

# CS 6474/CS 4803 Social Computing: Health and Well-Being

*Munmun De Choudhury*

[munmund@gatech.edu](mailto:munmund@gatech.edu)

Week 10 | March 16, 2022



Assignment II available (Due: Apr 13)





Le-shaker.fr



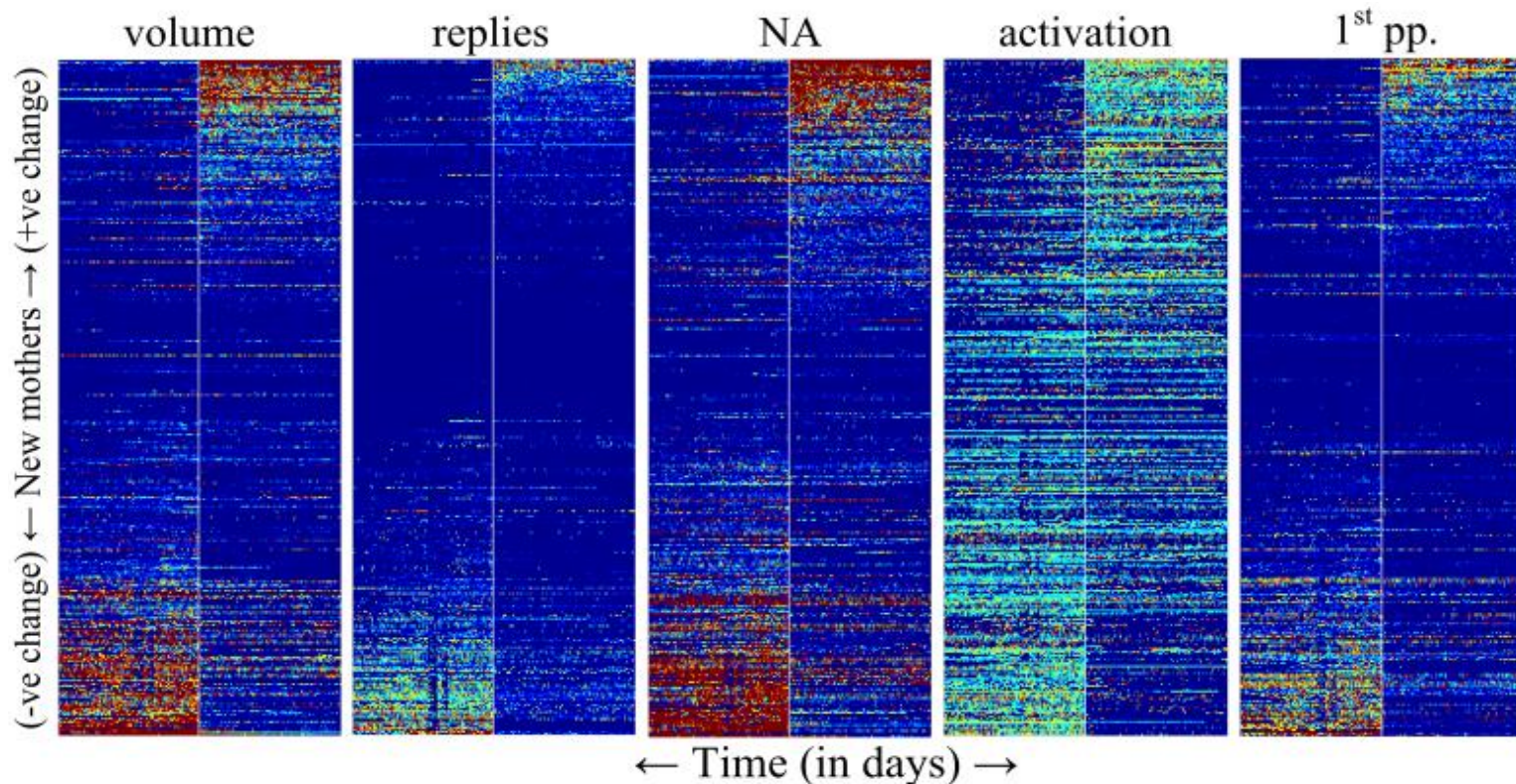
# Social Media Derived Behavioral and Affective Markers Predict Postpartum Changes



(De Choudhury, Counts, Horvitz, CSCW 2013; CHI 2013)



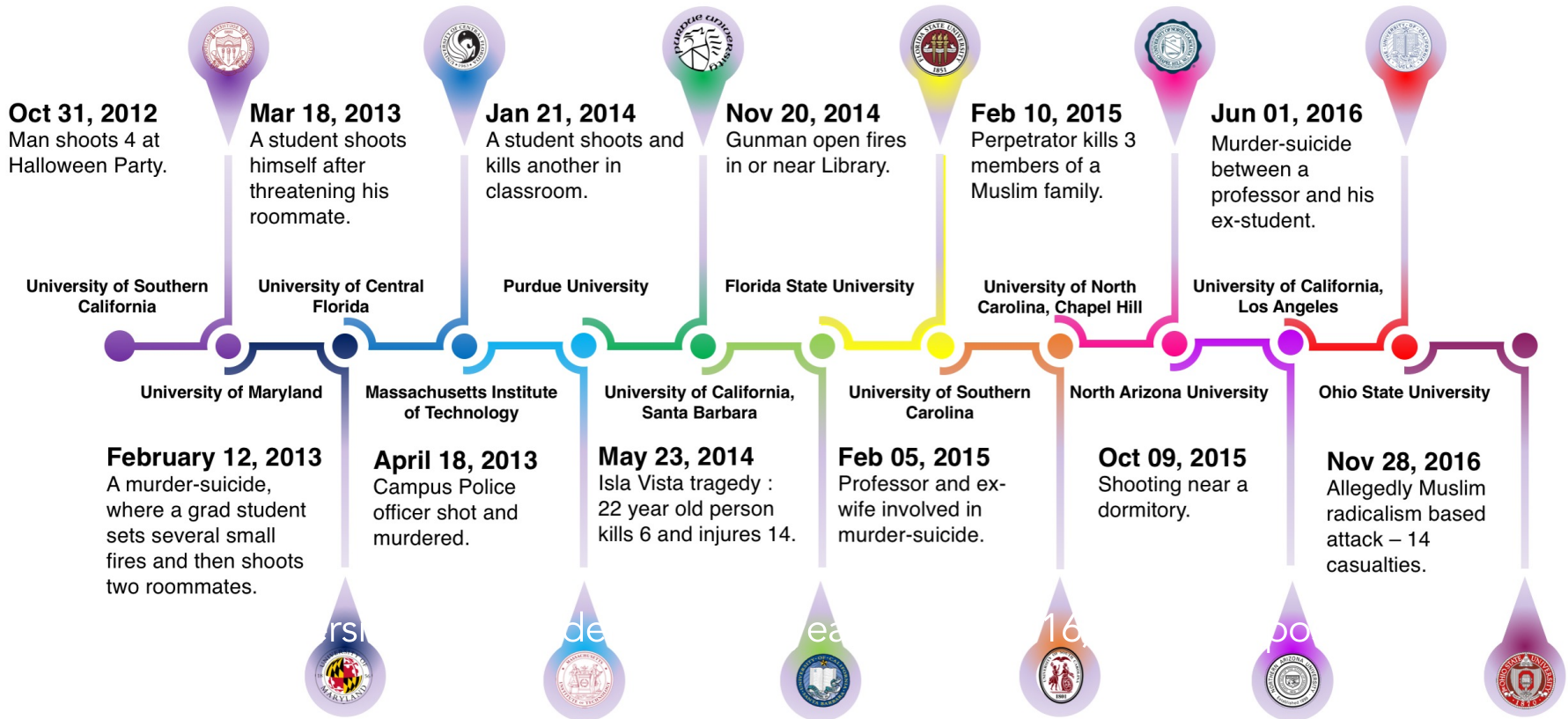
# Social Media Derived Behavioral and Affective Markers Predict Postpartum Changes



