CS 6474/CS 4803 Social Computing: Methodological Pitfalls

Munmun De Choudhury

munmund@gatech.edu

Week 13 | April 3, 2024

Examples of many successes...

"I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" A 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 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 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.

Submission history

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!"

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
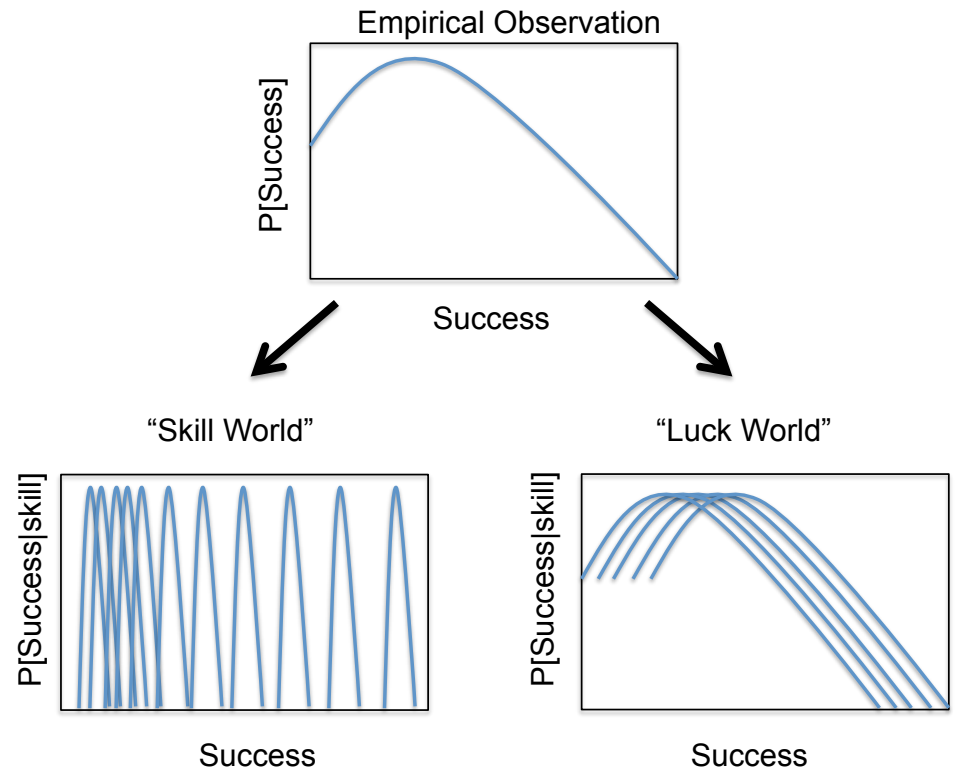
A unified framework: Luck vs. skill[2]

- 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:

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

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

Empirical Observation

P[Success]

Success

"Skill World"

"Luck World"

P[Success|skill]

Success

P[Success|skill]

Success

---

[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

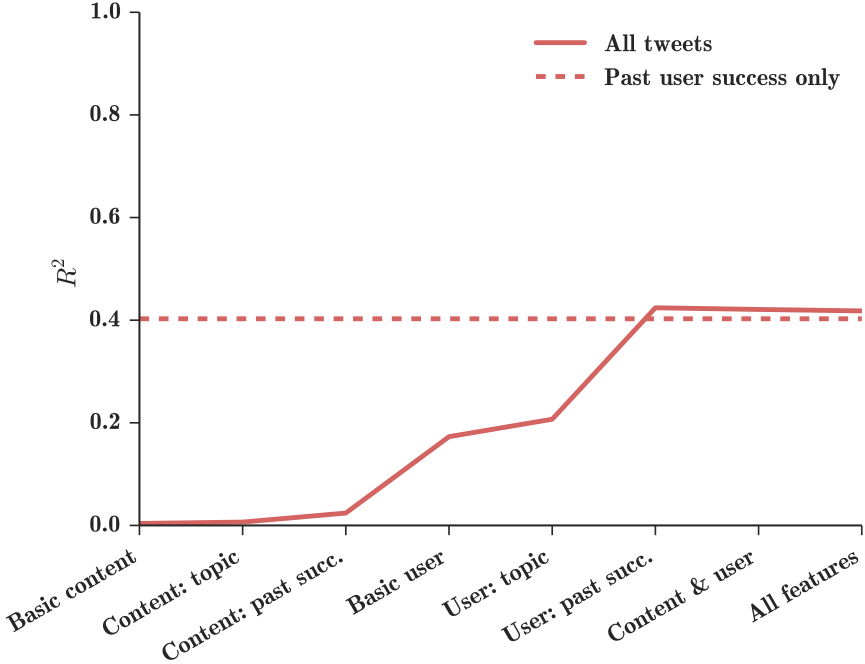- Measured the total cascade size for each seed tweet

