full Text
Authors & Info
Figures
Metrics

Abstract

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.

culture | predictions

This Issue

October 12, 2010
vol. 107 no. 41
Masthead (PDF)
Table of Contents

Don't Miss

PNAS Full-Text iOS App
Download the app for free from iTunes today!

Article Tools

Share

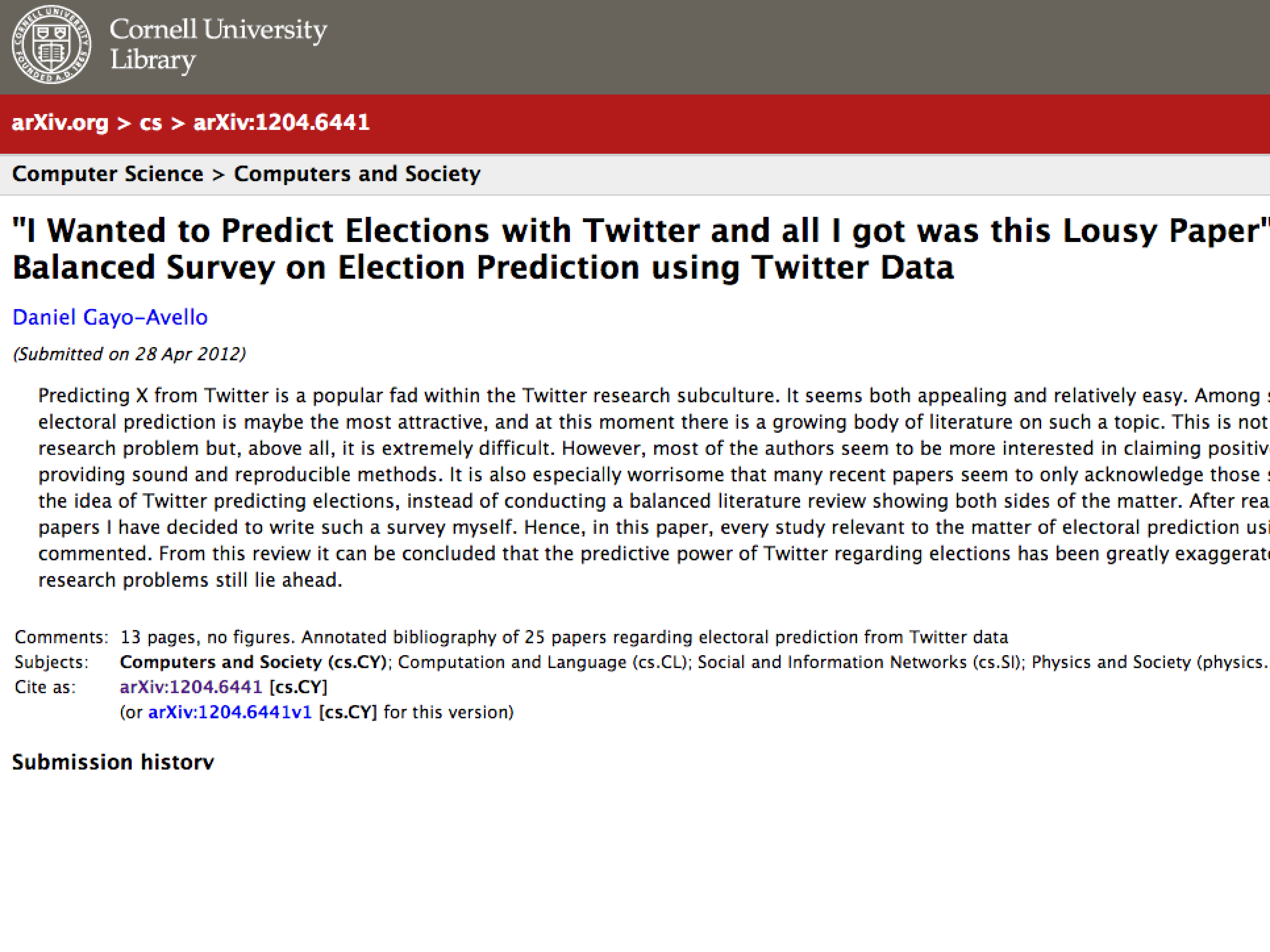
Published online before print
September 27, 2010, doi:

Golder et al.’s paper studied college-era Facebook. Do you think the kinds of temporal patterns of conversation they observed would hold on today’s Facebook? Is Facebook still clustered?