376 users (new mothers); 40,426 posts between March 2011 and July 2012

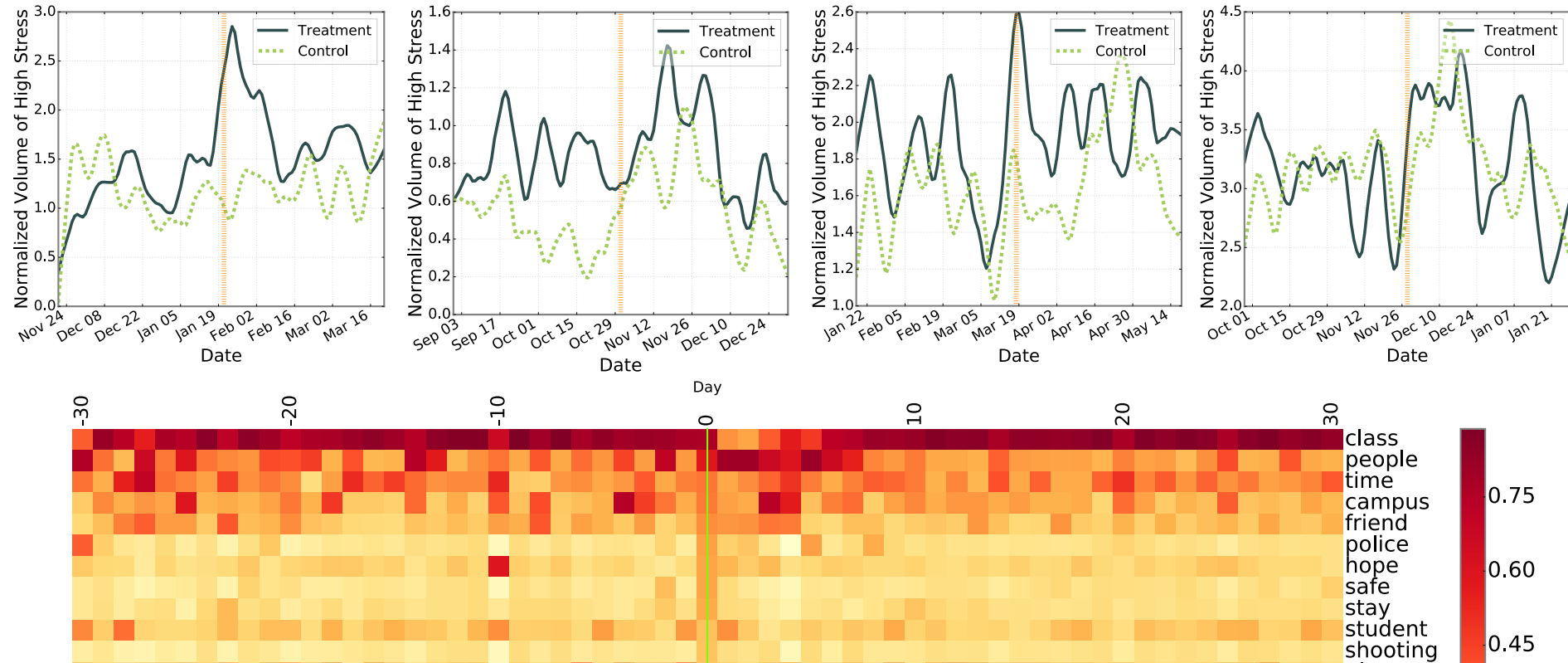


# Measuring Levels of Acute Stress in College Campuses with Social Media





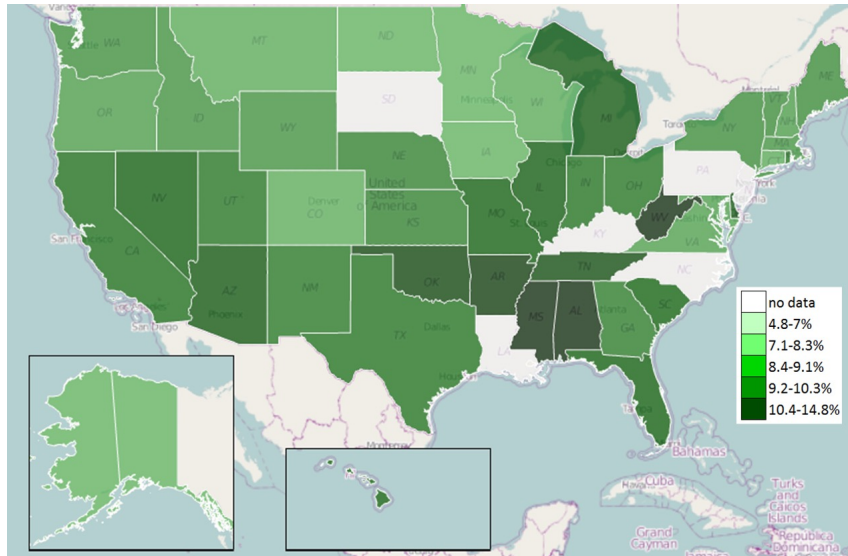
# Temporal and Linguistic Patterns of Stress



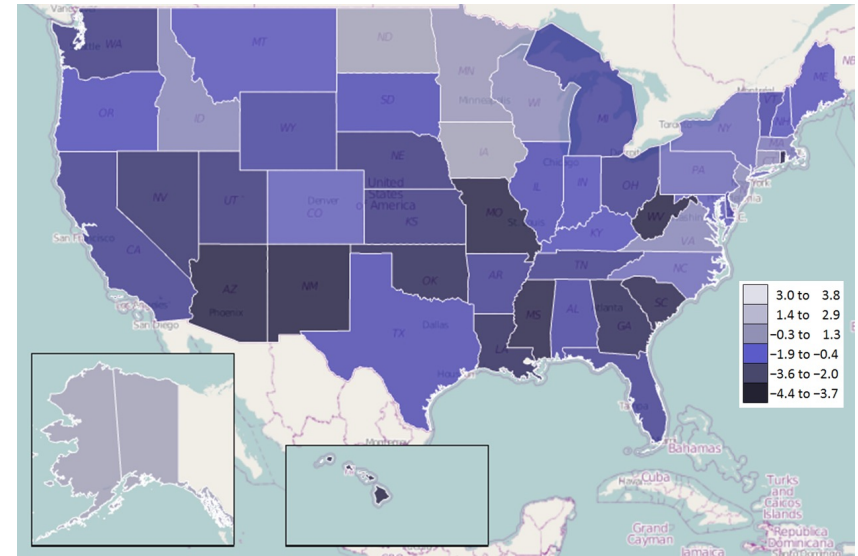


# Social media depression index

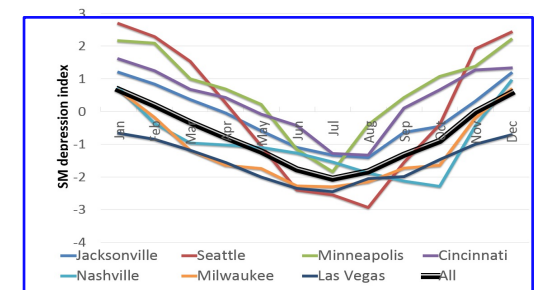
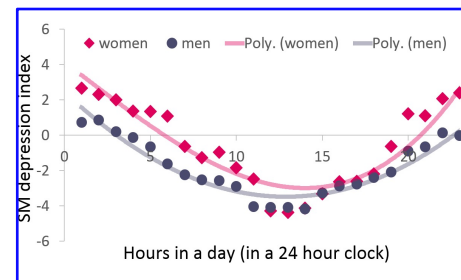
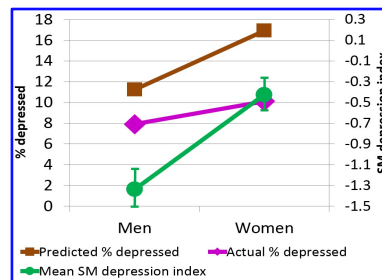
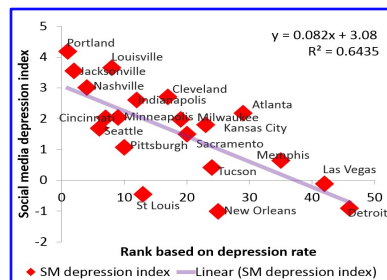
actual (BRFSS data)



predicted (SMDI)