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

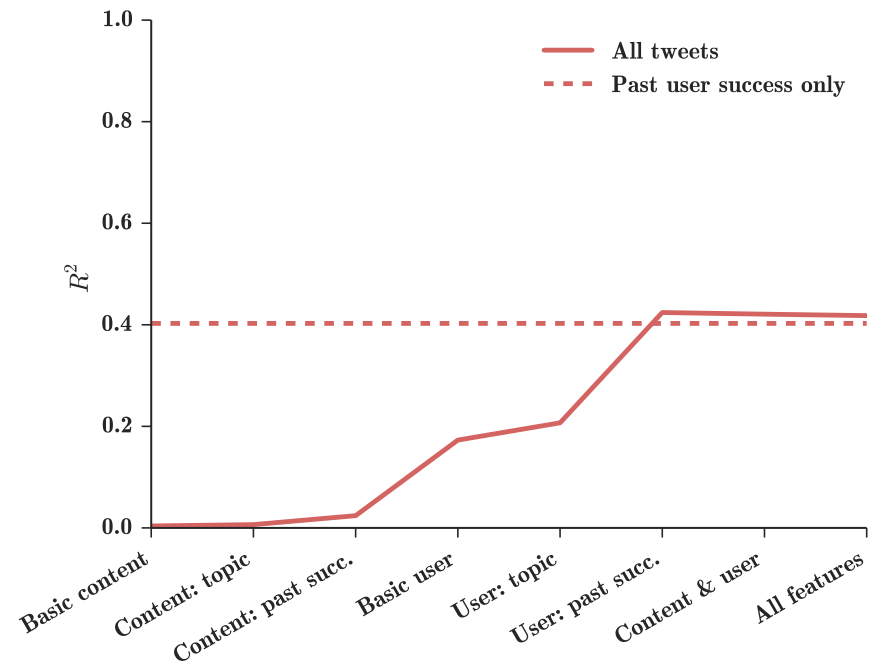best model explains roughly half of the variance in outcomes

