Carnegie Mellon

Thesis Defense

Institute for Software Research
Societal Computing

institute for SOFTWARE RESEARCH

# Bias and beyond in digital trace data

Momin M. Malik

http://mominmalik.com/defense.pdf

Thursday, 9 August 2018
9 am – 12 pm
Wean Hall 7500

| | |
|---|---|
| Jürgen Pfeffer *(co-chair)* | Institute for Software Research |
| Anind K. Dey *(co-chair)* | Human-Computer Interaction Institute |
| Cosma Rohilla Shalizi | Department of Statistics & Data Science |
| David Lazer | Northeastern University |

Eric Fisher (2011). European detail map of Flickr and Twitter locations, https://flic.kr/p/a1vp4W

*“We check our* ***e-mails*** *regularly, make* ***mobile phone calls****…*

David Lazer et al. (2009). Computational social science. *Science* 323 (5915), 721–723.
Eric Fisher (2011). European detail map of Flickr and Twitter locations, https://flic.kr/p/a1vp4W

*“We check our* ***e-mails*** *regularly, make* ***mobile phone calls****… We may post* ***blog entries*** *accessible to anyone, or maintain friendships through* ***online social networks****.*

David Lazer et al. (2009). Computational social science. *Science* 323 (5915), 721–723.
Eric Fisher (2011). European detail map of Flickr and Twitter locations, https://flic.kr/p/a1vp4W

*"We check our* ***e-mails*** *regularly, make* ***mobile phone calls****... We may post* ***blog entries*** *accessible to anyone, or maintain friendships through* ***online social networks****. Each of these transactions leaves* ***digital traces*** *that can be compiled into comprehensive pictures of both individual and group behavior,*

David Lazer et al. (2009). Computational social science. *Science* 323 (5915), 721–723.
Eric Fisher (2011). European detail map of Flickr and Twitter locations, https://flic.kr/p/a1vp4W

*"We check our* ***e-mails*** *regularly, make* ***mobile phone calls****… We may post* ***blog entries*** *accessible to anyone, or maintain friendships through* ***online social networks****. Each of these transactions leaves* ***digital traces*** *that can be compiled into comprehensive pictures of both individual and group behavior, with the* ***potential to transform our understanding of our lives, organizations, and societies****."*

David Lazer et al. (2009). Computational social science. *Science* 323 (5915), 721–723.
Eric Fisher (2011). European detail map of Flickr and Twitter locations, https://flic.kr/p/a1vp4W

Carnegie Mellon University
School of Computer Science



# What could go wrong?

**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 communication and culture, or narrow the palette of research options and alter what 'research' means? Given the rise of Big Data as a socio-technical phenomenon, we argue that it is necessary to critically interrogate its assumptions and biases. In this article, we offer six provocations to spark conversations about the issues of Big Data: a cultural, technological, and scholarly phenomenon that rests on the interplay of technology, analysis, and mythology that provokes extensive utopian and dystopian rhetoric.*

**Keywords** Big Data; analytics; social media; communication studies; social network sites; philosophy of science; epistemology; ethics; Twitter

(*Received 10 December 2011; final version received 20 March 2012*)

Carnegie Mellon University School of Computer Science



# What could go wrong?

contributed articles

DOI:10.1145/2001269.2001297

**The power to predict outcomes based on Twitter data is greatly exaggerated, especially for political elections.**

BY DANIEL GAYO-AVELLO

## Don't Turn Social Media Into Another 'Literary Digest' Poll

CONTENT PUBLISHED IN microblogging systems like Twitter can be data-mined to take the pulse of society, and a number of studies have praised the value of relatively simple approaches to sampling, opinion mining, and sentiment analysis. Here, I play devil's advocate, detailing a study I conducted late 2008/early 2009 in which such simple approaches largely overestimated President Barack Obama's victory in the

Many Twitter users do not protect their tweets, which then appear in the so-called public timeline. They are accessible through Twitter's own API, so are easily accessed and collected.

Twitter's original slogan—"What are you doing?"—encouraged users to share updates about the minutia of their daily activities with their friends. Twitter has since evolved into a complex information-dissemination platform, especially during situations of mass convergence.[8] Under certain circumstances, Twitter users not only provide information about themselves but also real-time updates of current events.[a]

Today Twitter is a source of information on such events, updated by millions of users[b] worldwide reacting to events as they unfold, often in real time. It was only a matter of time before the research community turned to it as a rich source of social, commercial, marketing, and political information.

My aim here is not a comprehensive survey on the topic but to focus on one of its most appealing applications: using its data to predict the outcome of current[c] and future events.

Such an application is natural in light of the excellent results obtained

a The 2008 Mumbai attacks and 2009 Iranian election protests are perhaps the best-known examples of Twitter playing such a role.
b As of mid-2009, Twitter reportedly had 41.74 million users.[7]
c Bill Tancer of Hitwise said predicting ongoing events should not be defined as "prediction" but rather as "data arbitrage."[13]

**key insights**

- **Using social media to predict future events is a hot research topic involving multiple challenges, including bias in its many forms.**

# What could go wrong?

**Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls**

**Zeynep Tufekci**
University of North Carolina, Chapel Hill
zeynep@unc.edu

**Abstract**

Large-scale databases of human activity in social media have captured scientific and policy attention, producing a flood of research and discussion. This paper considers methodological and conceptual challenges for this emergent field, with special attention to the validity and representativeness of social media big data analyses. Persistent issues include the over-emphasis of a single platform, Twitter, sampling biases arising from selection by hashtags, and vague and unrepresentative sampling frames. The sociocultural complexity of user behavior aimed at algorithmic invisibility (such as subtweeting, mock-retweeting, use of “screen captures” for text, etc.) further complicate interpretation of big data social media. Other challenges include accounting for field effects, i.e. broadly consequential events that do not diffuse only through the network under study but affect the whole society. The application of network methods from other fields to the study of human social activity may not always be appropriate. The paper concludes with a call to action on practical steps to improve our analytic capacity in this promising, rapidly-growing field.

**Introduction**

Very large datasets, commonly referred to as *big data*, have become common in the study of everything from genomes to galaxies, including, importantly, human behavior. Thanks to digital technologies, more and more human activities leave imprints whose collection, storage and aggregation can be readily automated. In particular, the use of social media results in the creation of datasets which may be obtained from platform providers or collected independently with relatively little effort as compared with traditional sociological methods.

*Social media big data* has been hailed as key to crucial insights into human behavior and extensively analyzed by scholars, corporations, politicians, journalists, and governments (Boyd and Crawford 2012; Lazer et al, 2009). Big data reveal fascinating insights into a variety of questions, and allow us to observe social phenomena at a previously unthinkable level, such as the mood oscillations of millions of people in 84 countries (Golder et al., 2011), or in cases where there is arguably no other feasible method of data collection, as with the study of ideological polarization on Syrian Twitter (Lynch, Freelon and Aday, 2014).

The emergence of big data from social media has had impacts in the study of human behavior similar to the introduction of the microscope or the telescope in the fields of biology and astronomy: it has produced a qualitative shift in the scale, scope and depth of possible analysis. Such a dramatic leap requires a careful and systematic examination of its methodological implications, including tradeoffs, biases, strengths and weaknesses.