Socio-demographic, spatio-temporal patterns of prevalence of depression





## Multi-Task Learning for Mental Health using Social Media Text

Adrian Benton  
Johns Hopkins University  
adrian@cs.jhu.edu

Margaret Mitchell  
Microsoft Research  
mmitchellai@google.com

### Abstract

We introduce initial groundwork for estimating suicide risk and mental health in a deep learning framework. By modeling multiple conditions, the system learns to make predictions about suicide risk and mental health at a low false positive rate. Conditions are modeled as tasks in a multi-task learning (MTL) framework, with gender prediction as an additional auxiliary task. We demonstrate the effectiveness of multi-task learning by comparison to a well-tuned single-task baseline with the same number of parameters. Our best MTL model predicts post-suicide risk with limited training data.

Automated monitoring of patients' language has traditional assessment measurements to additional support for mental health. This is the need for additional coverage, for example, benefits to clinicians. We explore some of learning and mental health media text that health conditions are learning methods that the opportunity to help health conditions.

## Facebook language predicts depression in medical records

Johannes C. Eichstaedt<sup>1,2</sup>, Robert J. Smith<sup>1,2</sup>, Raina M. Merchant<sup>1,2</sup>, Lyle H. Ungar<sup>1,2</sup>, Patrick Crutchley<sup>1,2</sup>, Daniel Protopopescu<sup>1,2</sup>, David A. Asch<sup>1,2</sup>, and H. Andrew Schwartz<sup>1</sup>

<sup>1</sup>Holistic Psychology Center, University of Pennsylvania, Philadelphia, PA 19104; <sup>2</sup>Perin Medicine Center for Digital Health, University of Pennsylvania, Philadelphia, PA 19104; <sup>3</sup>Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; <sup>4</sup>The Center for Health Equity Research and Promotion, Philadelphia Veterans Affairs Medical Center, Philadelphia, PA 19104; and <sup>5</sup>Computer Science Department, Stony Brook University, Stony Brook, NY 11794

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved September 11, 2016 (received for review February 26, 2016)

1 Introduction  
Suicide is one of the leading causes of death worldwide, and over 90% of suicides are preventable. However, detecting the risk of suicide is challenging. Currently, screening for suicide risk relies on both self-reports and clinical interviews during short sessions with a clinician. However, during short sessions with a clinician, it is often unclear when suicide risk is highest. Consequently, conditions like depression are often not adequately treated.

\*Now at Google Research.  
†https://www.nami.org/learn-more/Mental-Health-Conditions/Recovery/Conditions/SuicideRisk/AMH/2016/Communication-with-clinician-shop (Hollingshead, 2016).

Depression, the most prevalent mental illness, is underdiagnosed and undertreated, highlighting the need to extend the scope of current screening methods. Here, we use language from Facebook posts of consenting individuals to predict depression recorded in electronic medical records. We accessed the history of Facebook statuses posted by 683 patients visiting a large urban academic emergency department, 114 of whom had a diagnosis of depression in their medical records. Using only the language preceding their first documentation of a diagnosis of depression, we could identify depressed patients with fair accuracy (area under the curve (AUC) = 0.69), approximately matching the accuracy of screening surveys benchmarked against medical records. Restricting Facebook data to only the 6 months immediately preceding the first documented diagnosis of depression yielded a higher prediction accuracy (AUC = 0.72) for these users who had sufficient Facebook data. Significant prediction of future depression status was possible as far as 3 months before its first documentation. We found that language predictors of depression include emotional (sadness), interpersonal (loneliness, hostility), and cognitive (preoccupation with the self, rumination) processes. Unobtrusive depression assessment through social media of consenting individuals may become feasible as a scalable complement to existing screening and monitoring procedures.

big data | depression | social media | Facebook | screening

Each year, 7–26% of the US population experiences depression (1, 2), of whom only 15–49% receive minimally adequate treatment (3). By 2030, unipolar depressive disorders are predicted to be the leading cause of disability in high-income countries (4). The US Preventive Services Task Force recommends screening adults for depression in circumstances in which accurate diagnosis, treatment, and follow-up can be offered (5). These high rates of underdiagnosis and undertreatment suggest that existing procedures for screening and identifying depressed patients are inadequate. Novel methods are needed to identify and treat patients with depression.

By using Facebook language data from a sample of consenting patients who presented to a single emergency department, we built a method to predict the first documentation of a diagnosis of depression in the electronic medical record (EMR). Previous research has demonstrated the feasibility of using Twitter (6, 7) and Facebook language and activity data to predict depression (8), postpartum depression (9), suicidality (10), and post-traumatic stress disorder (11), relying on self-report or diagnoses on Twitter (12, 13) or the participants' responses to screening surveys (6, 7, 9) to establish health status. mental health status. In contrast to this prior work relying on self-report, we established a depression diagnosis by using medical codes from an EMR.

As described by Padner et al. (14), patients in a single urban academic emergency department (ED) were asked to share access to their medical records and the statuses from their Facebook timelines. We used depression-related International Classification of Diseases (ICD) codes in patients' medical records as a proxy for

the diagnosis of depression, which prior research has shown is feasible with moderate accuracy (15). Of the patients enrolled in the study, 114 had a diagnosis of depression in their medical records. For these patients, we determined the date at which their first documentation of a diagnosis of depression was recorded in the EMR at the hospital system. We analyzed the Facebook data generated by each user before this date. We sought to simulate a realistic screening scenario, and so, for each of these 114 patients, we identified 5 random control patients without a diagnosis of depression in the EMR, examining only the Facebook data they created before the corresponding depressed patient's first date of a recorded diagnosis of depression. This allowed us to compare depressed and control patients' data across the same time span and to model the prevalence of depression in the larger population (~14.7%).

### Results

**Prediction of Depression.** To predict the future diagnosis of depression in the medical record, we built a prediction model by using the textual content of the Facebook posts, post length, frequency of posting, temporal posting patterns, and demographics (Materials and Methods). We then evaluated the performance of this model by comparing the probability of depression estimated by our algorithm against the actual presence or absence of depression for each patient in the medical record (using 10-fold cross-validation to avoid overfitting). Varying the threshold of this probability for diagnosis

### Significance

Depression is disabling and treatable, but underdiagnosed. In this study, we show that the content shared by consenting users on Facebook can predict a future occurrence of depression in their medical records. Language predictive of depression includes references to typical symptoms, including sadness, loneliness, hostility, rumination, and increased self-reference. This study suggests that an analysis of social media data could be used to screen consenting individuals for depression. Further, social media content may point clinicians to specific symptoms of depression.

Author contributions: J.C.E., R.J.S., and H.A.S. designed research; J.C.E., R.J.S., D.P., and H.A.S. performed research; J.C.E. and H.A.S. analyzed data; and J.C.E., R.J.S., M.M., L.U., D.A., and H.A.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

This open access article is distributed under Creative Commons Attribution-NonCommercial 4.0 International license.

Supplementary Information: The data reported in this paper have been deposited in the Open Science Framework, 10.21203/rs.2016.10.01.v1.

\*To whom correspondence should be addressed. Email: jpeh@perin.med.upenn.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1602311115/-DC1.

Published online October 15, 2016.