Other than “poke” what other forms of social interaction are supported by today’s SNSes to support remote social ties?

What all do you think you can predict with social media, specifically with Twitter, with Facebook, other?

What is the primary challenge of social media based predictions over traditional predictions?



arXiv.org > cs > arXiv:1204.6441

Computer Science > Computers and Society

"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" Balanced Survey on Election Prediction using Twitter Data

Daniel Gayo-Avello

(Submitted on 28 Apr 2012)

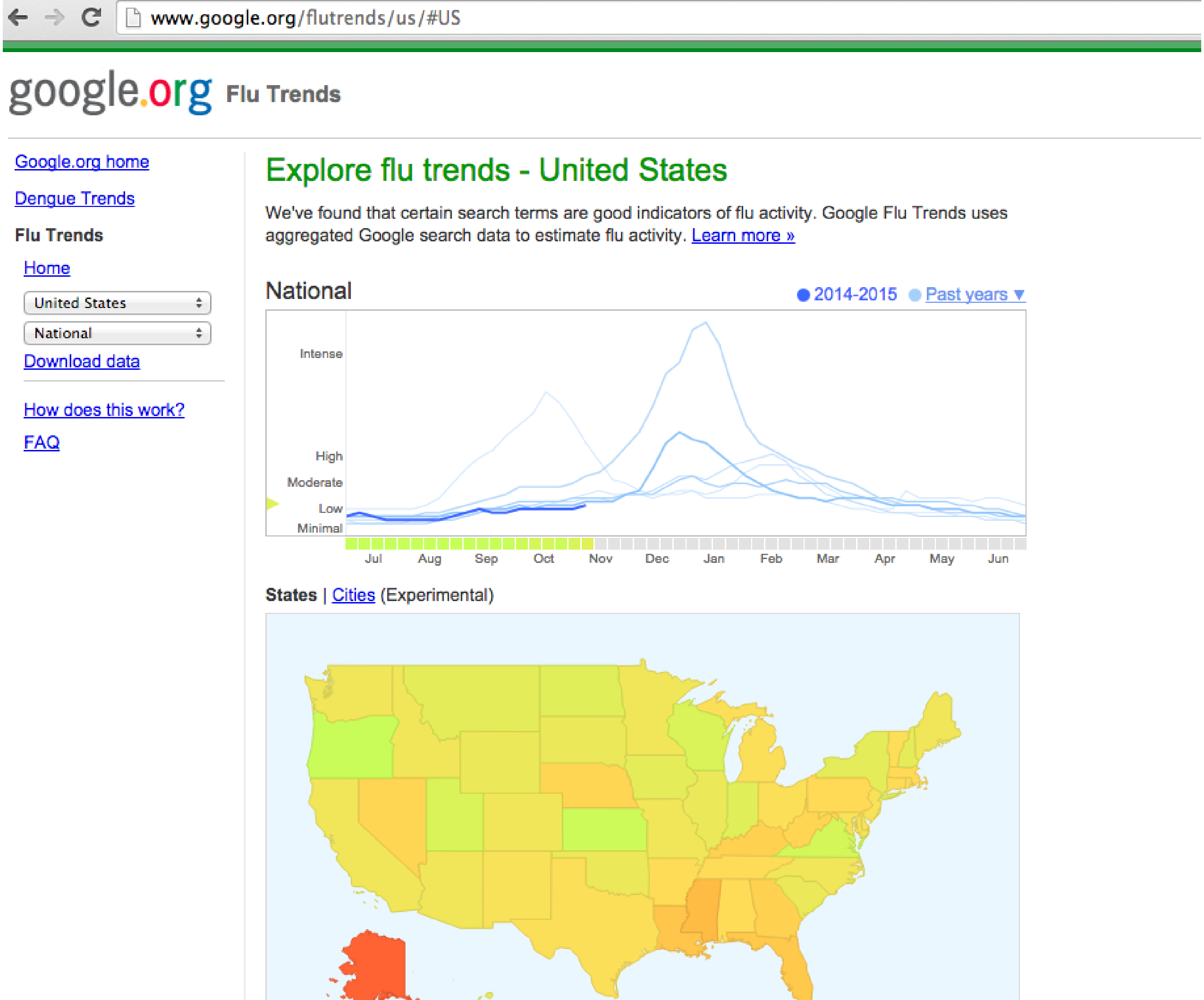
Predicting X from Twitter is a popular fad within the Twitter research subculture. It seems both appealing and relatively easy. Among s
electoral prediction is maybe the most attractive, and at this moment there is a growing body of literature on such a topic. This is not
research problem but, above all, it is extremely difficult. However, most of the authors seem to be more interested in claiming positiv
providing sound and reproducible methods. It is also especially worrisome that many recent papers seem to only acknowledge those s
the idea of Twitter predicting elections, instead of conducting a balanced literature review showing both sides of the matter. After rea
papers I have decided to write such a survey myself. Hence, in this paper, every study relevant to the matter of electoral prediction usi
commented. From this review it can be concluded that the predictive power of Twitter regarding elections has been greatly exaggerat
research problems still lie ahead.

