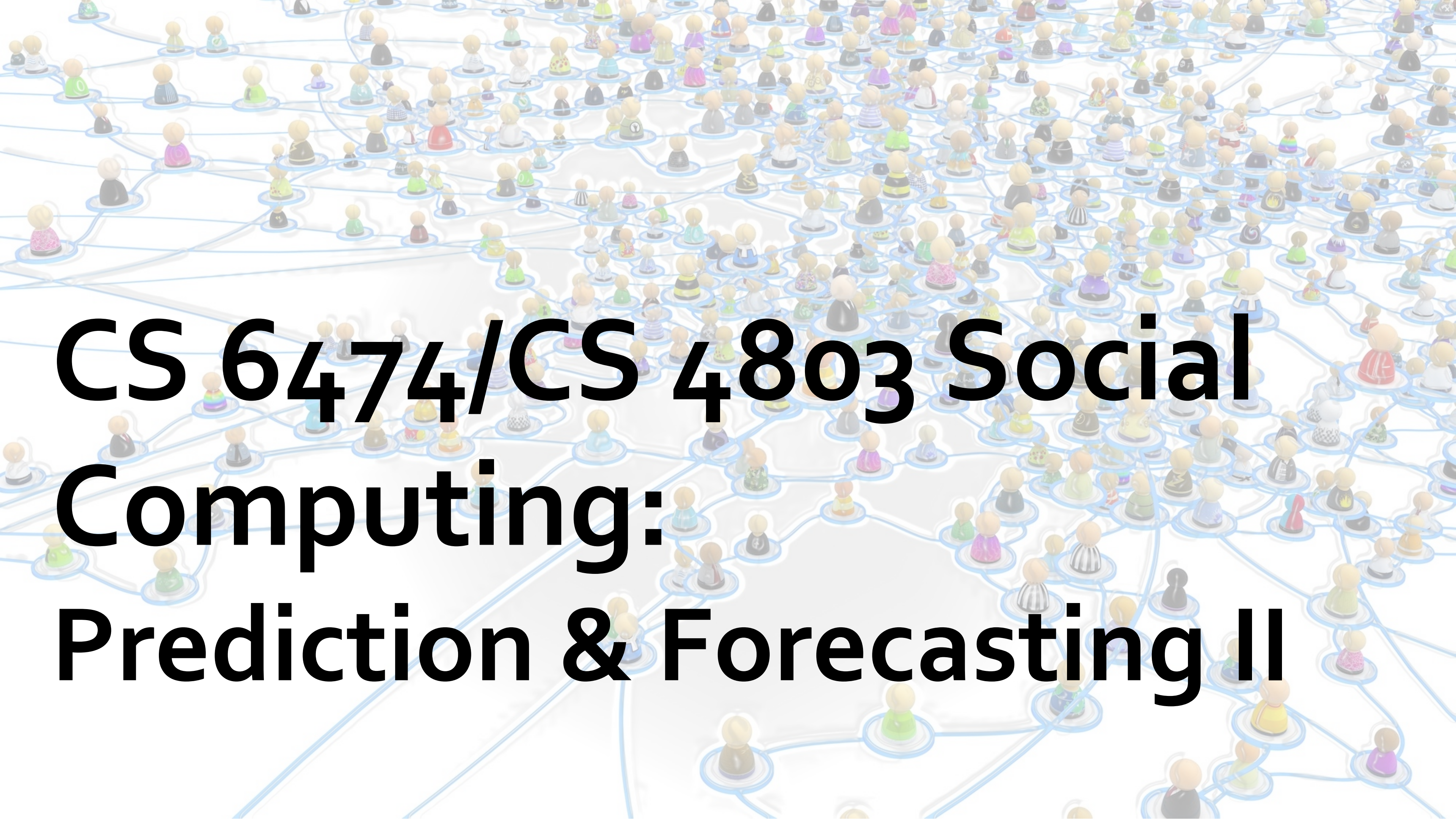


CS 6474/CS 4803 Social Computing: Prediction & Forecasting II

Munmun De Choudhury

munmund@gatech.edu

Week 13 | April 8, 2021

Final Presentations of Term Projects

- Scheduled for Apr 27
  - Signup document will be circulated soon
- Each team gets 15 minutes in all
  - 10-12 minutes of presentation
  - 3-5 minutes of Q&A
- Each team member needs to be present
- Structure:
  - Main idea
  - Background/Motivation
  - Research questions/Goals
  - Data/Social media platform
  - Method
  - Results
  - What you have learned

A checkered history…

Examples of many successes…

Incomplete history of cascade prediction

Who	Predicting	Features	Metric	Conclusion
HongD 10	Is item retweeted?	Topic Models	F1=0.47	Better than baseline
JendersKN 13	Will item reach some size *T*?	Content	F1>0.9	High accuracy
TanLP 14	Which of two does better?	Wording	Accu=65.6%	Computers are OK
ChengADKL 14	Will cascade double?	Temporal	AUC=0.88	Predictable
Lerman, Yang, Petrovic, Romero, Kupavskii, Ma, Weng, Zhao, Yu, etc				

Progress?

All of this work examines a different question with a different measure of success, evaluated on a different subset of data, making it difficult to assess overall progress[1]

[1] http://hunch.net/?p=22

Predicting success on Twitter?

Bakshy, Hofman, Mason, Watts (2011):

How viral will my tweet be?

“Cascades are unpredictable!”

Mason Porter @masonporter · Jan 19

I took a brief break from work. :)

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
We abhor vacuums. PAGE 181
2

Reasons behind the inconsistencies

Meaningless comparisons lead to false optimism in medical machine learning

Orianna DeMasi[1], Konrad Kording[2, 3], and Benjamin Recht[1]

[1]Department of Electrical Engineering and Computer Sciences, University of California Berkeley, Berkeley, CA, USA
[2]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA
[3]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA
odemasi@berkeley.edu, kording@upenn.edu, brecht@berkeley.edu

July 21, 2017

Abstract

A new trend in medicine is the use of algorithms to analyze big datasets, e.g. using everything your phone measures about you for diagnostics or monitoring. However, these algorithms are commonly compared against weak baselines, which may contribute to excessive optimism. To assess how well an algorithm works, scientists typically ask how well its output correlates with medically assigned scores. Here we perform a meta-analysis to quantify how the literature evaluates their algorithms for monitoring mental wellbeing. We find that the bulk of the literature ($\sim$77%) uses meaningless comparisons that ignore patient baseline state. For example, having an algorithm that uses phone data to diagnose mood disorders would be useful. However, it is possible to over 80% of the variance of some mood measures in the population by simply guessing that each patient has their own average mood - the patient-specific baseline. Thus, an algorithm that just predicts that our mood is like it usually is can explain the majority of variance, but is, obviously, entirely useless. Comparing to the wrong (population) baseline has a massive effect on the perceived quality of algorithms and produces baseless optimism in the field. To solve this problem we propose "user lift" that reduces these systematic errors in the evaluation of personalized medical monitoring.