Available online at www.sciencedirect.com

ScienceDirect



## Detecting depression and mental illness on social media: an integrative review

Sharath Chandra Guntuku<sup>1</sup>, David B Yaden<sup>1</sup>, Margaret L Kern<sup>2</sup>, Lyle H Ungar<sup>1</sup> and Johannes C Eichstaedt<sup>1</sup>

Although rates of diagnosing mental illness have improved over the past few decades, many cases remain undetected. Symptoms associated with mental illness are observable on Twitter, Facebook, and web forums, and automated methods are increasingly able to detect depression and other mental illnesses. In this paper, recent studies that aimed to predict mental illness using social media are reviewed. Mentally ill users have been identified using screening surveys, their public sharing of a diagnosis on Twitter, or by their membership in an online forum, and they were distinguishable from control users by patterns in their language and online activity. Automated detection methods may help to identify depressed or otherwise at-risk individuals through the large-scale passive monitoring of social media, and in the future may complement existing screening procedures.

### Addresses

<sup>1</sup>University of Pennsylvania, Philadelphia, PA, United States  
<sup>2</sup>The University of Melbourne, Melbourne, Australia

Corresponding author: Eichstaedt, Johannes C (johannes.eichstaedt@gmail.com)

### Current Opinion in Behavioral Science

This review comes from a themed

editorial by Michael Kosinski and Tamas

For a complete overview see the

Available online 31st July 2017

<http://dx.doi.org/10.1016/j.cobeha.2017.07.001>

2352-1546/© 2017 Elsevier Ltd. All

### Introduction

The widespread use of social media by users to help reduce growing number of studies of social media contexts, linking social media with stress, and other mental illnesses studies of this kind focus continues to be under-the-cases detected by primary 13–49% receiving minor

www.sciencedirect.com

## Natural Language Processing of Social Media as Screening for Suicide Risk

Glen Coppersmith, Ryan Leary, Patrick Crutchley and Alex Fine

Quincy, Boston, MA, USA

**ABSTRACT:** Suicide is among the 10 most common causes of death, as assessed by the World Health Organization. For every death by suicide, an estimated 138 people's lives are negatively affected, and almost any other disease around suicide death is usually alarming. The pervasiveness of social media—and the need-to-use mobile devices used to access social media networks—offers new types of data for understanding the behavior of those who (attempt to) take their own lives and suggests new possibilities for preventive intervention. We demonstrate the feasibility of using social media data to detect those at risk for suicide. Specifically, we use natural language processing and machine learning to identify deep learning techniques to detect quantifiable signals around suicide attempts, and describe designs for an automated system for estimating suicide risk, usable by those without specialized mental health training (e.g., a primary care doctor). We also discuss the ethical use of such technology and examine privacy implications. Currently, this technology is only used for intervention for individuals who have "opted in" for the analysis and intervention, but the technology enables scalable screening for suicide risk, potentially identifying many people who are at risk for suicide. This raises a significant cultural question about the trade-off between privacy and prevention—what have been the potential life-saving technology that is currently screening only a fraction of the people at risk because of respect for their privacy. In the current trade-off between privacy and prevention the right one?

**KEYWORDS:** Suicide, suicide screening, suicide prevention, social media, data science, natural language processing

RECEIVED February 28, 2016; ACCEPTED: June 25, 2016

NOTE: Proceedings from the Digital Mental Health Conference, London, 2017.

**FUNDING:** The authors declared receipt of the following financial support for the research, authorship, and/or publication of this article: Quincy is a for-profit company that designs analytic products related to mental health. Quincy funded this research in the interest of serving our customers with the best possible technology.

### Introduction

An estimated 16 million suicide attempts occur each year. Of these, approximately 800,000 people will die from those attempts.<sup>1</sup> Suicide deaths have increased by 24% in the past 20 years, making suicide one of the top 10 causes of death in the United States,<sup>2</sup> a pattern that seems to be constant across geographic region within the country.<sup>3</sup> Not only is the magnitude of the problem large and worsening, there has been little progress made over the past 50 years in understanding suicide and improving outcomes in at-risk individuals.<sup>4</sup> The subtlety of the problem reflects its complexity, and the diversity between causal factors underlying it. Here we focus on one piece of the puzzle: how can we identify those who are at risk of taking (or attempting to take) their own life, and how can this screening be used to foster effective interventions?

Assessing an individual's risk for suicidal behavior is difficult. Experienced and talented clinicians frequently struggle to correctly interpret signals in their patients' behavior that are indicative of suicide risk. Setting aside the profound difficulties associated with understanding an individual's personal history and in relationship to their capacity and motivations for self-harm, there is at least 2 practical reasons that assessing suicide risk is difficult: (1) the latency between the onset of suicide risk and the suicide attempt itself may be too small for interventions requiring contact with health professionals, and (2) most existing methods for detecting high risk of suicide require that individuals first be asked to disclose their suicidal thoughts to a health professional. In this article, we explore the possibility that digital life data—that is, the interactions that a person

has with digital devices, through the daily course of their life—collected passively but with consent might at least partially address each of these difficulties.

Individuals come to be at risk for suicide at different temporal intervals relative to suicide attempts. For instance, the kind of social isolation that is frequently associated with suicide can gradually accumulate over the course of a person's life or may become acute in a very short period of time after a traumatic life event such as the loss of a loved one.

Moreover, once an individual is engaged with a health care professional, standard methods of suicide intervention require both that the clinician administer a standardized assessment (often in the form of a questionnaire) and that the patients disclose their intention to harm themselves. Each of these presents its own challenges. First, administering a suicide screening tool may place an unreasonable burden on the health care provider. The standard for suicide screening within the health care system is a Beck's Scale for Suicide Ideation, a 5- or 19-item questionnaire examining the patient's actual and passive desire for suicide, and any specific plans they might have.<sup>5</sup> Many patients who are at risk for suicide only interact with primary care physicians (PCPs) or emergency departments (EDs) rather than those with psychiatric specialists. Such health care providers may lack the time or the training to administer a specific questionnaire for suicide risk. Indeed, enabling PCPs and EDs to better screen for suicide risk has been proposed as a means for reducing the suicide rate.<sup>6,7</sup> Second, patients cannot always be relied upon to disclose suicidal thoughts in the clinical setting.<sup>8</sup> These factors have the

## Identifying Social Media Markers of Machine Learning and Clinical

MS, Asra F Rizvi<sup>1,2\*</sup>, MA; Munmun De Choudhury<sup>1</sup>, PhD;

ed States

process in differentiating individuals who have included expert input to evaluate

media to more accurate identification linguistic analysis of shared content is

es of schizophrenia, was appraised for a classifier algorithm to distinguish users report appraisals on new, unseen Twitter

including greater use of interpersonal (on biological processes (P<0.01). The from control users with a mean accuracy of 0.71, precision, recall, and accuracy

erprise from multiple fields to strengthen online. These collaborations are crucial

analysis; Twitter

J Med Internet Res 2017 | vol. 19 | iss. 4 | e209 | p. 1



Creative Commons Attribution-NonCommercial 4.0 International license. This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International license (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial use, reproduction and distribution of the work under the same license, provided the original work is properly cited and the source is acknowledged. For more information, see the Creative Commons Attribution-NonCommercial 4.0 International license (http://creativecommons.org/licenses/by-nc/4.0/).