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

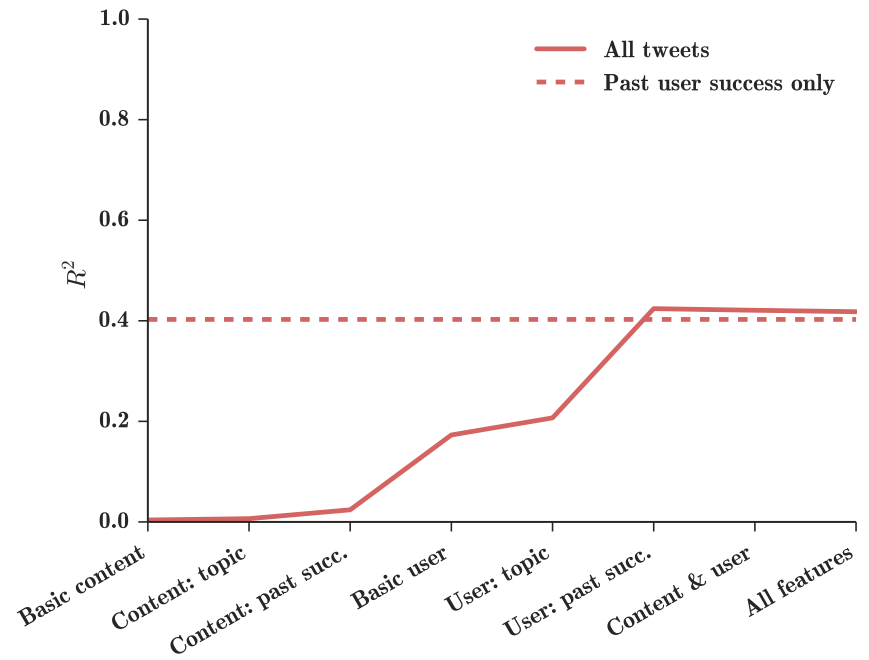## Content features alone perform poorly

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
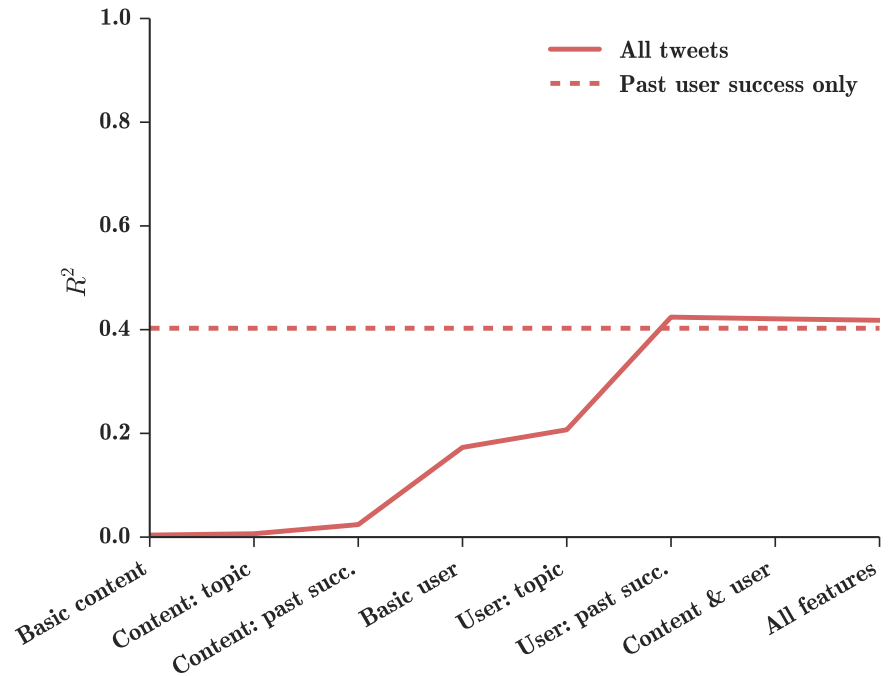

Model	Tweet time	Domain	Spam score	Category	Tweet topic	Past url success	User time	Followers	Friends	Statuses	User topic	Past user success	Topic interaction
1. Basic content	✓	✓	✓	✓									
2. Content, topic	✓	✓	✓	✓	✓								
3. Content, past succ.	✓	✓	✓	✓	✓	✓							
4. Basic user							✓	✓	✓	✓			
5.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

Detecting influenza epidemics using search engine query data

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

ⓘ 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

A United States

B Mid-Atlantic States

C New York City

Legend:
- 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

Y-axis: Weekly influenza-like illness (ILI) percent

X-axis: Date (influenza season)

Season labels: 2003/2004, 2004/2005, 2005/2006, 2006/2007, 2007/2008, 2008/2009, 2009/2010, 2010/2011, 2011/2012, 2012/2013