Exploring limits to prediction in complex social systems

A unified framework: Luck vs. skill[2]

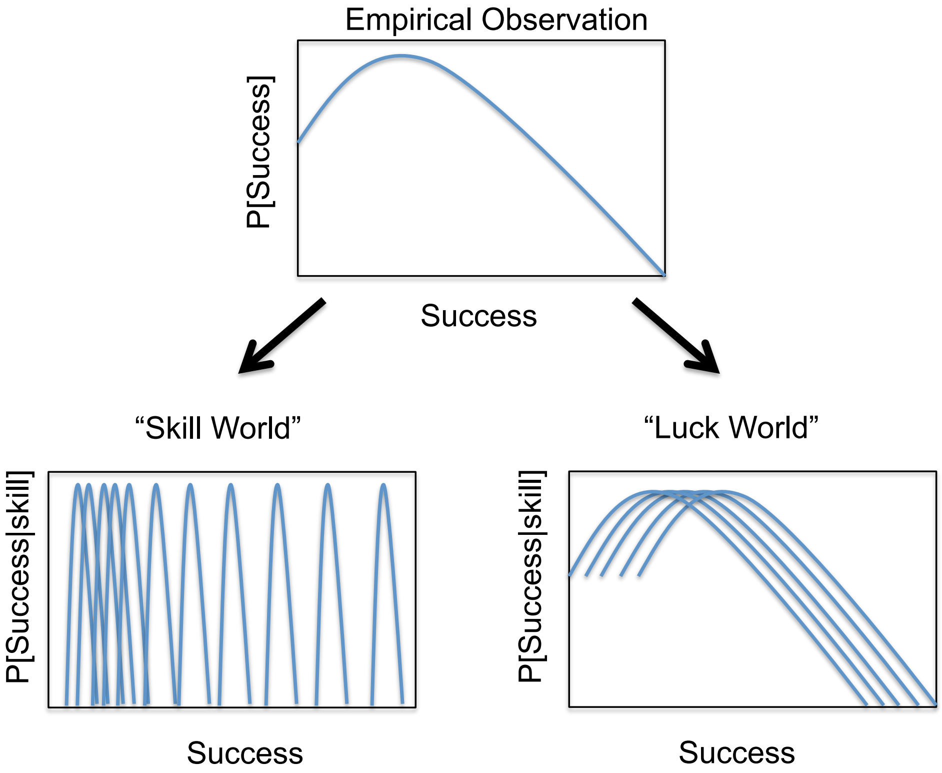
- Model success S as a mix of skill Q and luck ϵ:

$$S = f(Q) + \epsilon$$

- Measure the fraction of variance remaining after conditioning on skill:

$$F = \frac{\mathbb{E}[\mathrm{Var}(\mathrm{S}|Q)]}{\mathrm{Var}(\mathrm{S})} = 1 - R^2$$

- $R^2 = 1$ in a pure skill world, $R^2 = 0$ in pure luck world



[2] Formalizes Maboussin (2012)

Data

- Examined all 1.4B tweets containing URLs posted in February 2015
- Eliminated spam using internal Microsoft classifier
- Restricted attention to tweets containing URLs from the top 100 English-speaking domains with the most unique adopters
- Resulted in 850M tweets from 50M distinct users covering news, entertainment, videos, images, and products
- Measured the total cascade size for each seed tweet

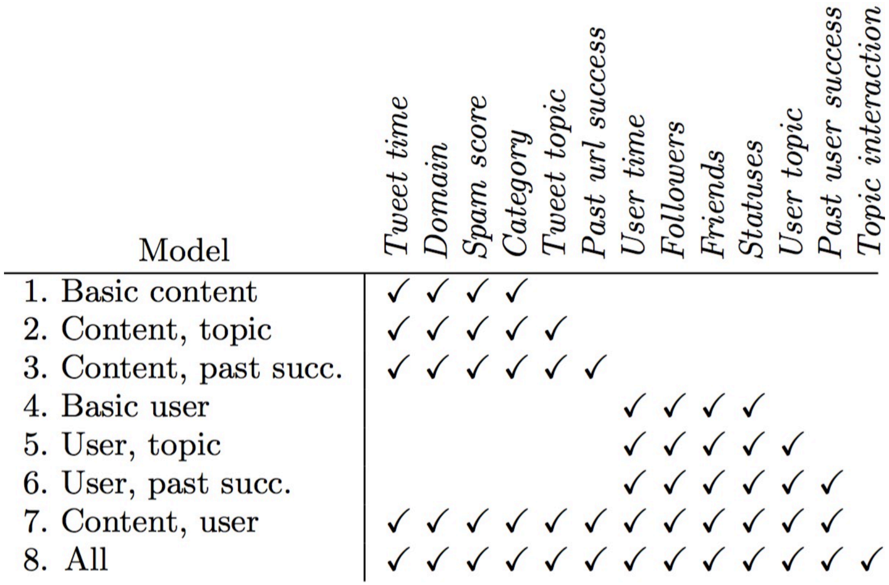
Predictive features

Used a random forest to estimate success (cascade size) given skill (available features)

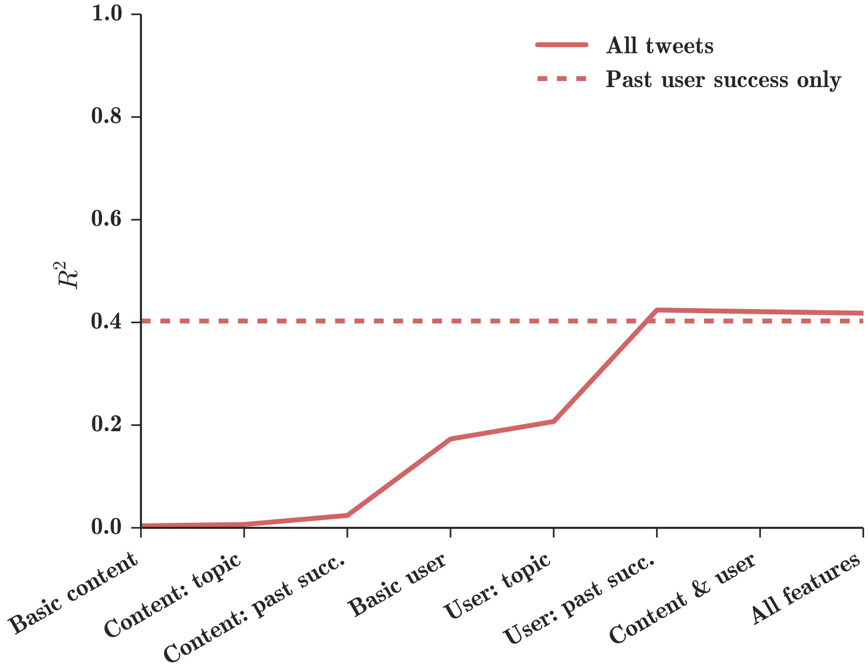
- Basic content features: URL domain, time of tweet, spam score, ODP category
- Basic user features: number of followers, number of friends, number of posts, account creation time
- Topic features: the most probable Latent Dirichlet Allocation topic for each user and tweet, along with an interaction term
- Past success: the average number of retweets received by each URL and user in the past