This paper examines methodological issues and questions of inference from social media big data. Methodological issues including the following: 1. The model organism problem, in which a few platforms are frequently used to generate datasets without adequate consideration of their structural biases. 2. Selecting on dependent variables without requisite precautions; many hashtag analyses, for example, fall in this category. 3. The denominator problem created by vague, unclear or unrepresentative sampling. 4. The prevalence of single platform studies which overlook the wider social ecology of interaction and diffusion.

There are also important questions regarding what we can legitimately infer from online imprints, which are but one aspect of human behavior. Issues include the following: 1. Online actions such as clicks, links, and retweets are complex social interactions with varying meanings, logics and implications, yet they may be aggregated together. 2. Users engage in practices that may be unintelligible to algorithms, such as subtweets (tweets referencing an unnamed but implicitly identifiable individual), quoting text via screen captures, and “hate-linking”—linking to denounce rather than endorse. 3. Network methods from other fields are often used to study human behavior without evaluating their appropriateness. 4. Social media data almost solely captures “node-to-node” interactions, while “field” effects—events that affect a society or a group in a wholesale fashion either through shared experience or through broadcast media—may often account for observed phenomena. 5. Human self-awareness needs to be taken into account; humans will alter behavior because they know they are being observed, and this change in behavior

# What could go wrong?


SOCIAL SCIENCES

## Social media for large studies of behavior

Large-scale studies of human behavior in social media need to be held to higher methodological standards

By Derek Ruths[1]* and Jürgen Pfeffer[2]

On 3 November 1948, the day after Harry Truman won the United States presidential elections, the *Chicago Tribune* published one of the most famous erroneous headlines in newspaper history: "Dewey Defeats Truman" (*1*, *2*). The headline was informed by telephone surveys, which had inadvertently undersampled Truman supporters (*1*). Rather than permanently discrediting the practice of polling, this event led to the development of more sophisticated techniques and higher standards that produce the more accurate and statistically rigorous polls conducted today (*3*).

Now, we are poised at a similar technological inflection point with the rise of online personal and social data for the study of human behavior. Powerful computational resources combined with the availability of massive social media data sets has given rise to a growing body of work that uses a combination of machine learning, natural language processing, network analysis, and statistics for the measurement of population structure and human behavior at unprecedented scale. However, mounting evidence suggests that many of the forecasts and analyses being produced misrepresent the real world (*4*–*6*). Here, we highlight issues that are endemic to the study of human behavior through large-scale social media data sets and discuss strategies that can be used to address them (see the table). Although some of the issues raised are very basic (and long-studied) in the social sciences, the new kinds of data and the entry of a variety of communities of researchers into the field make these issues worth revisiting and updating.

POLICY

**REPRESENTATION OF HUMAN POPULATIONS.** *Population bias.* A common assumption underlying many large-scale social media-based studies of human behavior

different social media platforms (*8*). For instance, Instagram is "especially appealing to adults aged 18 to 29, African-American, Latinos, women, urban residents" (*9*) whereas Pinterest is dominated by females, aged 25 to 34, with an average annual household income of $100,000 (*10*). These sampling biases are rarely corrected for (if even acknowledged).

*Proprietary algorithms for public data.* Platform-specific sampling problems, for example, the highest-volume source of public Twitter data, which are used by thousands of researchers worldwide, is not an accurate representation of the overall platform's data (*11*). Furthermore, researchers are left in the dark about when and how social media providers change the sampling and/or filtering of their data streams. So long as the algorithms and processes that govern these public data releases are largely dynamic, proprietary, and secret or undocumented, designing reliable and reproducible studies of human behavior that correctly account for the resulting biases will be difficult, if not impossible. Academic efforts to characterize aspects of the behavior of such proprietary systems can provide details needed to begin reporting biases.

The rise of "embedded researchers" (researchers who have special relationships with providers that give them elevated access to platform-specific data, algorithms, and resources) is creating a divided social media research community. Such researchers, for example, can see a platform's inner workings and make accommodations, but may not be able to reveal their corrections or the data used to generate their findings.

**REPRESENTATION OF HUMAN BEHAVIOR.** *Human behavior and online platform design.* Many social forces that drive the formation and dynamics of human behavior and relations have been intensively studied and are well-known (*12*–*14*). For instance, homophily ("birds of a feather flock together"), transitivity ("the friend of a friend is a friend"), and propinquity ("those close by form a tie") are all known by designers of social media platforms and, to increase platform use and adoption, have been incorporated in their link suggestion algorithms. Thus, it may be necessary to untangle psychosocial from platform-driven behavior. Unfortunately, few studies attempt this.

Social platforms also implicitly target

**Reducing biases and flaws in social media data**

DATA COLLECTION

- 1. Quantifies platform-specific biases (platform design, user base, platform-specific behavior, platform storage policies)
- 2. Quantifies biases of available data (access constraints, platform-side filtering)
- 3. Quantifies proxy population biases/mismatches

METHODS

- 4. Applies filters/corrects for nonhuman accounts in data
- 5. Accounts for platform and proxy population biases
  a. Corrects for platform-specific and proxy population biases
  OR
  b. Tests robustness of findings
- 6. Accounts for platform-specific algorithms
  a. Shows results for more than one platform
  OR

# What could go wrong?

SOCIAL SCIENCES

## *Social media for large studies of behavior*

Large-scale studies of human behavior in social media need to be held to higher methodological standards

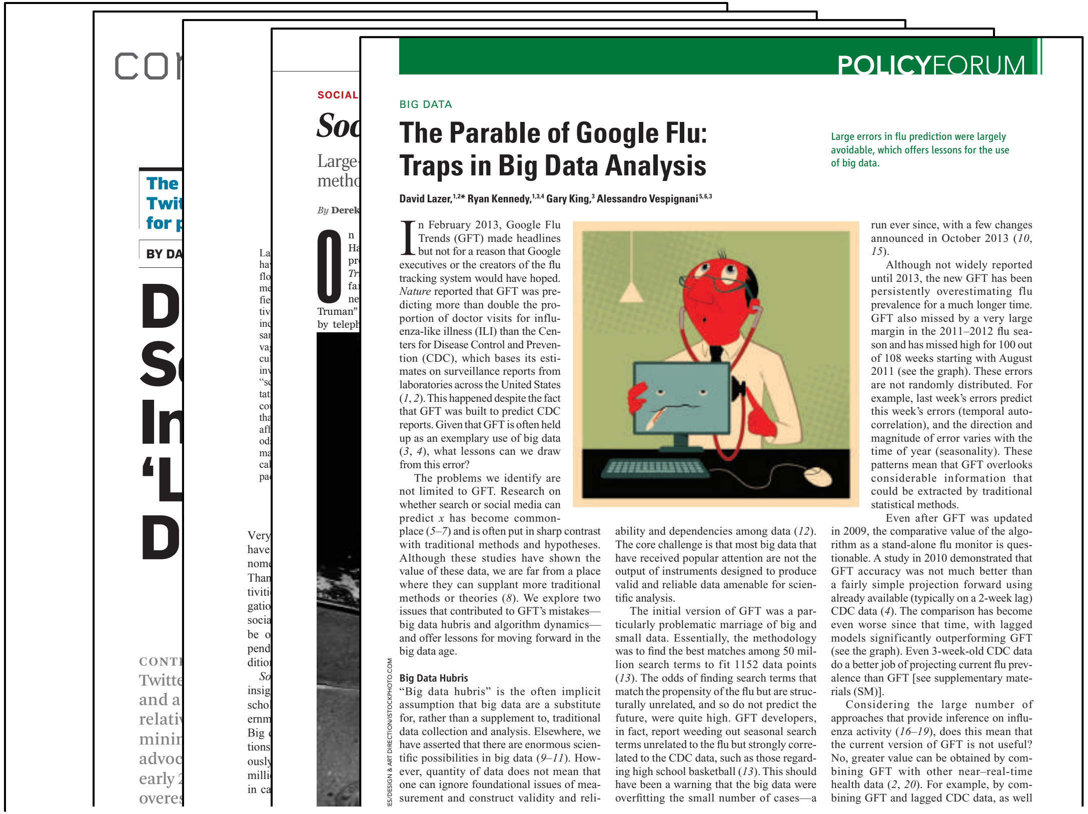
By **Derek Ruths**[1]* and **Jürgen Pfeffer**[2]