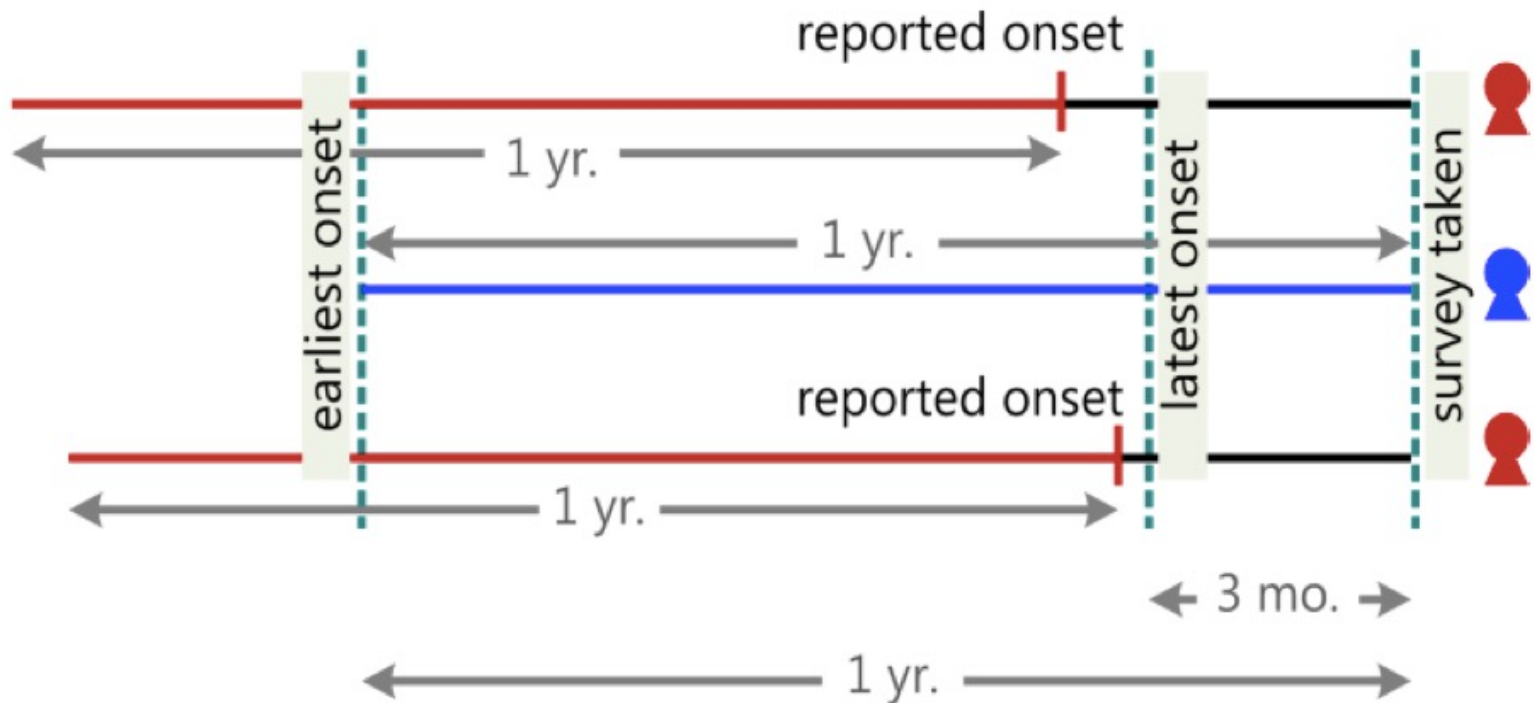


# Predicting Depression via Social Media



# 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



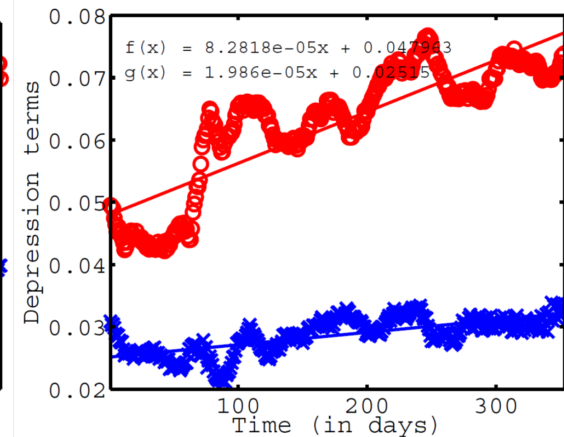
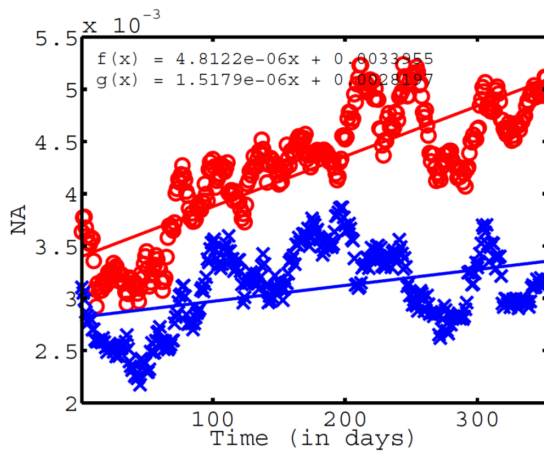
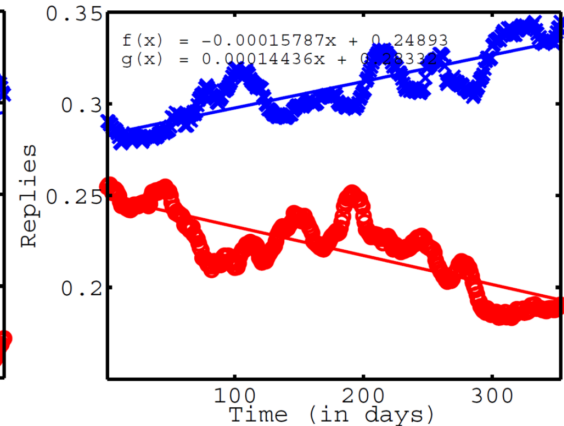
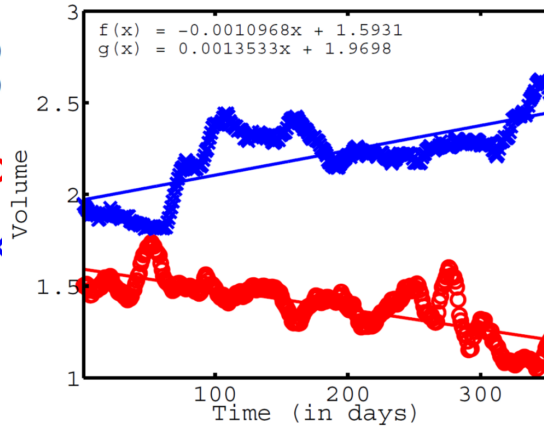
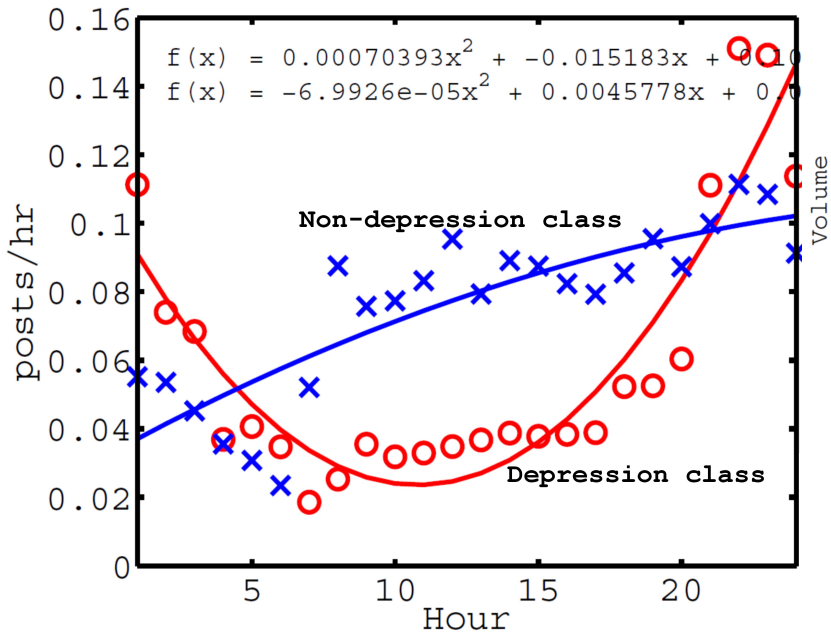


# Summary

- 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





# Summary

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



# Discussion Point I

In this 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 or gathering high quality ground truth?