Predictive performance

best model explains roughly half of the variance in outcomes



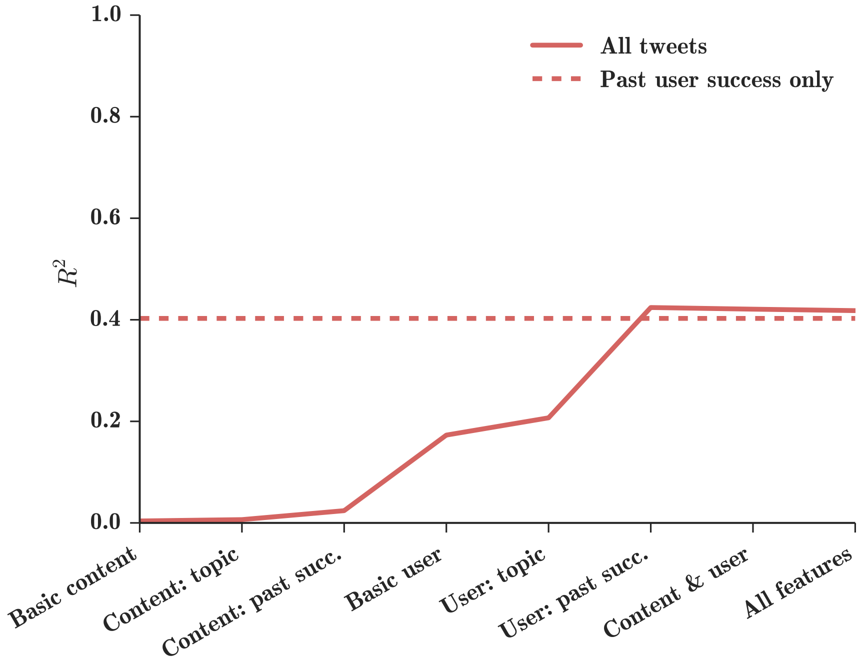
Model	*Tweet time*	*Domain*	*Spam score*	*Category*	*Tweet topic*	*Past url success*	*User time*	*Followers*	*Friends*	*Statuses*	*User topic*	*Past user success*	*Topic interaction*
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5. User, topic							✓	✓	✓	✓	✓		
6. User, past succ.							✓	✓	✓	✓	✓	✓	
7. Content, user	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
8. All	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓



Predictive performance

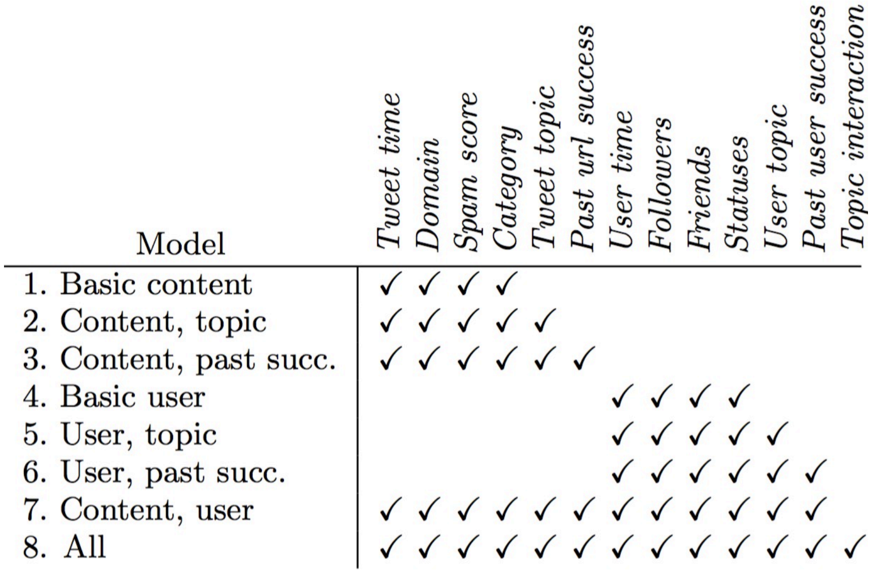
Content features alone perform poorly

Model	*Tweet time*	*Domain*	*Spam score*	*Category*	*Tweet topic*	*Past url success*	*User time*	*Followers*	*Friends*	*Statuses*	*User topic*	*Past user success*	*Topic interaction*
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5. User, topic							✓	✓	✓	✓	✓		
6. User, past succ.							✓	✓	✓	✓	✓	✓	
7. Content, user	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
8. All	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

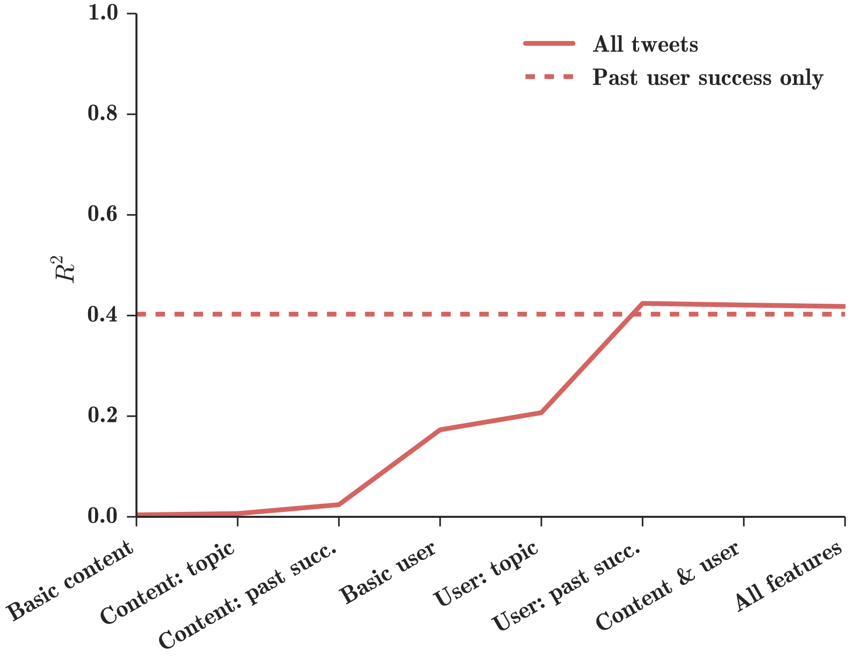


Predictive performance

Basic user features provide a reasonable boost in performance

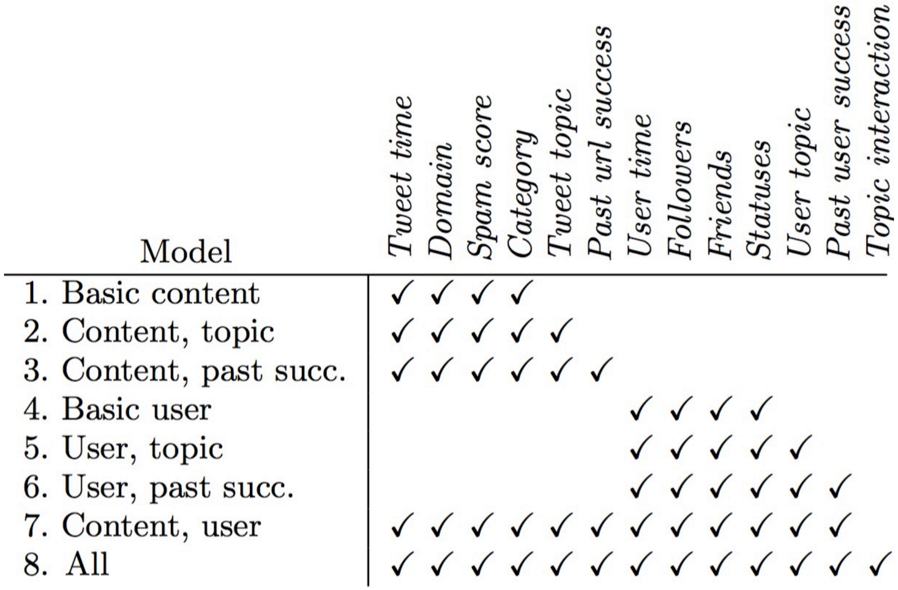


Model	*Tweet time*	*Domain*	*Spam score*	*Category*	*Tweet topic*	*Past url success*	*User time*	*Followers*	*Friends*	*Statuses*	*User topic*	*Past user success*	*Topic interaction*
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5. User, topic							✓	✓	✓	✓	✓		
6. User, past succ.							✓	✓	✓	✓	✓	✓	
7. Content, user	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
8. All	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