On 3 November 1948, the day after Harry Truman won the United States presidential elections, the *Chicago Tribune* published one of the most famous erroneous headlines in newspaper history: "Dewey Defeats Truman" (*1*, *2*). The headline was informed by telephone surveys, which had inadver- different social media platforms (*8*). For instance, Instagram is "especially appealing to adults aged 18 to 29, African-American, Latinos, women, urban residents" (*9*) whereas Pinterest is dominated by females, aged 25 to 34, with an average annual household income of $100,000 (*10*). These sampling biases are rarely corrected for (if even acknowledged).

*Proprietary algorithms for public data.* Platform-specific sampling problems, for example, the highest-volume source of pub-

The rise of "embedded researchers" (researchers who have special relationships with providers that give them elevated access to platform-specific data, algorithms, and resources) is creating a divided social media research community. Such researchers, for example, can see a platform's inner workings and make accommodations, but may not be able to reveal their corrections or the data used to generate their findings.

# What could go wrong?



POLICYFORUM

BIG DATA

## The Parable of Google Flu: Traps in Big Data Analysis

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

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

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

The problems we identify are not limited to GFT. Research on whether search or social media can predict *x* has become commonplace (*5*–*7*) and is often put in sharp contrast with traditional methods and hypotheses. Although these studies have shown the value of these data, we are far from a place where they can supplant more traditional methods or theories (*8*). We explore two issues that contributed to GFT's mistakes—big data hubris and algorithm dynamics—and offer lessons for moving forward in the big data age.

**Big Data Hubris**

"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. Elsewhere, we have asserted that there are enormous scientific possibilities in big data (*9*–*11*). However, quantity of data does not mean that one can ignore foundational issues of measurement and construct validity and reliability and dependencies among data (*12*). The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis.

The initial version of GFT was a particularly problematic marriage of big and small data. Essentially, the methodology was to find the best matches among 50 million search terms to fit 1152 data points (*13*). The odds of finding search terms that match the propensity of the flu but are structurally unrelated, and so do not predict the future, were quite high. GFT developers, in fact, report weeding out seasonal search terms unrelated to the flu but strongly correlated to the CDC data, such as those regarding high school basketball (*13*). This should have been a warning that the big data were overfitting the small number of cases—a

ES/DESIGN & ART DIRECTION/ISTOCKPHOTO.COM

run ever since, with a few changes announced in October 2013 (*10*, *15*).

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

Even after GFT was updated in 2009, the comparative value of the algorithm as a stand-alone flu monitor is questionable. A study in 2010 demonstrated that GFT accuracy was not much better than a fairly simple projection forward using already available (typically on a 2-week lag) CDC data (*4*). The comparison has become even worse since that time, with lagged models significantly outperforming GFT (see the graph). Even 3-week-old CDC data do a better job of projecting current flu prevalence than GFT [see supplementary materials (SM)].

Considering the large number of approaches that provide inference on influenza activity (*16*–*19*), does this mean that the current version of GFT is not useful? No, greater value can be obtained by combining GFT with other near–real-time health data (*2*, *20*). For example, by combining GFT and lagged CDC data, as well

# What could go wrong?

Commentary

BIG DATA & SOCIETY

Big Data & Society
July–December 2015: 1–4
© The Author(s) 2015
DOI: 10.1177/2053951715602495
bds.sagepub.com
SAGE

## Big Data and the danger of being precisely inaccurate

Daniel A McFarland and H Richard McFarland

**Abstract**

Social scientists and data analysts are increasingly making use of Big Data in their analyses. These data sets are often "found data" arising from purely observational sources rather than data derived under strict rules of a statistically designed experiment. However, since these large data sets easily meet the sample size requirements of most statistical procedures, they give analysts a false sense of security as they proceed to focus on employing traditional statistical methods. We explain how most analyses performed on Big Data today lead to "precisely inaccurate" results that hide biases in the data but are easily overlooked due to the enhanced significance of the results created by the data size. Before any analyses are performed on large data sets, we recommend employing a simple data segmentation technique to control for some major components of observational data biases. These segments will help to improve the accuracy of the results.

**Keywords**

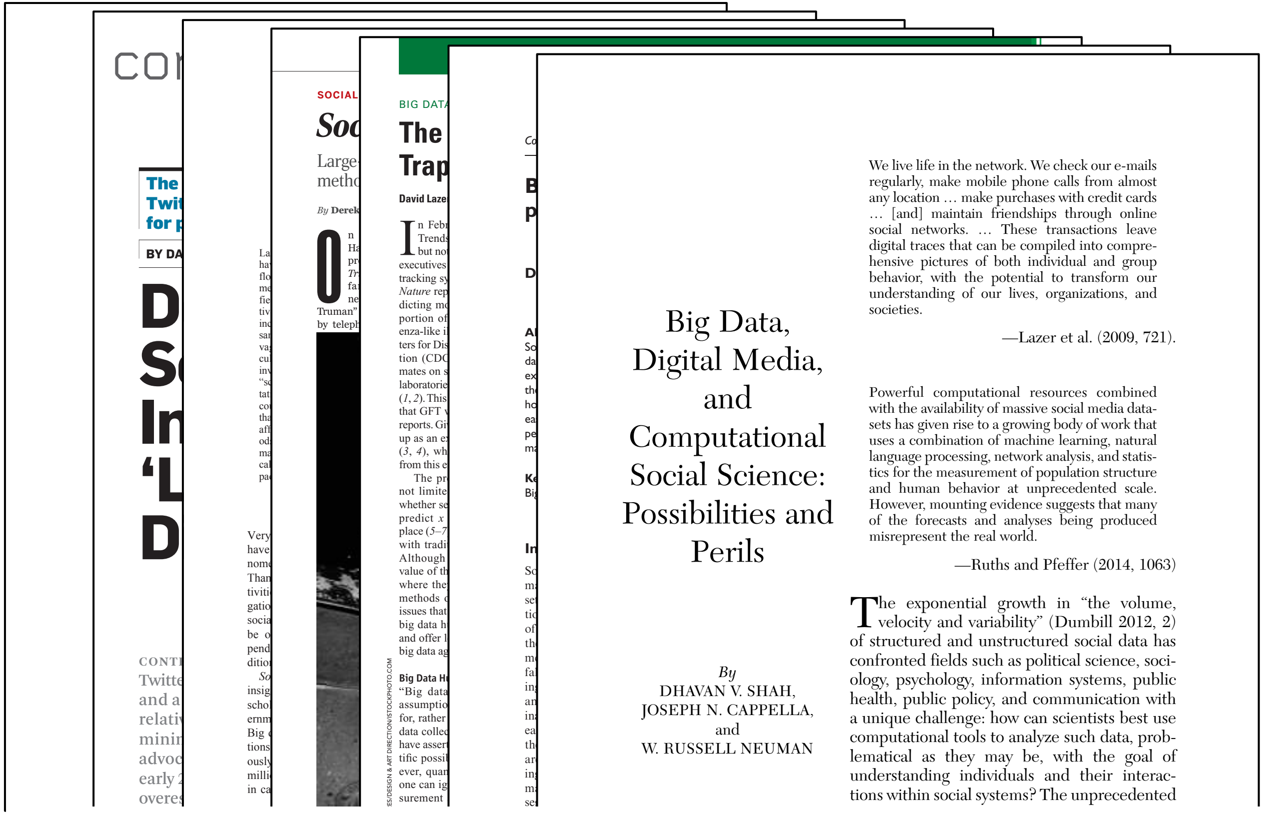
Big Data, bias, segmentation, sociology, statistics, inaccuracy