PLOS COMPUTATIONAL BIOLOGY

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

Article	Authors	Metrics	Comments	Media Coverage

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

Keyword, Author, or DO

Home | **Articles** | Front Matter | News | Podcasts | Authors | Sub

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

Article | Figures & SI | Info & Metrics | PDF

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
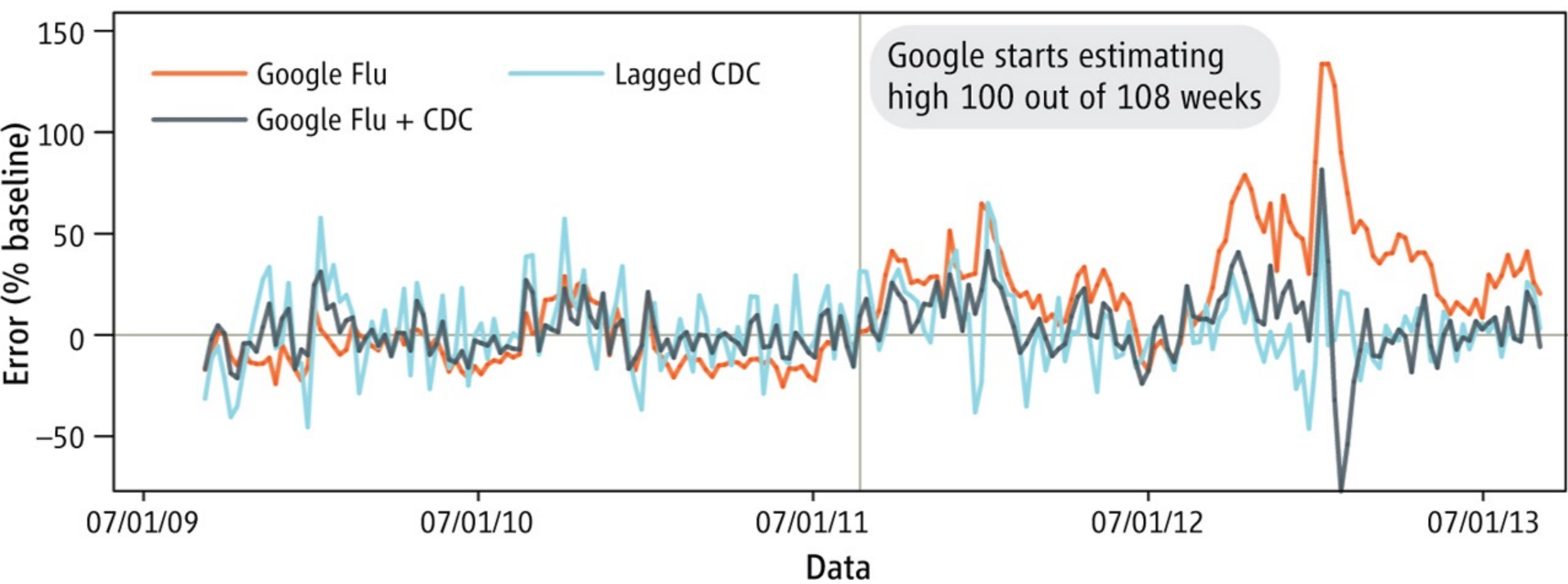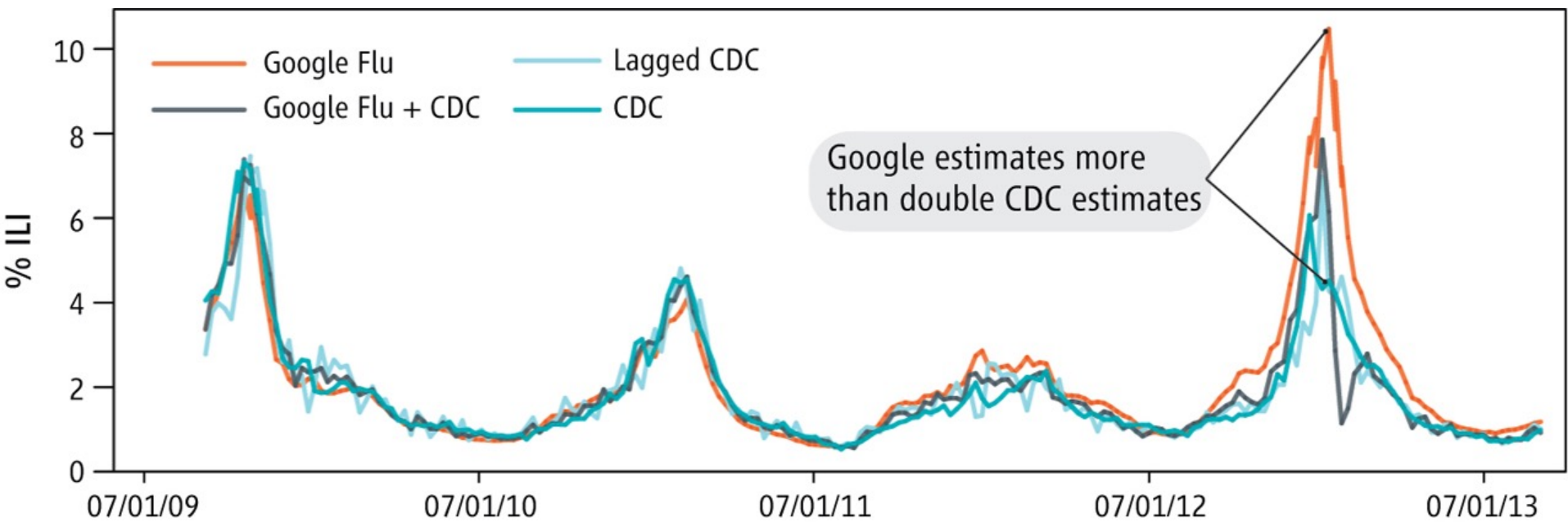
Big data hubris