Comments: 13 pages, no figures. Annotated bibliography of 25 papers regarding electoral prediction from Twitter data
Subjects: **Computers and Society (cs.CY)**; Computation and Language (cs.CL); Social and Information Networks (cs.SI); Physics and Society (physics.
Cite as: arXiv:1204.6441 [cs.CY]
(or arXiv:1204.6441v1 [cs.CY] for this version)

Submission historv

Why Social Media Can’t Predict Elections

- Post-hoc analysis, not *real* prediction
- Demographics not considered
- The people who tweet may not be the people who vote
- There’s no way to count votes on Twitter – even your neighbor’s dog has a Twitter profile
- Chance is not a valid baseline because incumbency tends to play a major role in most of the elections
- All the tweets are assumed to be trustworthy. That is, the presence of rumors, propaganda, misleading information, sarcasm, humor is ignored.
- Self-selection bias is simply ignored. People tweet on a voluntary basis and, therefore, data is produced only by those politically active.

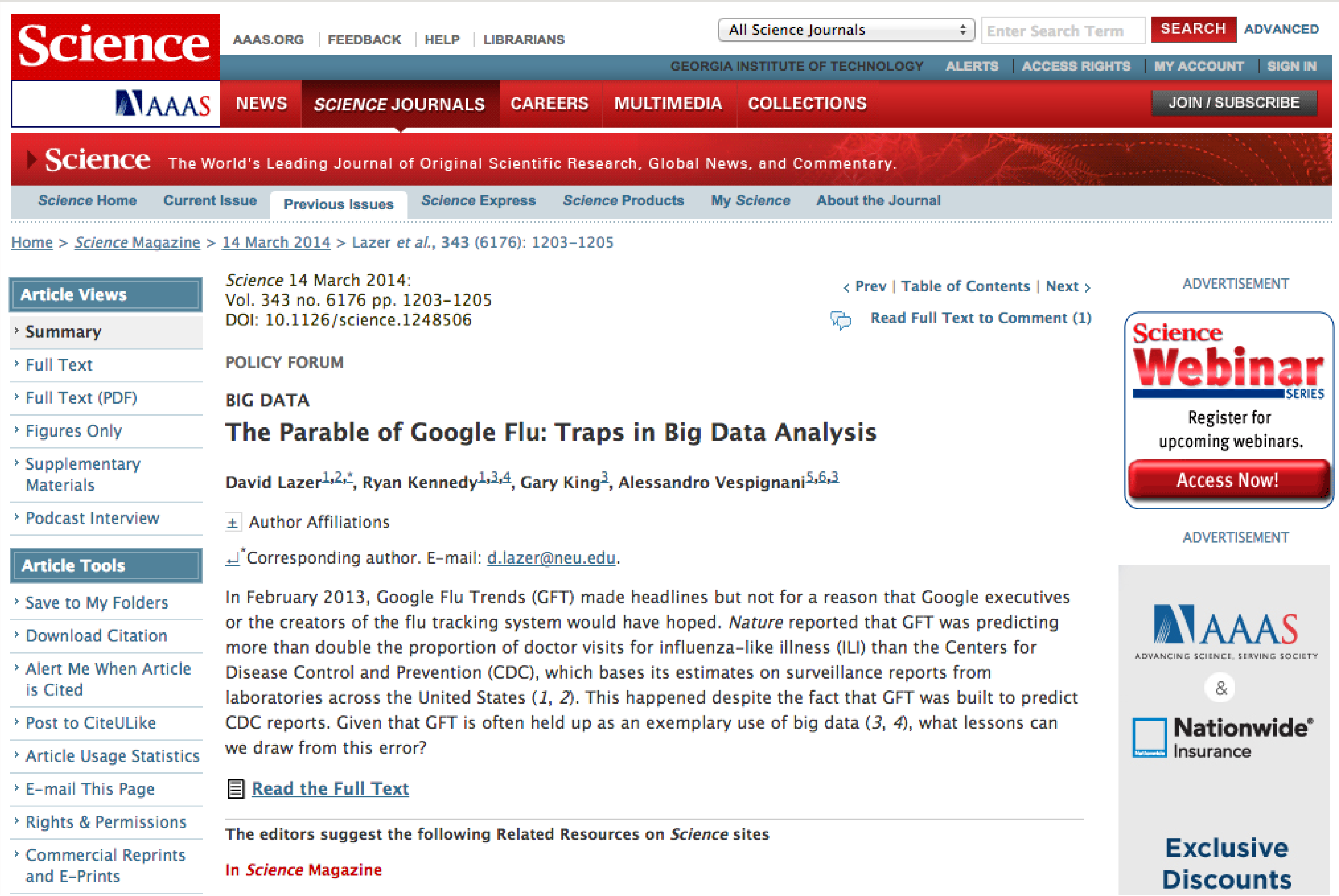


Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

States | Cities (Experimental)



Science 14 March 2014:
Vol. 343 no. 6176 pp. 1203-1205
DOI: 10.1126/science.1248506