### Introduction

Social scientists and data analysts are increasingly making use of Big Data in their analyses. These data sets are often "found data"[1] arising from purely observational sources rather than data derived under strict rules of a statistically designed experiment. However, since these large data sets easily meet the sample size requirements of most statistical procedures, they give analysts a false sense of security as they proceed to focus on employing traditional statistical methods. We explain how most analyses performed on Big Data today lead to "precisely inaccurate" results that hide biases in the data but are easily overlooked due to the enhanced significance of the results created by the data size. Before any analyses are performed on large data sets, we recommend employing data segmentation techniques to control for some major components of observational data biases. These segments will help improve the accuracy of results.

data collected on website traffic, sensor data, or any large-scale source of user activity. This data is often labeled as "big" because it can easily contain many millions of records reflecting user behaviors on a website such as viewing, clicking, downloading, uploading, evaluating, and purchasing of digital resources. In most cases, these data sets are snapshots of time that are collected on an entire sample of individuals who are active in that particular moment. Examples of this sort of data are website log files or traffic data, social media data dumps (e.g. Twitter, Facebook, LinkedIn, etc.), online professional networks (e.g. where teachers learn about jobs and post teaching resources), or even massive open online courses with user interactions and performance (e.g. Coursera). As cell phones and wearable devices begin to collect sensor data, Big Data will only get bigger and the problem referenced in this article will

# What could go wrong?



Big Data, Digital Media, and Computational Social Science: Possibilities and Perils

*By*
DHAVAN V. SHAH,
JOSEPH N. CAPPELLA,
and
W. RUSSELL NEUMAN

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location … make purchases with credit cards … [and] maintain friendships through online social networks. … These transactions leave digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

—Lazer et al. (2009, 721).

Powerful computational resources combined with the availability of massive social media datasets has given rise to a growing body of work that uses a combination of machine learning, natural language processing, network analysis, and statistics for the measurement of population structure and human behavior at unprecedented scale. However, mounting evidence suggests that many of the forecasts and analyses being produced misrepresent the real world.

—Ruths and Pfeffer (2014, 1063)

The exponential growth in “the volume, velocity and variability” (Dumbill 2012, 2) of structured and unstructured social data has confronted fields such as political science, sociology, psychology, information systems, public health, public policy, and communication with a unique challenge: how can scientists best use computational tools to analyze such data, problematical as they may be, with the goal of understanding individuals and their interactions within social systems? The unprecedented

Carnegie Mellon University
School of Computer Science



# What could go wrong?

Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites

By ESZTER HARGITTAI

This article discusses methodological challenges of using big data that rely on specific sites and services as their sampling frames, focusing on social network sites in particular. It draws on survey data to show that people do not select into the use of such sites randomly. Instead, use is biased in certain ways yielding samples that limit the generalizability of findings. Results show that age, gender, race/ethnicity, socioeconomic status, online experiences, and Internet skills all influence the social network sites people use and thus where traces of their behavior show up. This has implications for the types of conclusions one can draw from data derived from users of specific sites. The article ends by noting how big data studies can address the shortcomings that result from biased sampling frames.

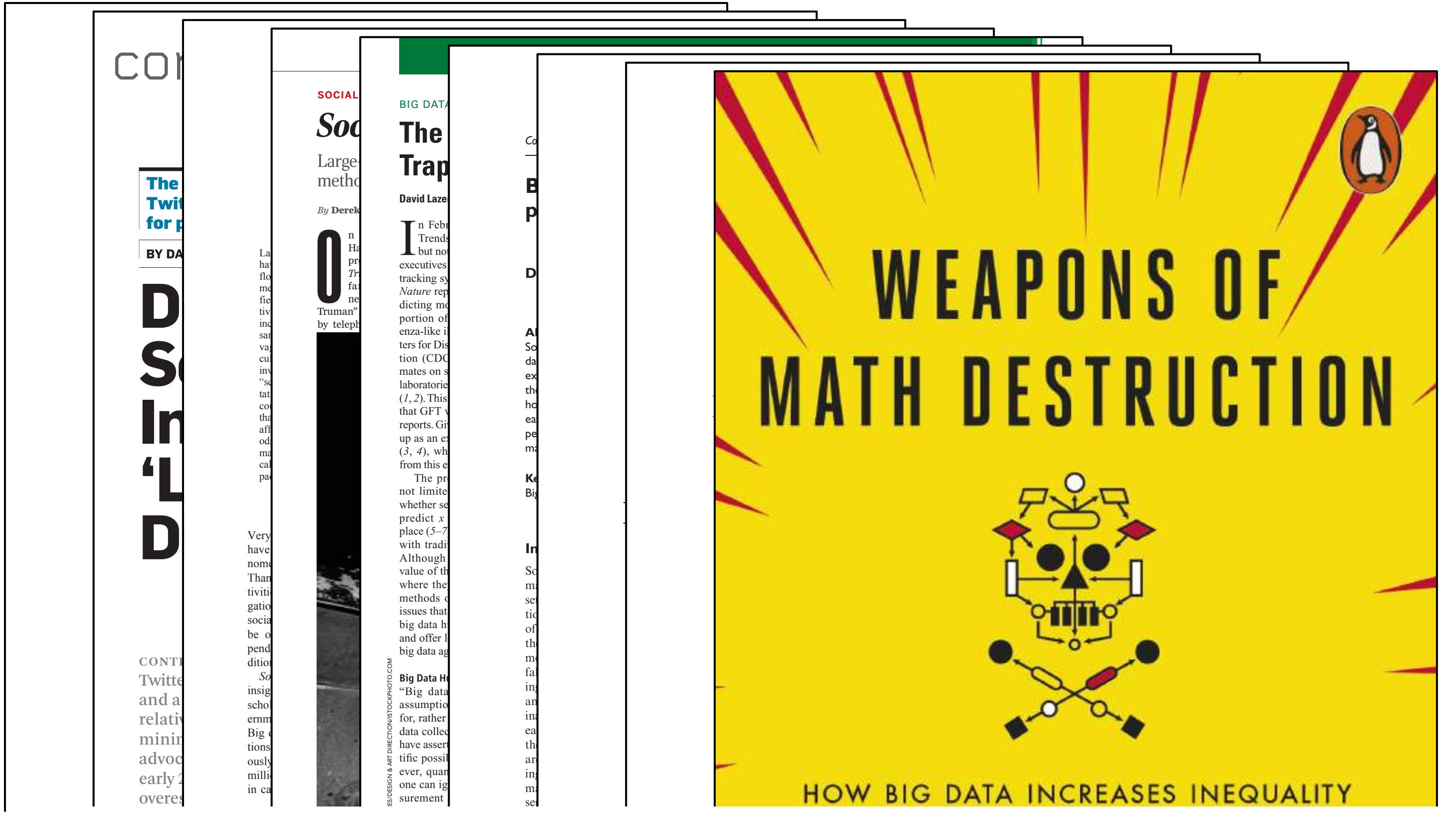
*Keywords:* big data; Internet skills; digital inequality; social network sites; sampling frame; biased sample; sampling