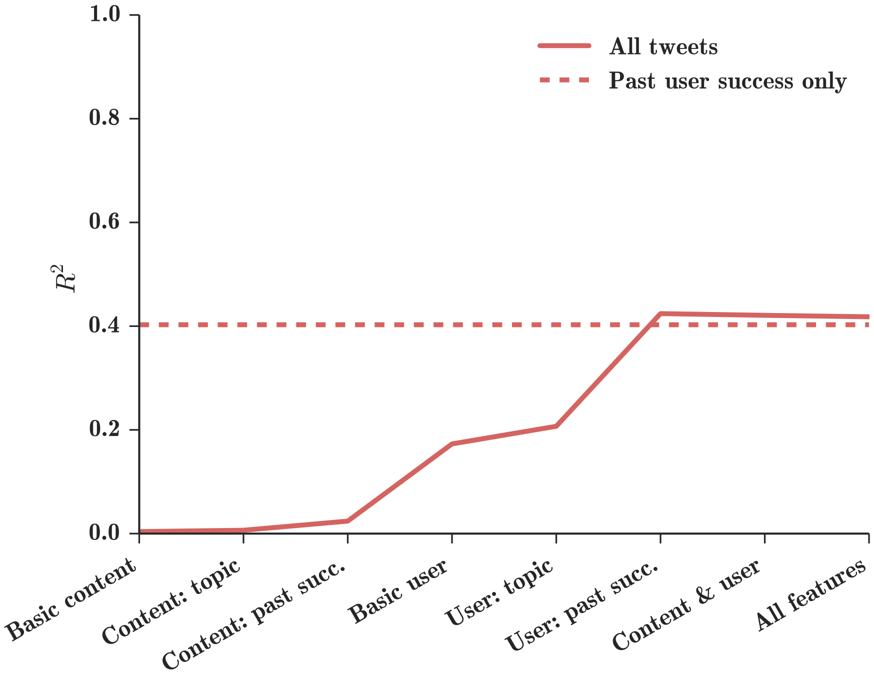


Predictive performance

Past user success alone accounts for almost all of predictive power



Model	*Tweet time*	*Domain*	*Spam score*	*Category*	*Tweet topic*	*Past url success*	*User time*	*Followers*	*Friends*	*Statuses*	*User topic*	*Past user success*	*Topic interaction*
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5. User, topic							✓	✓	✓	✓	✓		
6. User, past succ.							✓	✓	✓	✓	✓	✓	
7. Content, user	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
8. All	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓



- Both models derive their predictive power from the same simple feature: a user's past success

- Content features are only weakly informative

- Performance plateaus as we add more features, suggesting a possible limit to the predictability of diffusion outcomes

How can you *prove* a limit?

- Results robust to other ML models
  - Decision tree, linear regression
- Consistent with prior work
- Asymptote, dependency between features
- Can’t rule everything out
  - Simulation

Simulation

- SIR disease model
- Scale free network similar to Twitter
  - 7M users, α = 2.05
  - 8B simulated cascades
- *Quality*: R_0 = average neighbors infected
  - *p*(infect over edge) x *mean-degree*
- Prediction task
  - Given (possibly noisy) estimate of R_0 and the seed node, predict cascade size

Conclusions

Most things don't spread, but when they do, it's difficult to predict success

Conclusions

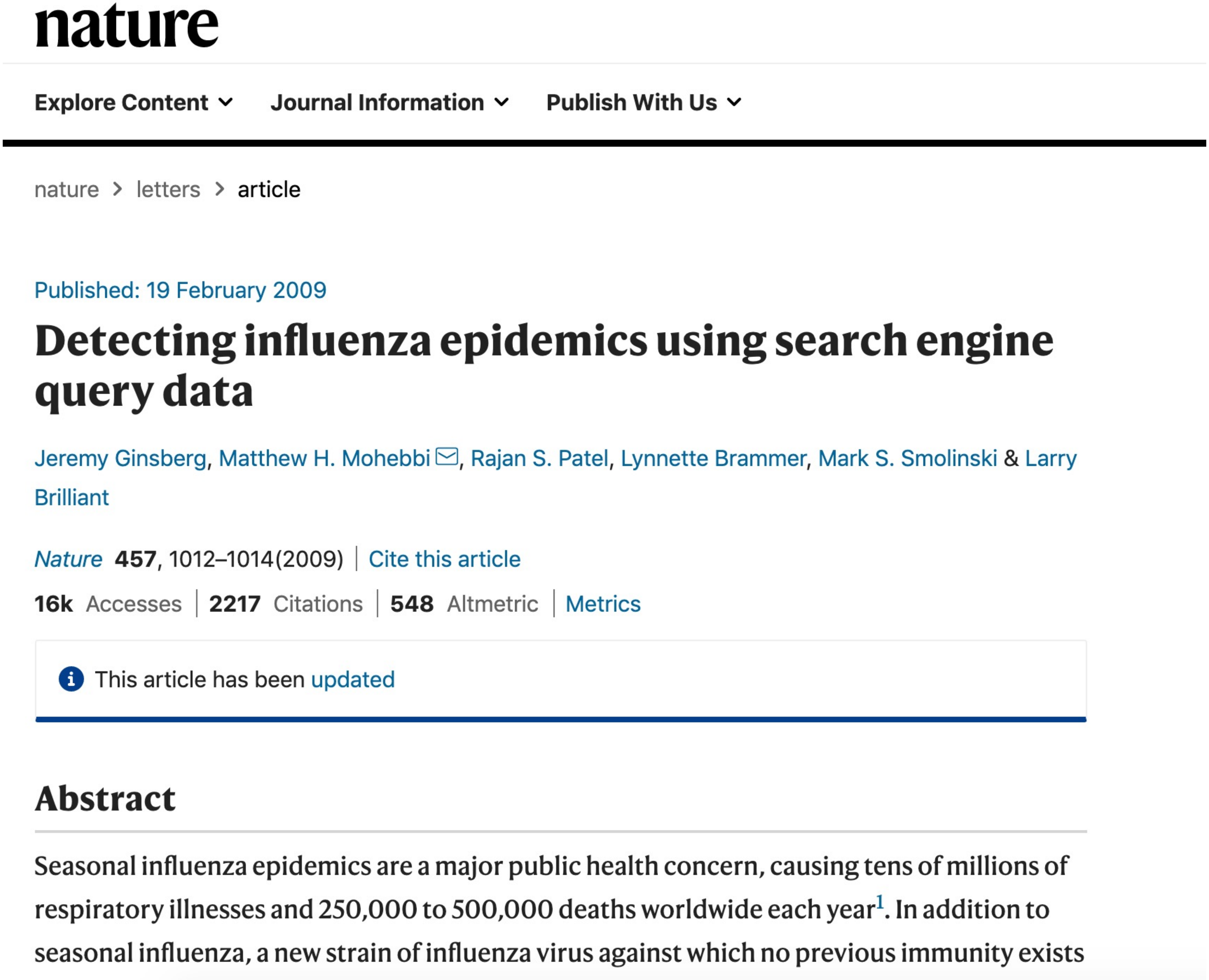
Despite a great deal of research on the topic, it's difficult to assess long-term progress in predicting success

Conclusions

State-of-the-art models explain roughly half of the variance in outcomes, based primarily on past success

Conclusions

This is likely due to randomness in diffusion process itself, rather than our ability to estimate or model it



nature

Published: 19 February 2009

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg, Matthew H. Mohebbi ✉, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski & Larry Brilliant