POLICY FORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

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

± Author Affiliations

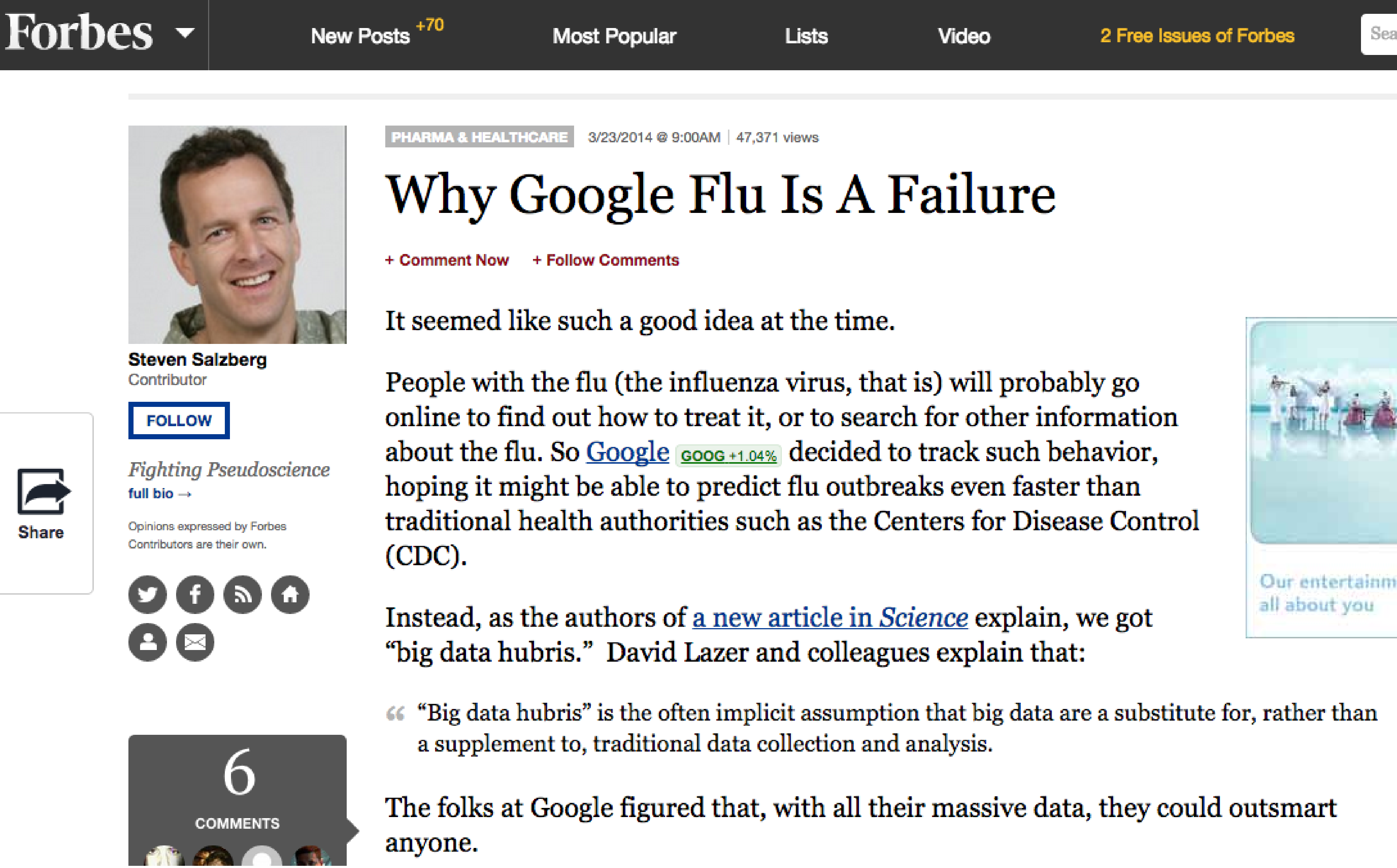
[*]Corresponding author. E-mail: d.lazer@neu.edu.

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?

Read the Full Text

The editors suggest the following Related Resources on *Science* sites

In *Science* Magazine



PHARMA & HEALTHCARE 3/23/2014 @ 9:00AM | 47,371 views

Why Google Flu Is A Failure

Steven Salzberg
Contributor

Fighting Pseudoscience
full bio →

Opinions expressed by Forbes Contributors are their own.