As people incorporate digital media into increasing parts of their everyday lives, a growing number of their actions leave digital traces. This information is available to businesses, government agencies, and beyond. Researchers have analyzed such large-scale trace data to address a myriad of social behavioral questions from the political (e.g., Tumasjan

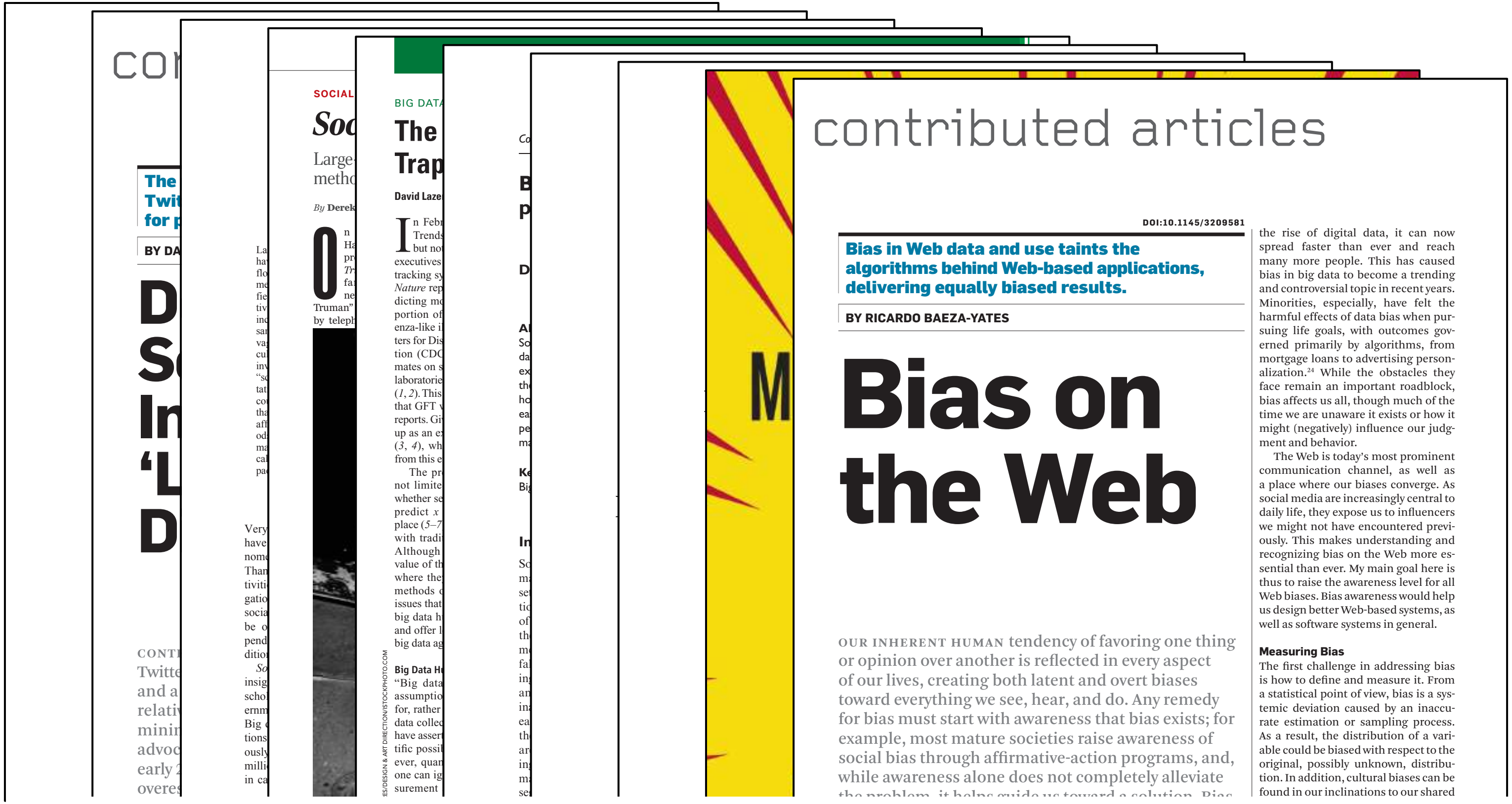
*Eszter Hargittai is Delaney Family Professor of Communication Studies and faculty associate of the Institute for Policy Research at Northwestern University, where she heads the Web Use Project.*

NOTE: The author greatly appreciates the generous support of the John D. and Catherine T. MacArthur

# What could go wrong?



# What could go wrong?


contributed articles

DOI:10.1145/3209581

**Bias in Web data and use taints the algorithms behind Web-based applications, delivering equally biased results.**

BY RICARDO BAEZA-YATES

# Bias on the Web

OUR INHERENT HUMAN tendency of favoring one thing or opinion over another is reflected in every aspect of our lives, creating both latent and overt biases toward everything we see, hear, and do. Any remedy for bias must start with awareness that bias exists; for example, most mature societies raise awareness of social bias through affirmative-action programs, and, while awareness alone does not completely alleviate

the rise of digital data, it can now spread faster than ever and reach many more people. This has caused bias in big data to become a trending and controversial topic in recent years. Minorities, especially, have felt the harmful effects of data bias when pursuing life goals, with outcomes governed primarily by algorithms, from mortgage loans to advertising personalization.[24] While the obstacles they face remain an important roadblock, bias affects us all, though much of the time we are unaware it exists or how it might (negatively) influence our judgment and behavior.

The Web is today's most prominent communication channel, as well as a place where our biases converge. As social media are increasingly central to daily life, they expose us to influencers we might not have encountered previously. This makes understanding and recognizing bias on the Web more essential than ever. My main goal here is thus to raise the awareness level for all Web biases. Bias awareness would help us design better Web-based systems, as well as software systems in general.