# Discussion Point II

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?



RESEARCH ARTICLE

# Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance

**Mauricio Santillana<sup>1,2,3\*</sup>, André T. Nguyen<sup>1</sup>, Mark Dredze<sup>4</sup>, Michael J. Paul<sup>5</sup>, Elaine O. Nsoesie<sup>6,7</sup>, John S. Brownstein<sup>2,3</sup>**

**1** Harvard School of Engineering and Applied Sciences, Cambridge, Massachusetts, United States of America, **2** Boston Children's Hospital Informatics Program, Boston, Massachusetts, United States of America, **3** Harvard Medical School, Boston, Massachusetts, United States of America, **4** Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, United States of America, **5** Department of Information Science, University of Colorado, Boulder, Colorado, United States of America, **6** Department of Global Health, University of Washington, Seattle, Washington, United States of America, **7** Institute for Health Metrics and Evaluation, Seattle, Washington, United States of America

\* [msantill@fas.harvard.edu](mailto:msantill@fas.harvard.edu)



CrossMark  
click for updates

## OPEN ACCESS

**Citation:** Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS (2015) Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance. PLoS Comput Biol 11(10): e1004513. doi:10.1371/journal.pcbi.1004513

## Abstract

We present a machine learning-based methodology capable of providing real-time (“now-cast”) and forecast estimates of influenza activity in the US by leveraging data from multiple data sources including: Google searches, Twitter microblogs, nearly real-time hospital visit records, and data from a participatory surveillance system. Our main contribution consists of combining multiple influenza-like illnesses (ILI) activity estimates, generated indepen-





PDF



More ▾



Cite



Permissions

**Original Investigation** | Health Informatics

December 23, 2020

# Development of a Machine Learning Model Using Multiple, Heterogeneous Data Sources to Estimate Weekly US Suicide Fatalities

Daejin Choi, PhD<sup>1</sup>; Steven A. Sumner, MD<sup>2</sup>; Kristin M. Holland, PhD<sup>3</sup>; John Draper, PhD<sup>4</sup>; Sean Murphy, PhD<sup>4</sup>; Daniel A. Bowen, MPH<sup>3</sup>; Marissa Zwald, PhD<sup>3</sup>; Jing Wang, MD<sup>3</sup>; Royal Law, PhD<sup>5</sup>; Jordan Taylor, BS<sup>6</sup>; Chaitanya Konjeti, BS<sup>6</sup>; Munmun De Choudhury, PhD<sup>6</sup>

» [Author Affiliations](#) | [Article Information](#)

*JAMA Netw Open.* 2020;3(12):e2030932. doi:10.1001/jamanetworkopen.2020.30932

## Key Points

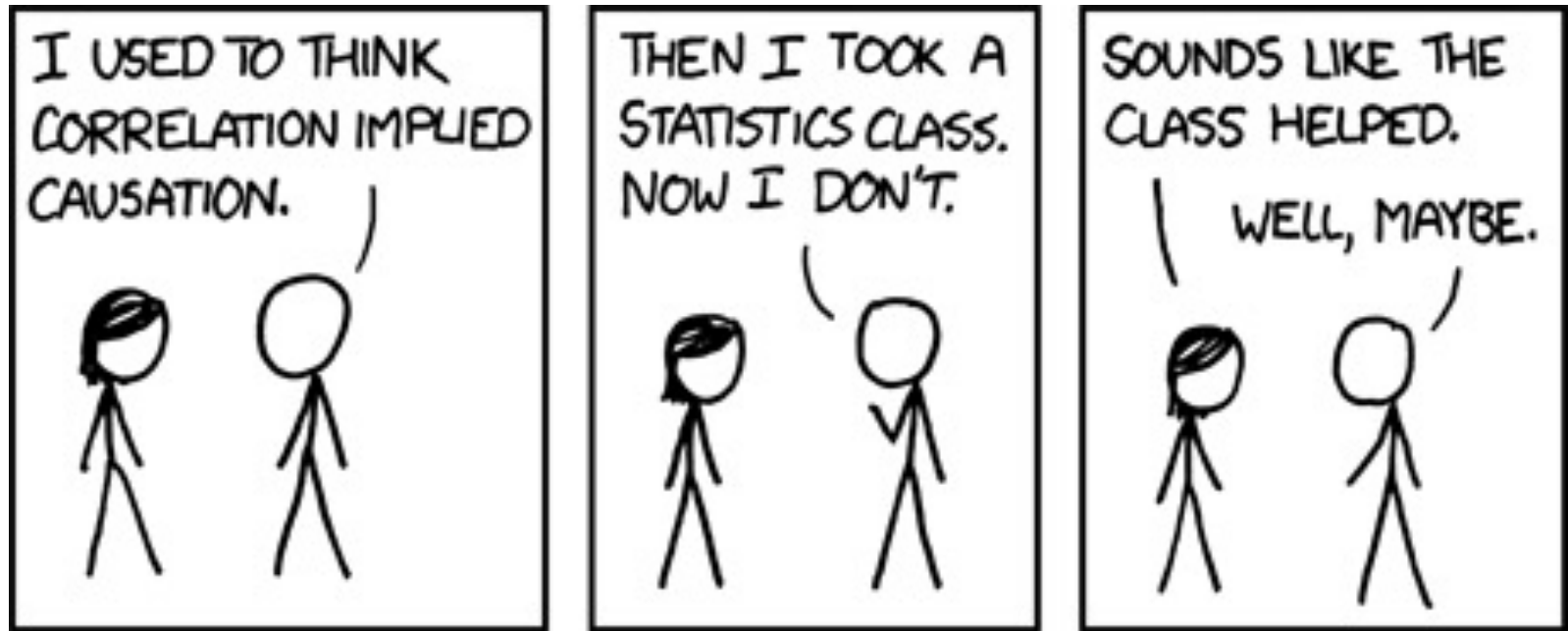
**Question** Can real-time streams of secondary information related to suicide be used to accurately estimate suicide fatalities in the US in real time?



# Discussion Point III

But are models trained on aggregated group-level differences useful at the individual level?





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.





The slide features a solid black background. At the top edge, there is a horizontal bar composed of three colored segments: blue on the left, purple in the middle, and yellow on the right. At the bottom edge, there is a similar horizontal bar with three colored segments: yellow on the left, purple in the middle, and blue on the right.

What comes next?