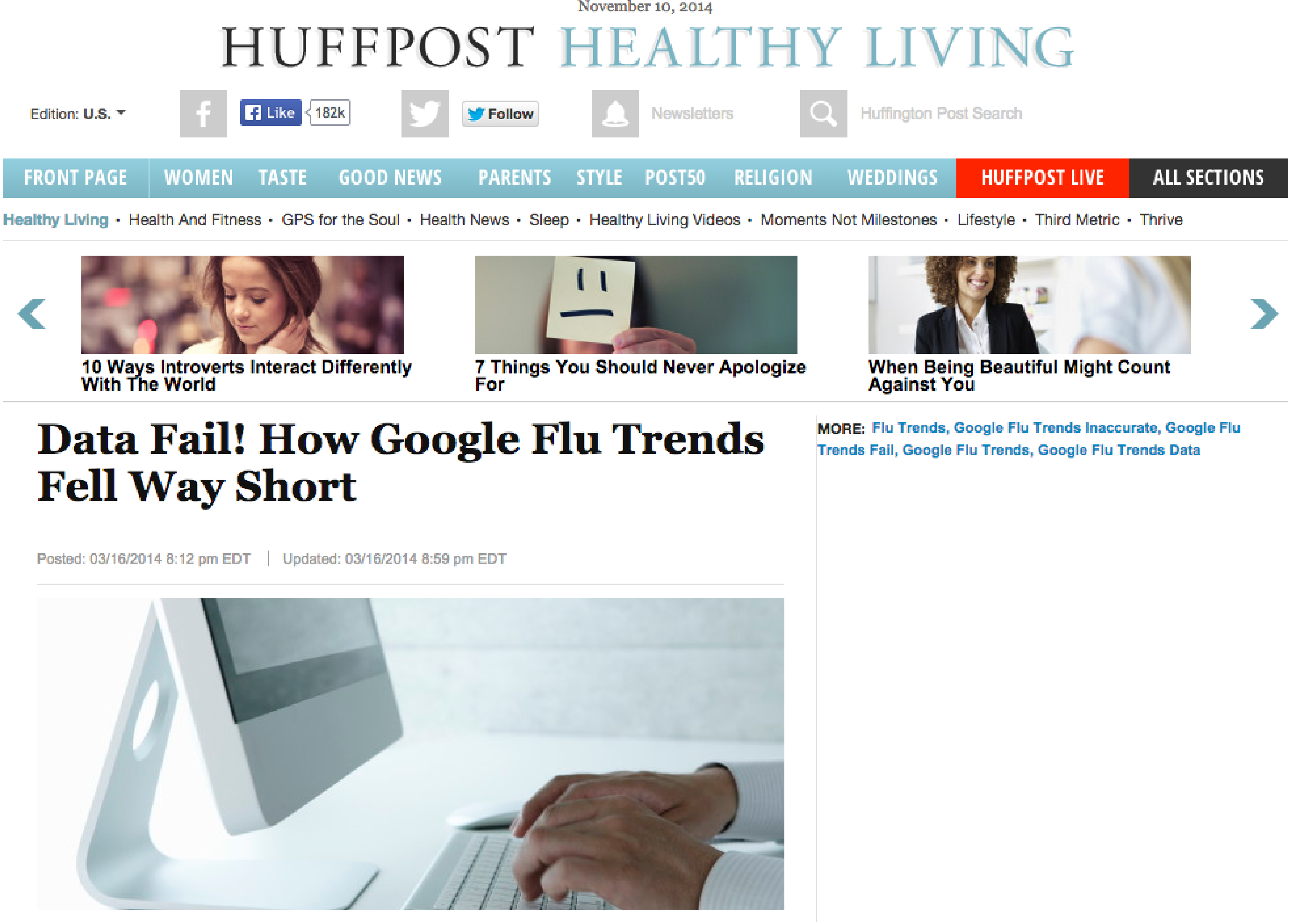
It seemed like such a good idea at the time.

People with the flu (the influenza virus, that is) will probably go online to find out how to treat it, or to search for other information about the flu. So Google GOOG +1.04% decided to track such behavior, hoping it might be able to predict flu outbreaks even faster than traditional health authorities such as the Centers for Disease Control (CDC).

Instead, as the authors of a new article in *Science* explain, we got "big data hubris." David Lazer and colleagues explain that:

> "Big data hubris" is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.

The folks at Google figured that, with all their massive data, they could outsmart anyone.