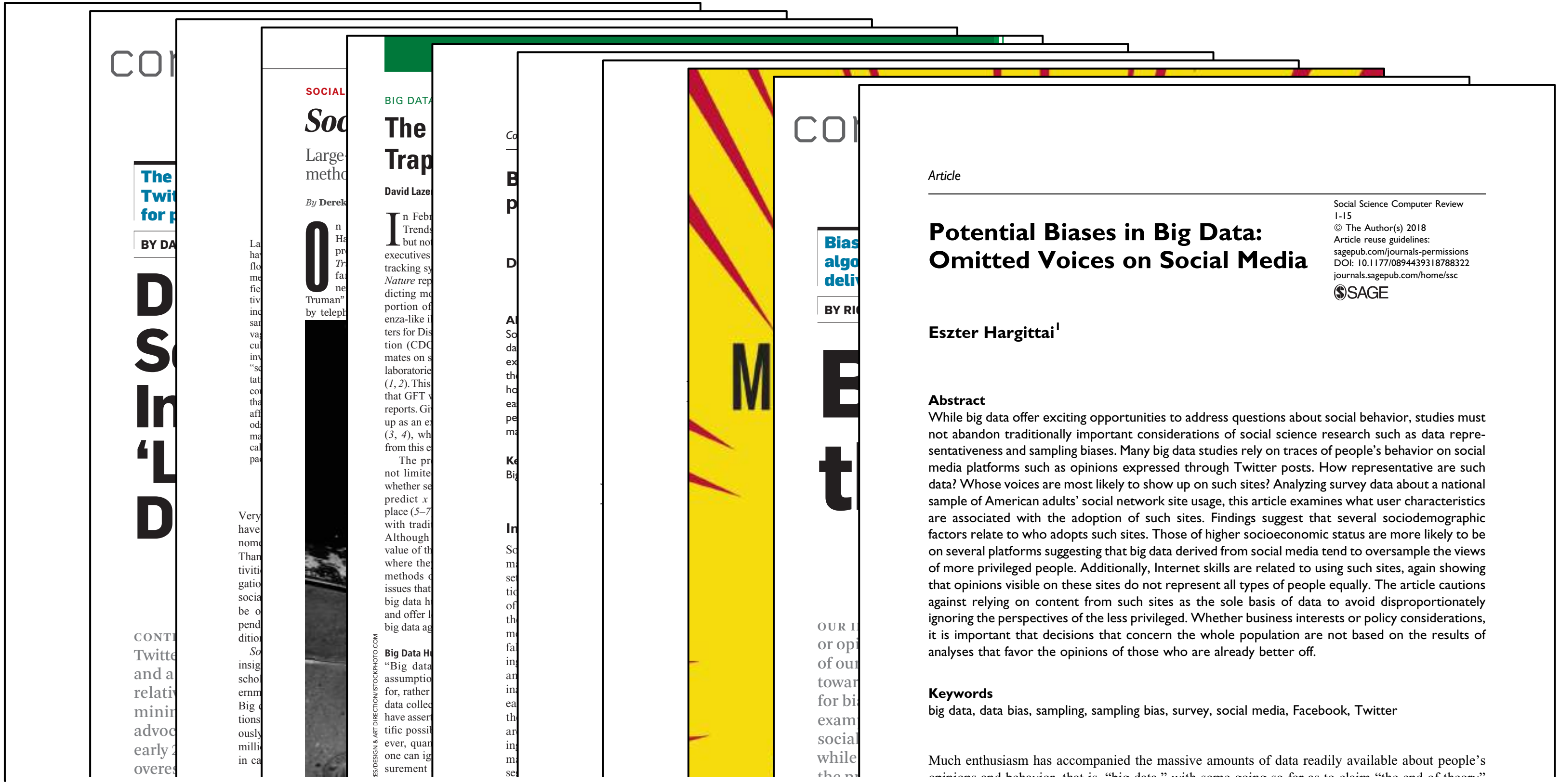
**Measuring Bias**

The first challenge in addressing bias is how to define and measure it. From a statistical point of view, bias is a systemic deviation caused by an inaccurate estimation or sampling process. As a result, the distribution of a variable could be biased with respect to the original, possibly unknown, distribution. In addition, cultural biases can be found in our inclinations to our shared

Carnegie Mellon University
School of Computer Science



# What could go wrong?



*Article*

Social Science Computer Review
1-15
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0894439318788322
journals.sagepub.com/home/ssc
SAGE

**Potential Biases in Big Data: Omitted Voices on Social Media**

**Eszter Hargittai**[1]

**Abstract**

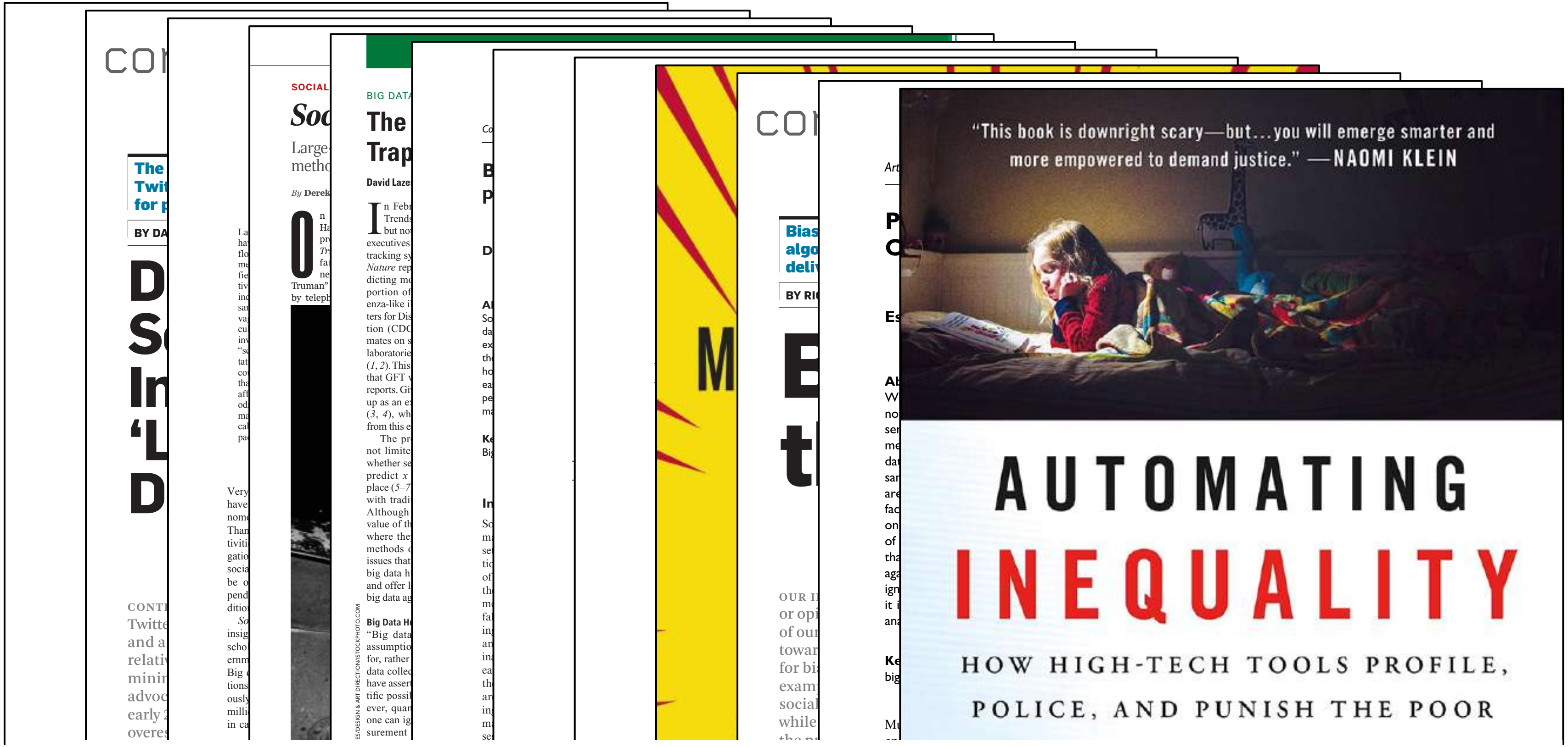
While big data offer exciting opportunities to address questions about social behavior, studies must not abandon traditionally important considerations of social science research such as data representativeness and sampling biases. Many big data studies rely on traces of people's behavior on social media platforms such as opinions expressed through Twitter posts. How representative are such data? Whose voices are most likely to show up on such sites? Analyzing survey data about a national sample of American adults' social network site usage, this article examines what user characteristics are associated with the adoption of such sites. Findings suggest that several sociodemographic factors relate to who adopts such sites. Those of higher socioeconomic status are more likely to be on several platforms suggesting that big data derived from social media tend to oversample the views of more privileged people. Additionally, Internet skills are related to using such sites, again showing that opinions visible on these sites do not represent all types of people equally. The article cautions against relying on content from such sites as the sole basis of data to avoid disproportionately ignoring the perspectives of the less privileged. Whether business interests or policy considerations, it is important that decisions that concern the whole population are not based on the results of analyses that favor the opinions of those who are already better off.