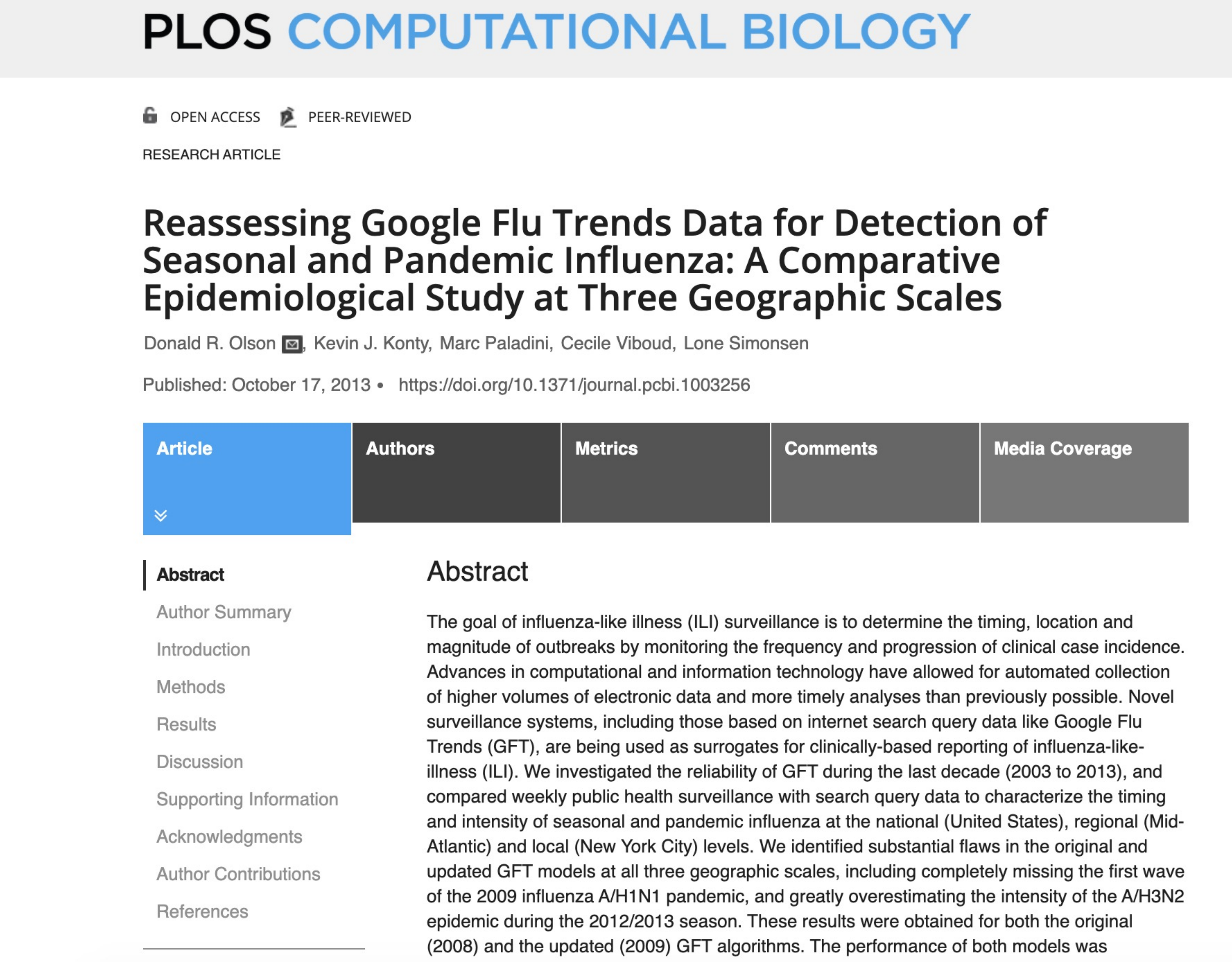
Nature **457**, 1012–1014(2009) | Cite this article

16k Accesses | **2217** Citations | **548** Altmetric | Metrics

ℹ This article has been updated

Abstract

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year[1]. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists



PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

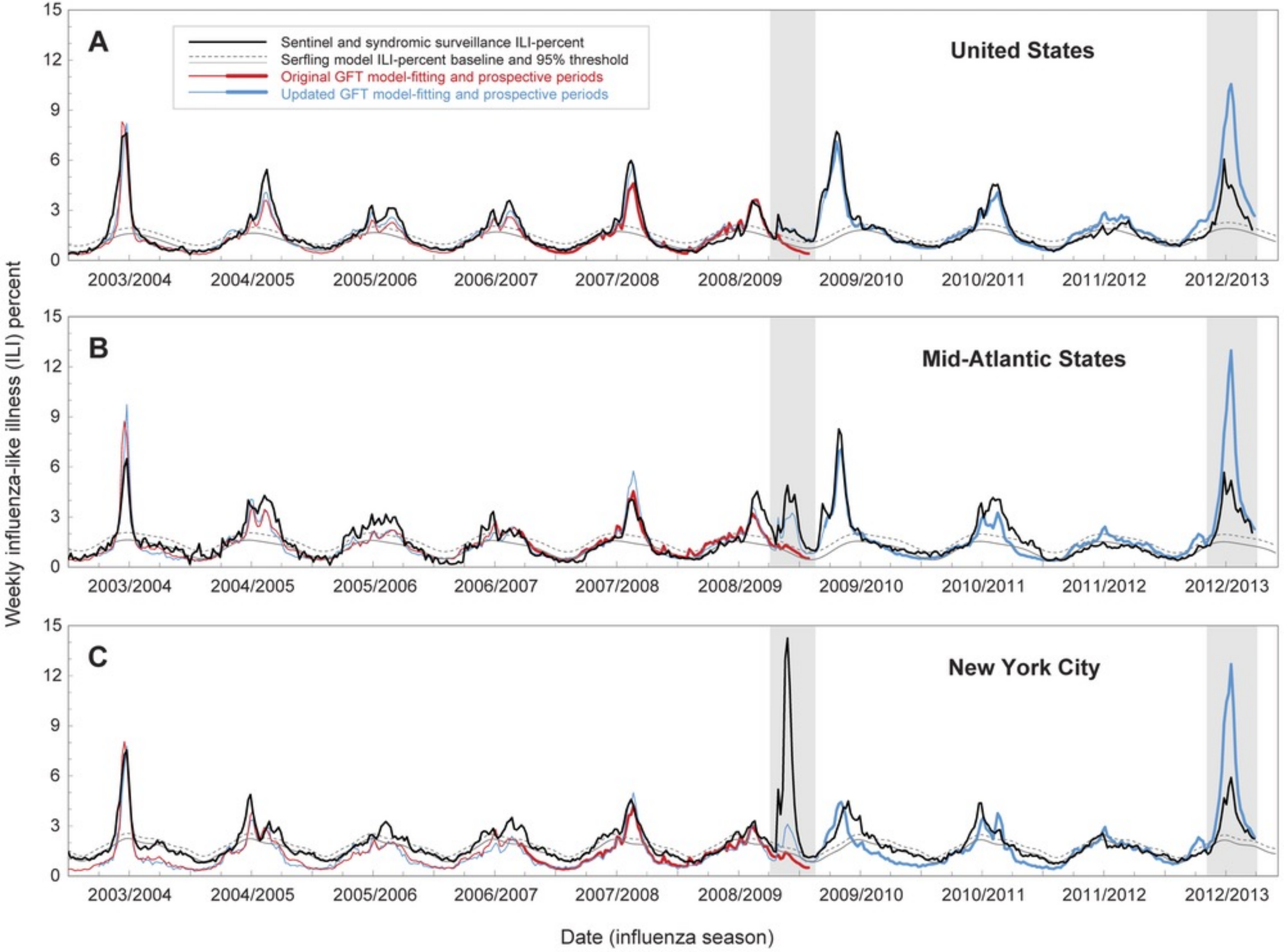
Reassessing Google Flu Trends Data for Detection of Seasonal and Pandemic Influenza: A Comparative Epidemiological Study at Three Geographic Scales