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-

Journal Home   Browse Journal ⌄   Journal Info ⌄   Stay Connected ⌄   **Submit Paper**   Search 🔍

Article Menu   Close ⌃

Download PDF 📄

Open EPUB

Did you struggle to get access to this article? This product could help you

LEAN LIBRARY
A SAGE Publishing Company

📄 **Full Article**

Content List   ⌃

Abstract

Notes

References

Deeper data: a response to boyd and Crawford

Andre Brock

First Published August 24, 2015 | Research Article | 🔴 Check for updates
https://doi.org/10.1177/0163443715594105

Article information ⌄

Altmetric 5 🔒

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,

Class Exercise

Assess what "small" or "deep" data, in each of the following cases might be considered.

i) Predict how people react on a new product release (e.g., the latest version of iPhone), as observed on social media

ii) Predict whether greater anonymity leads to greater hate speech on social media

iii) Predict whether following recommended videos on YouTube leads down a more politically extreme rabbit hole

iv) Predict whether deplatforming reduces misinformation on social media

Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development

MORGAN KLAUS SCHEUERMAN*, University of Colorado Boulder, USA

EMILY DENTON, Google Research, USA

ALEX HANNA, Google Research, USA

Data is a crucial component of machine learning. The field is reliant on data to train, validate, and test models. With increased technical capabilities, machine learning research has boomed in both academic and industry settings, and one major focus has been on computer vision. Computer vision is a popular domain of machine learning increasingly pertinent to real-world applications, from facial recognition in policing to object detection for autonomous vehicles. Given computer vision's propensity to shape machine learning research and impact human life, we seek to understand disciplinary practices around dataset documentation—how data is collected, curated, annotated, and packaged into datasets for computer vision researchers and practitioners to use for model tuning and development. Specifically, we examine what dataset documentation communicates about the underlying values of vision data and the larger practices and goals of computer vision as a field. To conduct this study, we collected a corpus of about 500 computer vision datasets, from which we sampled 114 dataset publications across different vision tasks. Through both a structured and thematic content analysis, we document a number of values around accepted data practices, what makes desirable data, and the treatment of humans in the dataset construction process. We discuss how computer vision datasets authors value efficiency at the expense of care; universality at the expense of contextuality; impartiality at the expense of positionality; and model work at the expense of data work. Many of the silenced values we identify sit in opposition with social computing practices. We conclude with suggestions on how to better incorporate silenced values into the dataset creation and curation process.