**Keywords**

big data, data bias, sampling, sampling bias, survey, social media, Facebook, Twitter

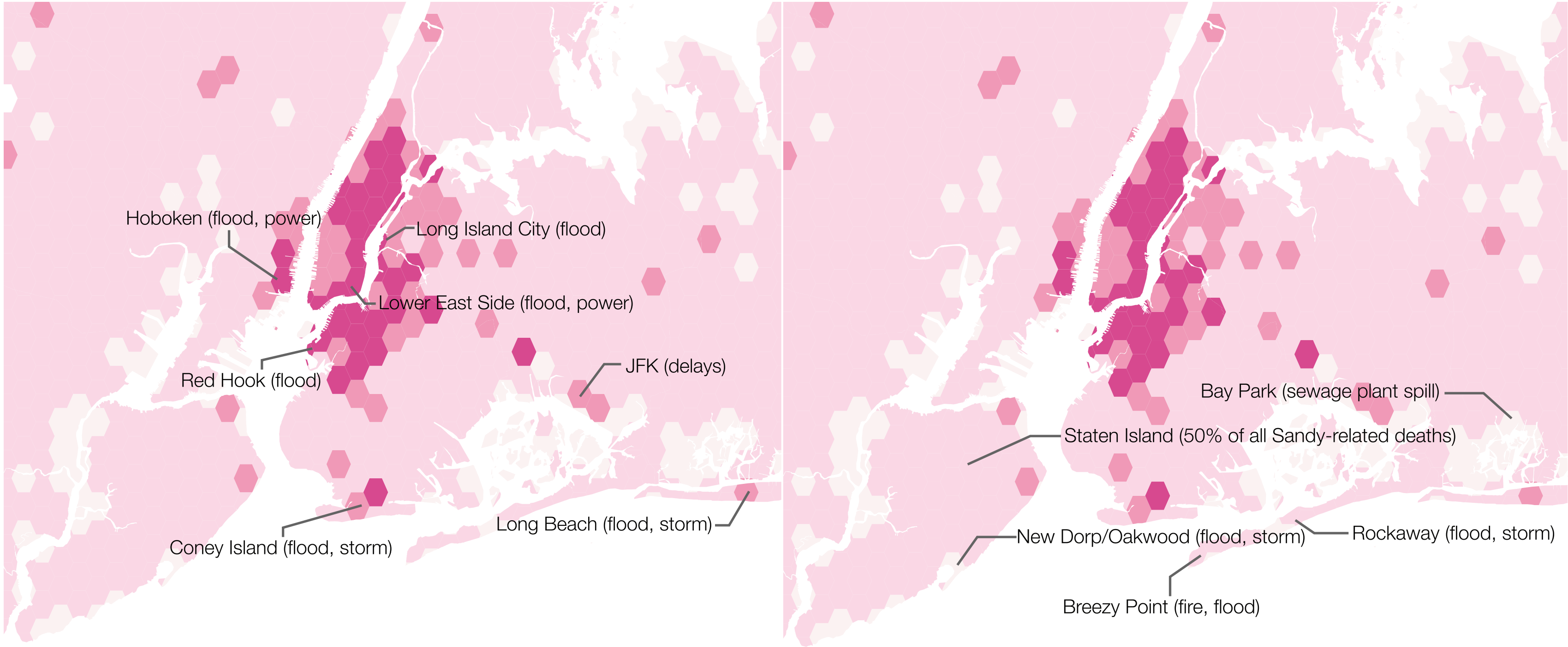
Much enthusiasm has accompanied the massive amounts of data readily available about people's

# What could go wrong?



# Misdirected resources

## Hurricane Sandy, tweets vs. damage/deaths




Taylor Shelton, Ate Poorthuis, Mark Graham, and Matthew Zook (2014). Mapping the data shadows of Hurricane Sandy. *Geoforum* 52, 167–179.

# Unjust targeting

**Identifying Networks of Criminals**

"Facebook has helped me by identifying suspects that were friends or associates of other suspects in a crime and all brought in and interviewed and later convicted of theft and drug offenses."

"My biggest use for social media has been to locate and identify criminals. I have started to utilize it to piece together local drug networks."

| | |
|---|---|
| **Goal:** | **Predict, Monitor, and Prevent Risk In/Around Protests** |
| Anticipated Activity: | Protests, Riots, Looting |
| Overt Threats: | Unions, Activist Groups, Etc. |
| Locations: | Schools, Public Spaces, Malls, High-Rent Districts |
| Actions Taken: | During Event(s), Post-Event |

**Geofeedia**

LexisNexis® Risk Solutions (2014). Survey of law enforcement personnel and their use of social media.