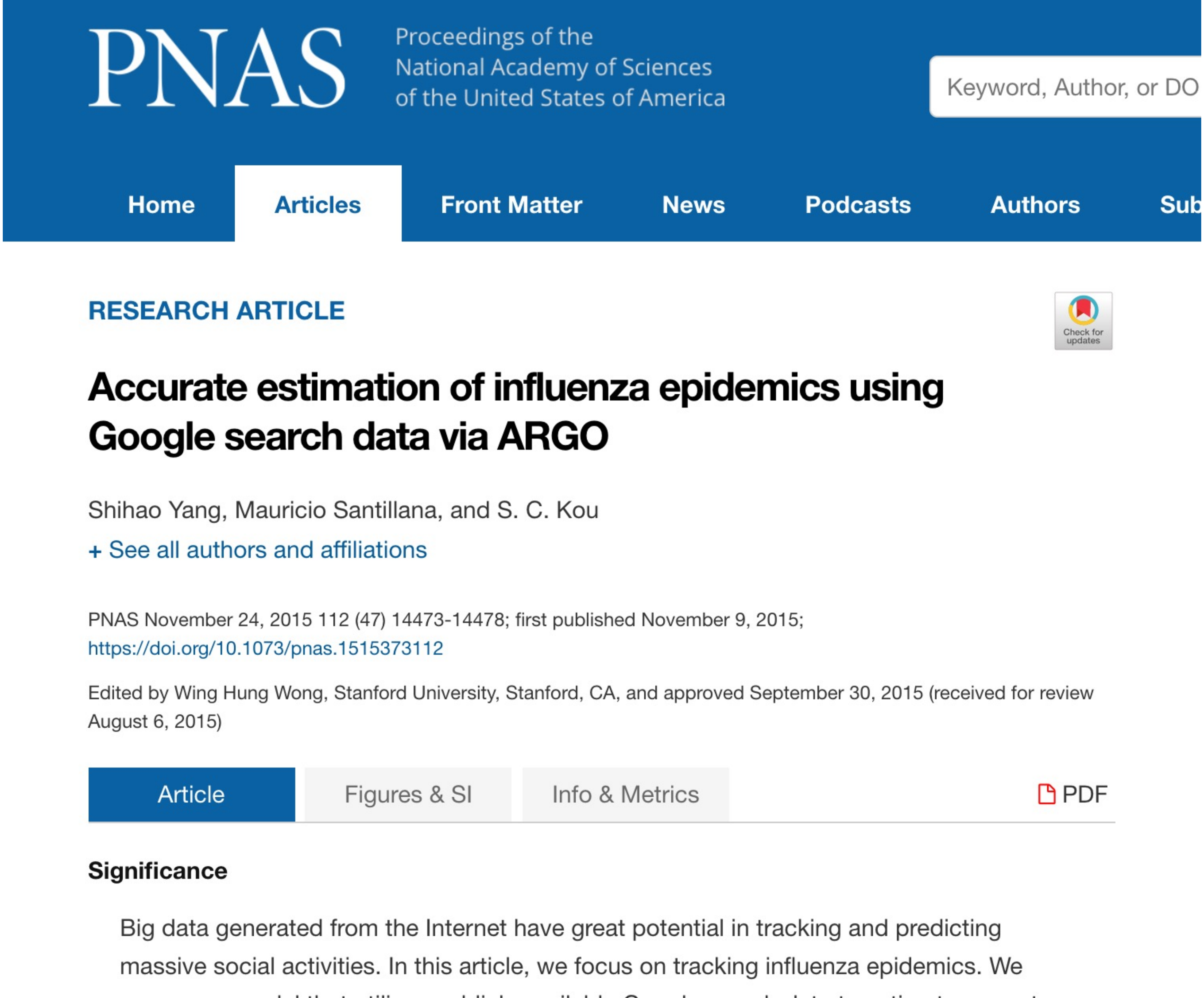
Donald R. Olson, Kevin J. Konty, Marc Paladini, Cecile Viboud, Lone Simonsen

Published: October 17, 2013 • https://doi.org/10.1371/journal.pcbi.1003256

Abstract

The goal of influenza-like illness (ILI) surveillance is to determine the timing, location and magnitude of outbreaks by monitoring the frequency and progression of clinical case incidence. Advances in computational and information technology have allowed for automated collection of higher volumes of electronic data and more timely analyses than previously possible. Novel surveillance systems, including those based on internet search query data like Google Flu Trends (GFT), are being used as surrogates for clinically-based reporting of influenza-like-illness (ILI). We investigated the reliability of GFT during the last decade (2003 to 2013), and compared weekly public health surveillance with search query data to characterize the timing and intensity of seasonal and pandemic influenza at the national (United States), regional (Mid-Atlantic) and local (New York City) levels. We identified substantial flaws in the original and updated GFT models at all three geographic scales, including completely missing the first wave of the 2009 influenza A/H1N1 pandemic, and greatly overestimating the intensity of the A/H3N2 epidemic during the 2012/2013 season. These results were obtained for both the original (2008) and the updated (2009) GFT algorithms. The performance of both models was


A
B
C
United States
Mid-Atlantic States
New York City
Sentinel and syndromic surveillance ILI-percent
Serfling model ILI-percent baseline and 95% threshold
Original GFT model-fitting and prospective periods
Updated GFT model-fitting and prospective periods
Weekly influenza-like illness (ILI) percent
Date (influenza season)
0
3
6
9
12
15
2003/2004
2004/2005
2005/2006
2006/2007
2007/2008
2008/2009
2009/2010
2010/2011
2011/2012
2012/2013



RESEARCH ARTICLE

Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang, Mauricio Santillana, and S. C. Kou