1 Introduction

Data powers modern machine learning and artificial intelligence. Major tech corporations have built their intellectual and financial wealth by monetizing massive caches of text, images, transactions, and relationships. Modern nation-states and municipalities are increasingly data-driven, like when

---

Treading with caution

Attention to noise, bias, and "provenance" — broadly, where did data arise, what inferences were drawn from the data, and how relevant are those inferences to the present situation?
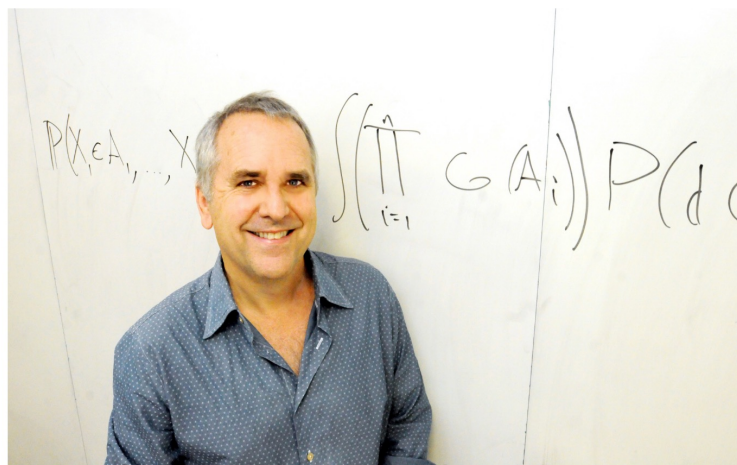

▶ Listen to this story
--:--   22:31

Photo credit: Peg Skorpinski

Artificial Intelligence — The Revolution Hasn't Happened Yet

Michael Jordan [Follow]
Apr 19, 2018 · 16 min read

Artificial Intelligence (AI) is the mantra of the current era. The phrase is intoned by technologists, academicians, journalists and venture capitalists alike. As with many phrases that cross over from technical academic fields into general circulation, there is significant misunderstanding

Science

Current Issue   First release papers   Archive   About ⌄   Submit manuscript

SPECIAL ISSUE ESSAY

Prediction and explanation in social systems

JAKE M. HOFMAN , AMIT SHARMA , AND , DUNCAN J. WATTS   Authors Info & Affiliations

🔒 GET ACCESS

Abstract

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

Stanford ML Group

CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar*, Jeremy Irvin*, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng

We develop an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists.

Chest X-rays are currently the best available method for diagnosing pneumonia, playing a crucial role in clinical care and epidemiological studies. Pneumonia is responsible for more than 1 million hospitalizations and 50,000 deaths per year in the US alone.
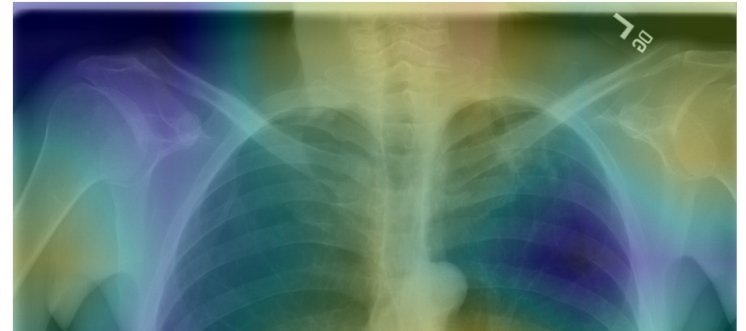
Predictions wouldn't make us superhumans; in fact we would still need the humans