10 Ways Introverts Interact Differently With The World

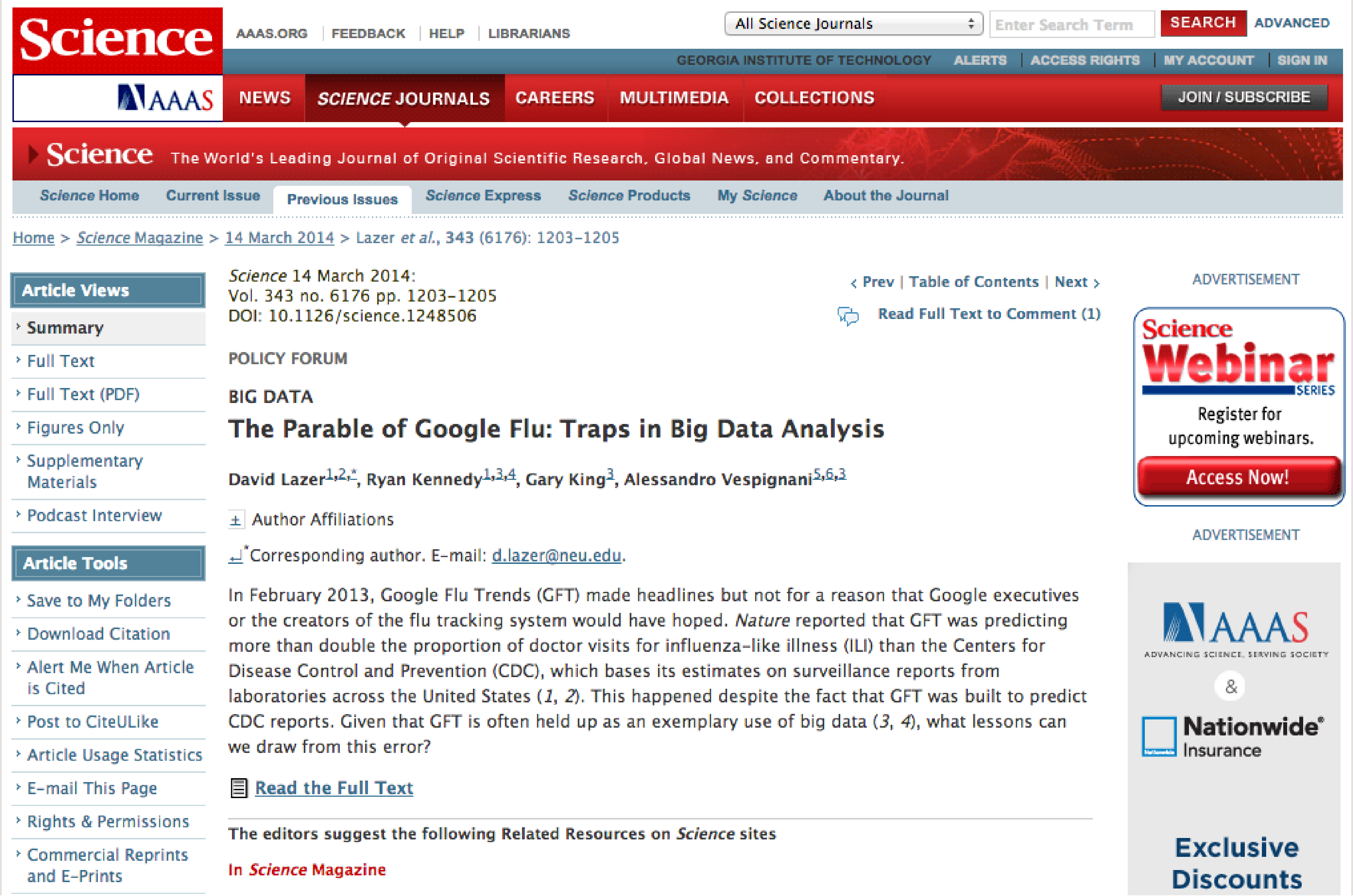
7 Things You Should Never Apologize For

When Being Beautiful Might Count Against You

Data Fail! How Google Flu Trends Fell Way Short

Posted: 03/16/2014 8:12 pm EDT | Updated: 03/16/2014 8:59 pm EDT

MORE: Flu Trends, Google Flu Trends Inaccurate, Google Flu Trends Fail, Google Flu Trends, Google Flu Trends Data



Science 14 March 2014:
Vol. 343 no. 6176 pp. 1203-1205
DOI: 10.1126/science.1248506

POLICY FORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

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

± Author Affiliations

*Corresponding author. E-mail: d.lazer@neu.edu.