+ See all authors and affiliations

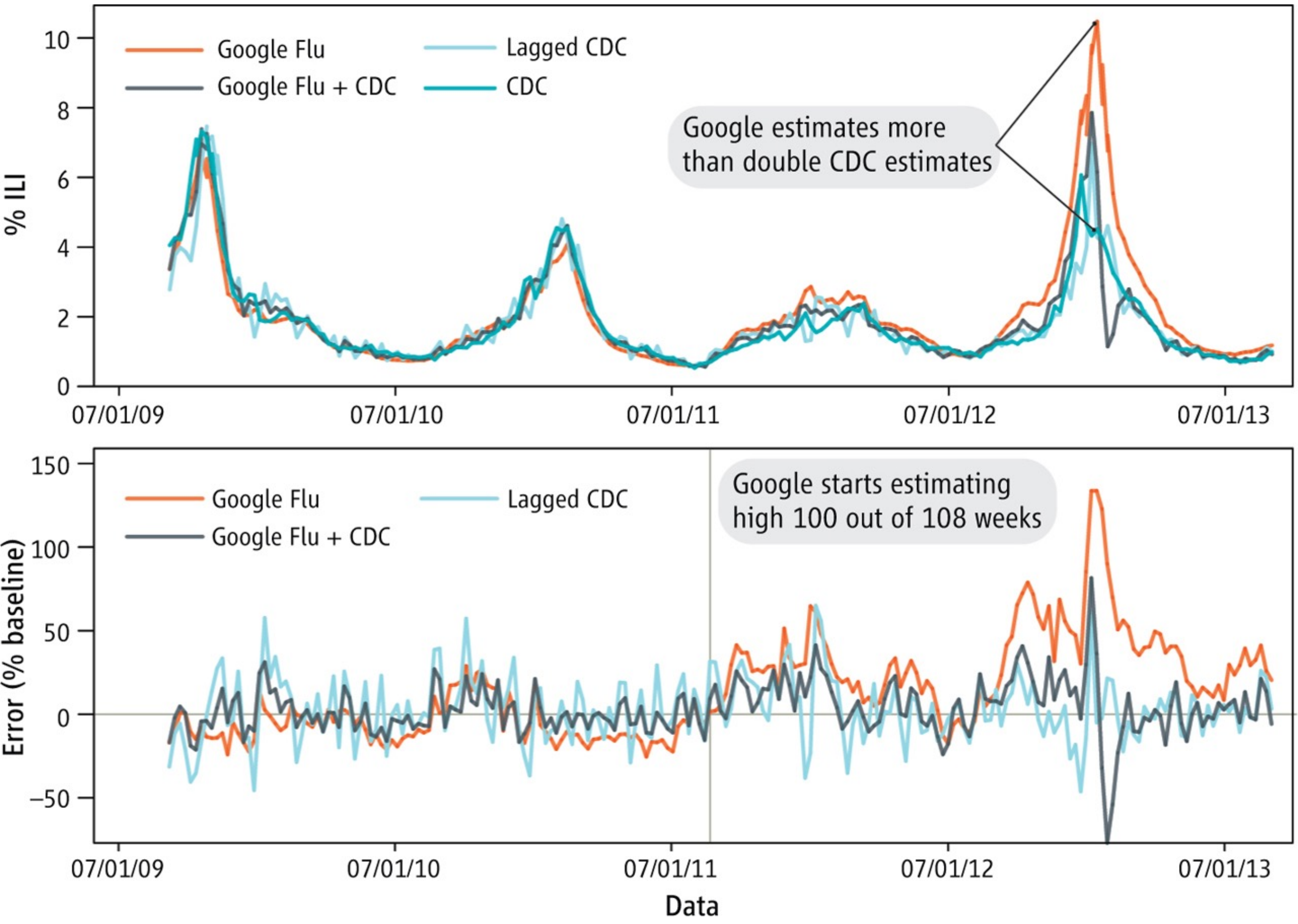
PNAS November 24, 2015 112 (47) 14473-14478; first published November 9, 2015; https://doi.org/10.1073/pnas.1515373112

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

Significance

Big data generated from the Internet have great potential in tracking and predicting massive social activities. In this article, we focus on tracking influenza epidemics. We

The parable of google flu


Google Flu
Lagged CDC
Google Flu + CDC
CDC
Google estimates more than double CDC estimates
% ILI
0
2
4
6
8
10
07/01/09
07/01/10
07/01/11
07/01/12
07/01/13
Google Flu
Lagged CDC
Google Flu + CDC
Google starts estimating high 100 out of 108 weeks
Error (% baseline)
150
100
50
0
−50
07/01/09
07/01/10
07/01/11
07/01/12
07/01/13
Data

Big data hubris

Algorithmic Dynamics

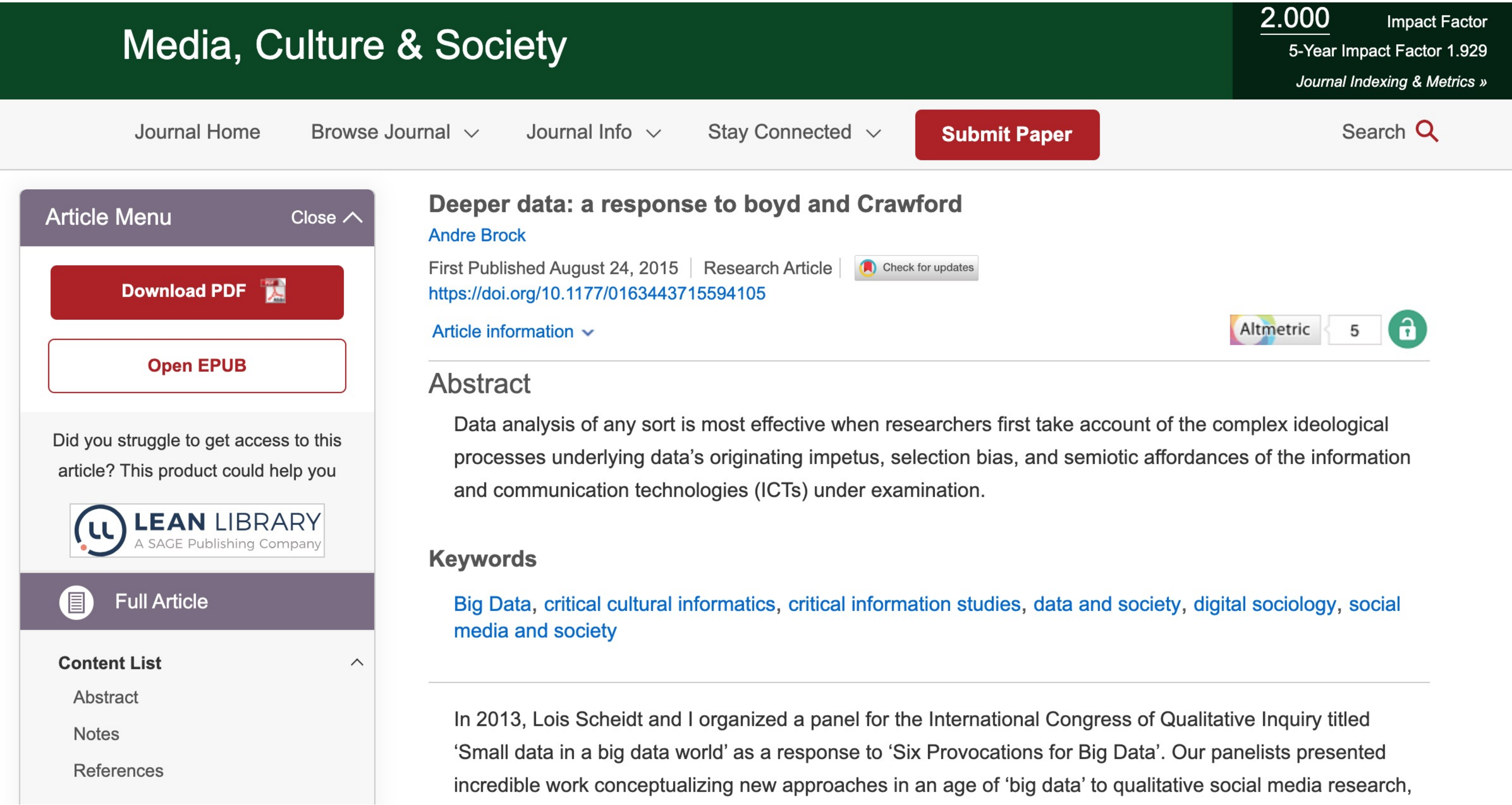
It's Not Just About Size of the Data

danah boyd & Kate Crawford

CRITICAL QUESTIONS FOR BIG DATA

Provocations for a cultural, technological, and scholarly phenomenon

The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions. Diverse groups argue about the potential benefits and costs of analyzing genetic sequences, social media interactions, health records, phone logs, government records, and other digital traces left by people. Significant questions emerge. Will large-scale search data help us create better tools, services, and public goods? Or will it usher in a new wave of privacy incursions and invasive marketing? Will data analytics help us understand online communities and political movements? Or will it be used to track protesters and suppress speech? Will it transform how we study human communi-