# What comes next?

Social Media + Machine Learning for clinical interventions

Efficacy

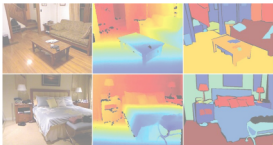


Validity

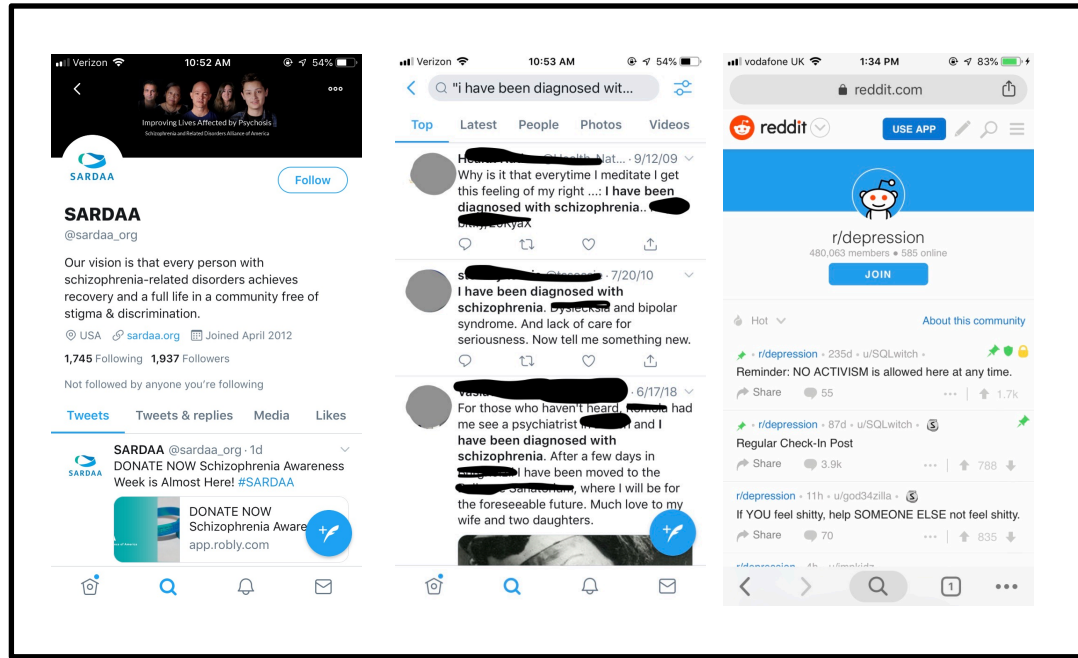




# SOCIAL MEDIA + MACHINE LEARNING



Ground truth label:  
Readily available



Proxy ground truth  
labels



Ground truth label:  
clinical assessment



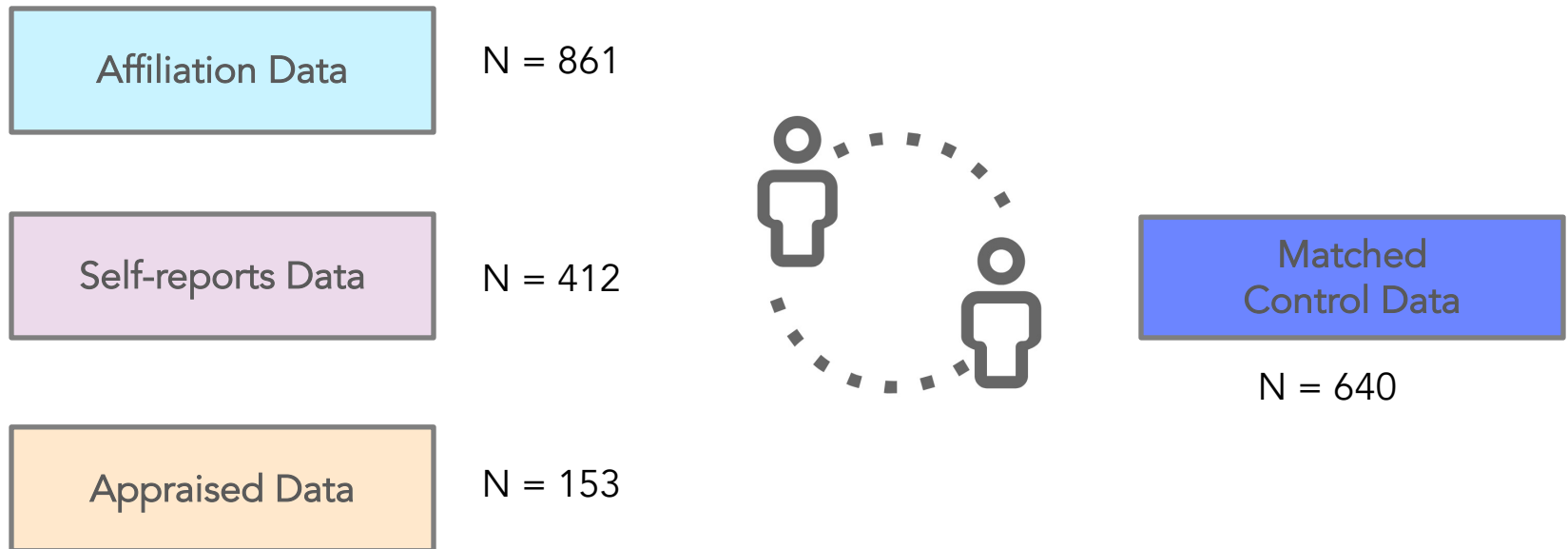
Construct Validity: Do the proxy diagnostic signals  
objectively and accurately measure what they  
claim to measure (clinical mental illness diagnosis)



Theoretical/Clinical grounding: Is what is being  
measured by the proxy diagnostic signals  
**valid in itself?**



# Proxy data sets: diagnostic signals for schizophrenia on Twitter





# Patient's social media data

Schizophrenia  
Patient Data

N = 88

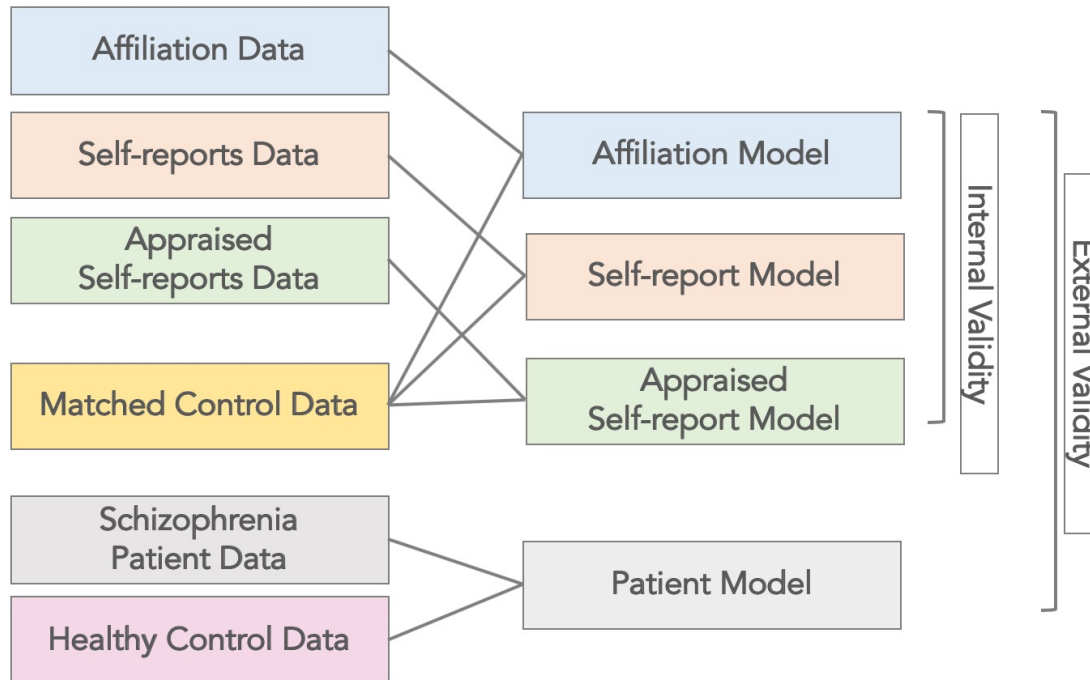
Healthy Control Data

N = 55





# Methodology: Triangulation



Binary classification task:

Distinguishing those with schizophrenia from control populations



# Efficacy

High internal validity  
Very low external validity

Cross Validation

Testing on patient data

Affiliation Model

0.89

0.21

Self-report Model

0.72

0.48

Appraised Model

0.80

0.55














































Patient Model

0.72

0.76



# Issues with Construct Validity

<i>Affiliation</i>	$\beta$	<i>Appraised</i>	$\beta$	<i>Patient</i>	$\beta$
i'm	 -0.825	NegAffect	 0.063	cog mech	 -0.003
stigma	 0.665	negation	 0.074	present	 -0.002
mhchat	 0.696	present	 0.40	body	 -0.002
body	 0.729	help	 0.401	verbs	 -0.002
bipolar	 0.774	thought	 0.41	social	 -0.002
work	 0.919	i'm	 0.44	aux verbs	 -0.002
self	 0.961	die	 0.45	help	 0.0002
social	 1.109	alone	 0.45	feeling	 0.001
care	 1.111	hard	 0.457	i'm	 0.002
depression	 1.116	cry	 0.50	gonna	 0.002
suicide	 1.133	body	 0.52	angel	 0.002
thanks	 1.445	feeling	 0.523	burning	 0.002
illness	 1.447	verbs	 0.58	pray	 0.003
help	 1.632	sorry	 0.662	lifetime	 0.005
mental health	 1.866	gonna	 0.63	attack	 0.006



# Main Takeaway

If the broader research agenda is to use social media data to inform clinical decision-making, such as early diagnosis, treatment or patient-provider interventions, **(social media) data collection and machine learning model development should happen in context**

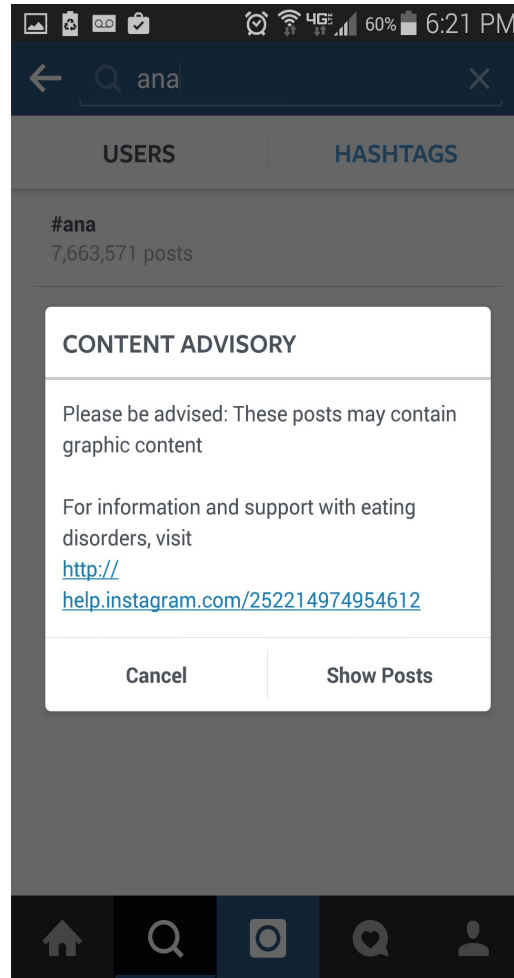
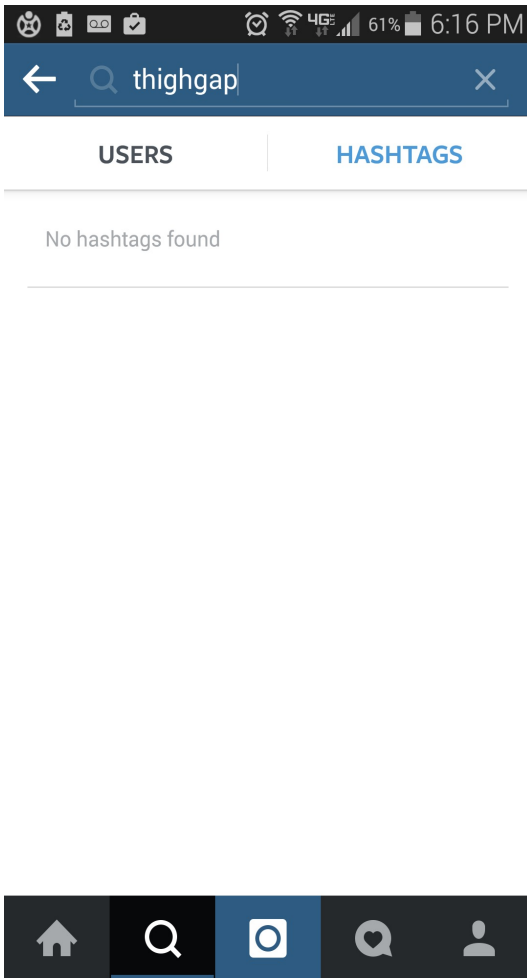


# Class Exercise

Describe a design idea where we can use social media based depression (or other mental health condition like schizophrenia) predictors to help people. How would it negotiate privacy and ethical issues?



# Improving “Blanket” Interventions



## 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](http://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.

[Go back](#)

[View search results](#)





suicide

Web

News

Images

Videos

Books

More ▾

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](http://www.suicidepreventionlifeline.org)

●●○○ AT&T LTE

2:19 PM



facebook.com



facebook



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.

See Post

Continue





# A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media

Stevie Chancellor  
Georgia Tech  
Atlanta, GA, US  
schancellor3@gatech.edu

Michael L Birnbaum  
Northwell Health  
Glen Oaks, NY, US  
mbirnbaum@northwell.edu

Eric D. Caine  
University of Rochester  
Rochester, NY, US  
Eric\_Caine@urmc.rochester.edu

Vincent M. B. Silenzio  
University of Rochester  
Rochester, NY, US  
vincent.silenzio@rochester.edu

Munmun De Choudhury  
Georgia Tech  
Atlanta, GA, US  
munmund@gatech.edu

## ABSTRACT

Powered by machine learning techniques, social media provides an unobtrusive lens into individual behaviors, emotions, and psychological states. Recent research has successfully employed social media data to predict mental health states of individuals, ranging from the presence and severity of mental disorders like depression to the risk of suicide. These algorithmic inferences hold great potential in supporting early detection and treatment of mental disorders and in the design of interventions. At the same time, the outcomes of this research can pose great risks to individuals, such as issues of incorrect, opaque algorithmic predictions, involvement of bad or unaccountable actors, and potential biases from intentional or inadvertent misuse of insights. Amplifying these tensions, there are also divergent and sometimes inconsistent methodological gaps and under-explored ethics and privacy dimensions. This paper presents a taxonomy of these concerns and ethical challenges, drawing from existing literature, and poses questions to be resolved as this research gains traction. We identify three areas of tension: ethics committees and the gap of social media research; questions of validity, data, and machine learning; and implications of this research for key stakeholders. We conclude with calls to action to begin resolving these interdisciplinary dilemmas.

## CCS CONCEPTS

• **Human-centered computing** → Collaborative and social computing; Social media; • **Applied computing** → Psychology;

*Conference on Fairness, Accountability, and Transparency (FAT\* '19)*. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3287560.3287587>

## 1 INTRODUCTION

Last year, Facebook unveiled automated tools to identify individuals contemplating suicide or self-injury [75, 62]. The company claims that they “use pattern recognition technology to help identify posts and live streams as likely to be expressing thoughts of suicide,” which then can deploy resources to assist the person in crisis [75]. Reactions to Facebook’s suicide prevention artificial intelligence (AI) are mixed, with some concerned about the use of AI to detect suicidal ideation as well as potential privacy violations [86]. Other suicide prevention AIs, however, have been met with stronger public backlash. Samaritan’s Radar, an app that scanned a person’s friends for concerning Twitter posts, was pulled from production, citing concerns for data collection without user permission [54], as well as enabling harassers to intervene when someone was vulnerable [4].

Since 2013, a new area of research has incorporated techniques from machine learning, natural language processing, and clinical psychology to categorize individuals’ moods and expressed well-being from social media data. These algorithms are powerful enough to infer with high accuracy whether an individual might be suffering from disorders such as major depression [28, 19, 84, 73, 78], postpartum depression [26, 27], post-traumatic stress [21], schizophrenia [60, 6], and suicidality [15, 22]. These algorithms can also reveal symptomatology linked to psychiatric challenges,