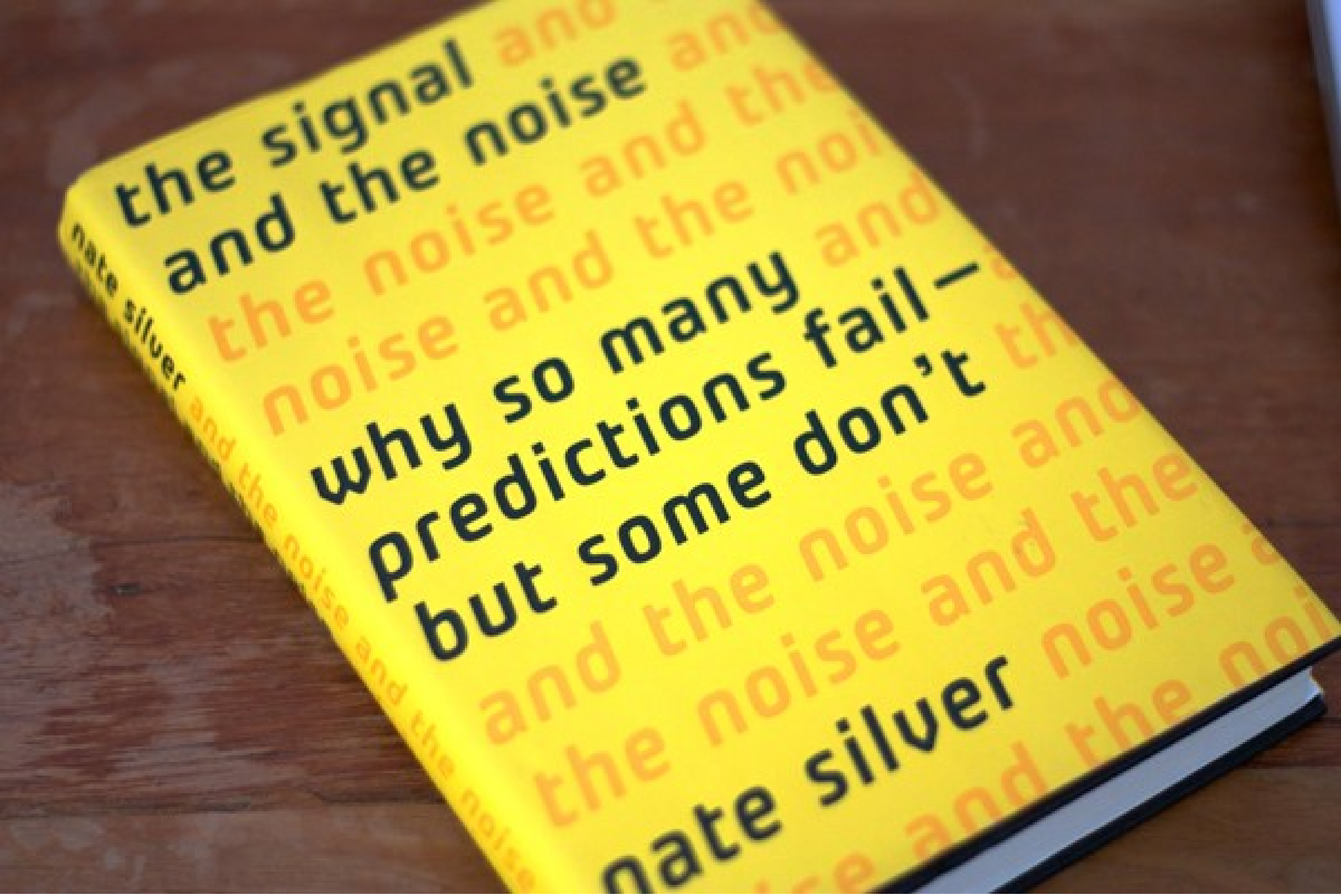
In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers 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 editors suggest the following Related Resources on *Science* sites

In *Science* Magazine

Combine multiple data “sensors”

The papers we read primarily use observational data for prediction. Note all focus on retrospective prediction. What are the problems with this approach? How to fix this problem?



Hypothesis testing, statistical significance, descriptive methods, experimental approach

Next class

- Wednesday 11/12
- Topic: “Event and News Analytics”
- Assigned readings due by 11:59 pm Tuesday