Media, Culture & Society

2.000 Impact Factor
5-Year Impact Factor 1.929
Journal Indexing & Metrics »

Deeper data: a response to boyd and Crawford

Andre Brock

First Published August 24, 2015 | Research Article

https://doi.org/10.1177/0163443715594105

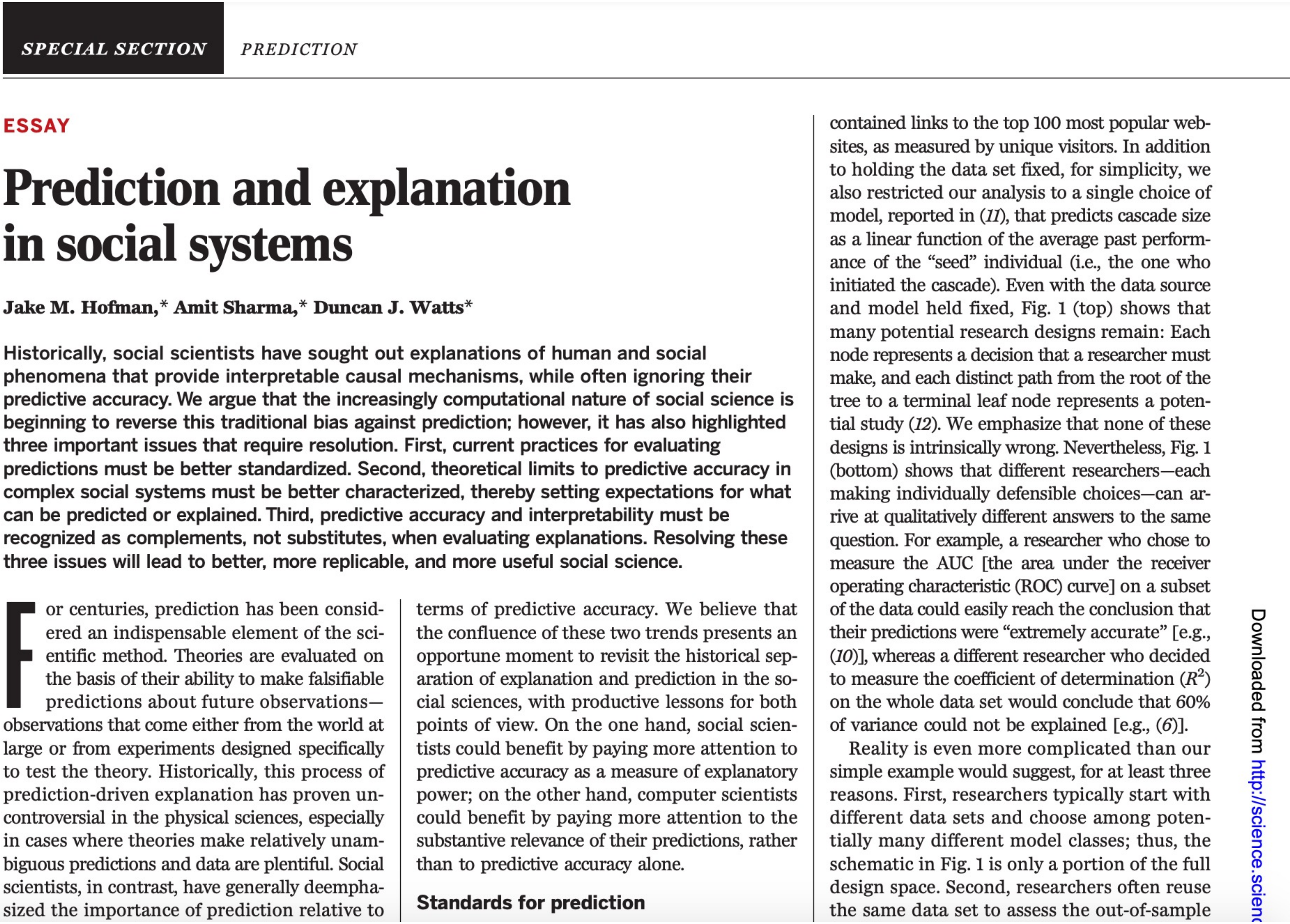
Abstract

Data analysis of any sort is most effective when researchers first take account of the complex ideological processes underlying data's originating impetus, selection bias, and semiotic affordances of the information and communication technologies (ICTs) under examination.

Keywords

Big Data, critical cultural informatics, critical information studies, data and society, digital sociology, social media and society

In 2013, Lois Scheidt and I organized a panel for the International Congress of Qualitative Inquiry titled 'Small data in a big data world' as a response to 'Six Provocations for Big Data'. Our panelists presented incredible work conceptualizing new approaches in an age of 'big data' to qualitative social media research,



ESSAY

Prediction and explanation in social systems

Jake M. Hofman,* Amit Sharma,* Duncan J. Watts*

Historically, social scientists have sought out explanations of human and social phenomena that provide interpretable causal mechanisms, while often ignoring their predictive accuracy. We argue that the increasingly computational nature of social science is beginning to reverse this traditional bias against prediction; however, it has also highlighted three important issues that require resolution. First, current practices for evaluating predictions must be better standardized. Second, theoretical limits to predictive accuracy in complex social systems must be better characterized, thereby setting expectations for what can be predicted or explained. Third, predictive accuracy and interpretability must be recognized as complements, not substitutes, when evaluating explanations. Resolving these three issues will lead to better, more replicable, and more useful social science.

For centuries, prediction has been considered an indispensable element of the scientific method. Theories are evaluated on the basis of their ability to make falsifiable predictions about future observations—observations that come either from the world at large or from experiments designed specifically to test the theory. Historically, this process of prediction-driven explanation has proven uncontroversial in the physical sciences, especially in cases where theories make relatively unambiguous predictions and data are plentiful. Social scientists, in contrast, have generally deemphasized the importance of prediction relative to terms of predictive accuracy. We believe that the confluence of these two trends presents an opportune moment to revisit the historical separation of explanation and prediction in the social sciences, with productive lessons for both points of view. On the one hand, social scientists could benefit by paying more attention to predictive accuracy as a measure of explanatory power; on the other hand, computer scientists could benefit by paying more attention to the substantive relevance of their predictions, rather than to predictive accuracy alone.

Standards for prediction

contained links to the top 100 most popular websites, as measured by unique visitors. In addition to holding the data set fixed, for simplicity, we also restricted our analysis to a single choice of model, reported in (11), that predicts cascade size as a linear function of the average past performance of the "seed" individual (i.e., the one who initiated the cascade). Even with the data source and model held fixed, Fig. 1 (top) shows that many potential research designs remain: Each node represents a decision that a researcher must make, and each distinct path from the root of the tree to a terminal leaf node represents a potential study (12). We emphasize that none of these designs is intrinsically wrong. Nevertheless, Fig. 1 (bottom) shows that different researchers—each making individually defensible choices—can arrive at qualitatively different answers to the same question. For example, a researcher who chose to measure the AUC [the area under the receiver operating characteristic (ROC) curve] on a subset of the data could easily reach the conclusion that their predictions were "extremely accurate" [e.g., (10)], whereas a different researcher who decided to measure the coefficient of determination (R^2) on the whole data set would conclude that 60% of variance could not be explained [e.g., (6)].

Reality is even more complicated than our simple example would suggest, for at least three reasons. First, researchers typically start with different data sets and choose among potentially many different model classes; thus, the schematic in Fig. 1 is only a portion of the full design space. Second, researchers often reuse the same data set to assess the out-of-sample

Class Discussion

Assess whether in each of the following cases interpretation or prediction (or both) is/are preferred.

i) How people react on a new product release (e.g., an iphone), as observed on social media
ii) Whether greater anonymity leads to greater hate speech on social media
iii) Whether following recommended videos on YouTube leads down a more politically extreme rabbit hole
iv) Whether deplatforming reduces misinformation on social media