Nicole Ozer (2016). Police use of social media surveillance software is escalating, and activists are in the digital crosshairs. ACLU of Northern CA. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/police-use-social-media-surveillance-software

# Enabling exclusion and expulsion

WIRED

Privacy

## Europe is using smartphone data as a weapon to deport refugees

European leaders need to bring immigration numbers down, and metadata on smartphones could be just what they need to start sending migrants back

*By* **MORGAN MEAKER**

*02 Jul 2018*

# Surveillance for punishment

## On Their Last Legs

SMARTPHONES SHOULD REPLACE GPS ANKLE BRACELETS FOR MONITORING OFFENDERS +++ BY ROBERT S. GABLE

**IMAGINE THIS**: It's early morning, and you're sleeping alone in your bed. Suddenly your ankle vibrates, and a voice blurts out from beneath the sheets: "This is the monitoring center. You are not in your inclusion zone. Do you have permission to be outside this area?" ¶ That's what happened to a man named Jeffrey B. when his GPS-equipped ankle bracelet went berserk. The California Department of Corrections and Rehabilitation had strapped a tracking anklet on Jeffrey for good reason. He had pleaded guilty to 26 counts of peeping into windows and video recording young women while they were undressing. After three

*IEEE Spectrum*, August 2017.

# Identify and avoid risks to achieve positive outcomes

# Thesis

*Social media and sensor data are biased.*

*But we can identify, study and understand the forms of bias.*

*Once we understand these, we can identify <u>scopes</u> within which findings are meaningful and robust.*

# Outline

**Part I: Critiques**

Chapter 1. Demographic biases
Chapter 2. Platform effects
Chapter 3. Sensors and social networks

**Part II: Responses**

(Thesis)
Chapter 4. Social media for public health outreach
Chapter 5. Mobile phone sensors for cohort studies

# Contributions

**Part I: Critiques**

Chapter 1. First national, multivariate, spatial model of Census and social media data

Chapter 2. First empirical demonstration of theorized platform effects

Chapter 3. First theorization of sensor data for social networks

**Part II: Responses**

Chapter 4. Demonstration of rigorous use of Twitter for public health

Chapter 5. First sensor study to combine rigorously validated modeling and social network theory

# Part I: Critiques


YOU KEEP ON USING THESE DATA
I DO NOT THINK THEY MEAN WHAT YOU THINK THEY MEAN

# Chapter 1: Demographic biases

with Hemank Lamba, Constantine Nakos, Jürgen Pfeffer

# What “geotagged tweets” are


https://api.twitter.com/1.1/statuses/show/12345678901234567 8.json


```
{
    "created_at": "Wed Apr 01 00:47:05
                   +00002015",
    "text": "This view tho
             \uE106\uE00E,
    "user": {
        "followers_count": 36000,
        "friends_count": 25000,
        "geo_enabled": true,
    },
    "geo": {
        "type": "Point",
        "coordinates":
        [36.11570625,-115.17407114]
    }
}
```

Carnegie Mellon University
School of Computer Science

Introduction

Part I: Critiques

1. Demographic biases

2. Platform effects

3. Sensors and social networks

Part II: Responses

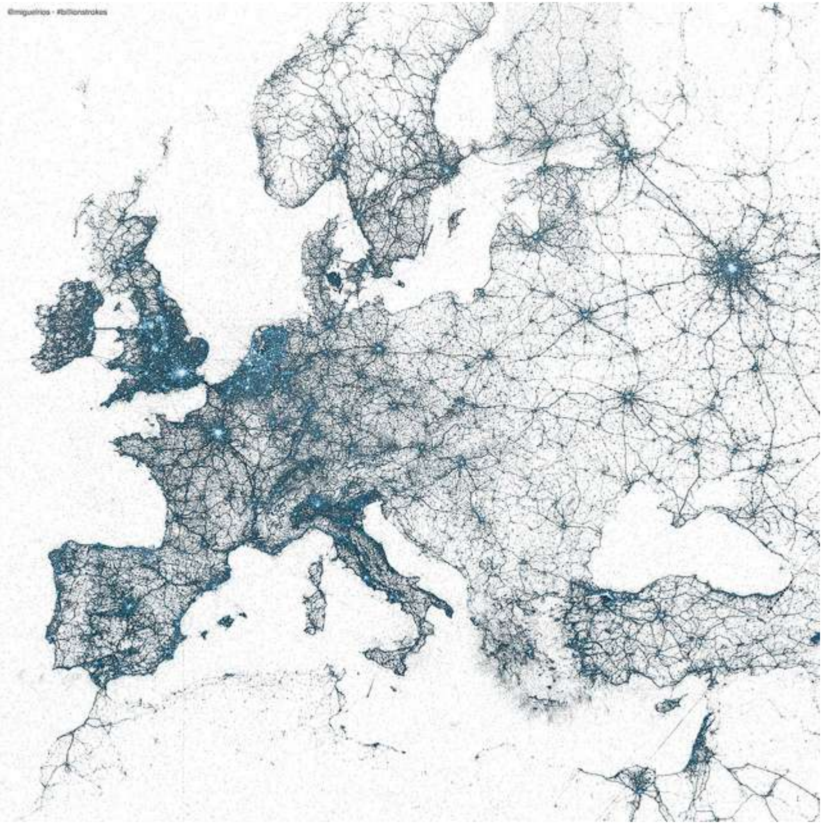
(Thesis)

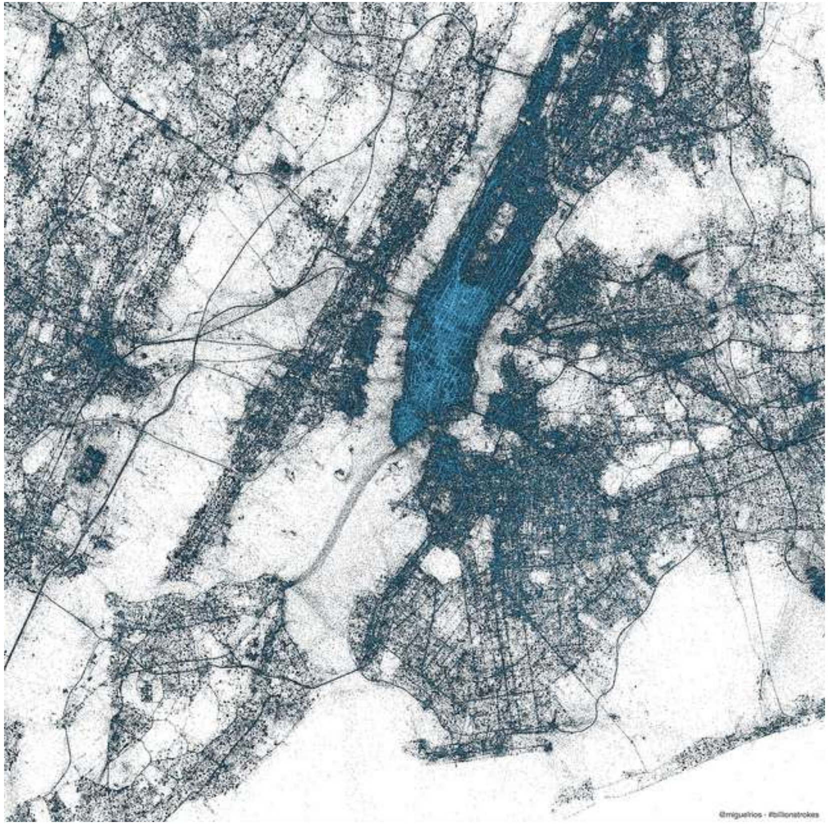
4. Public health outreach

5. Mobile phone sensors and cohorts

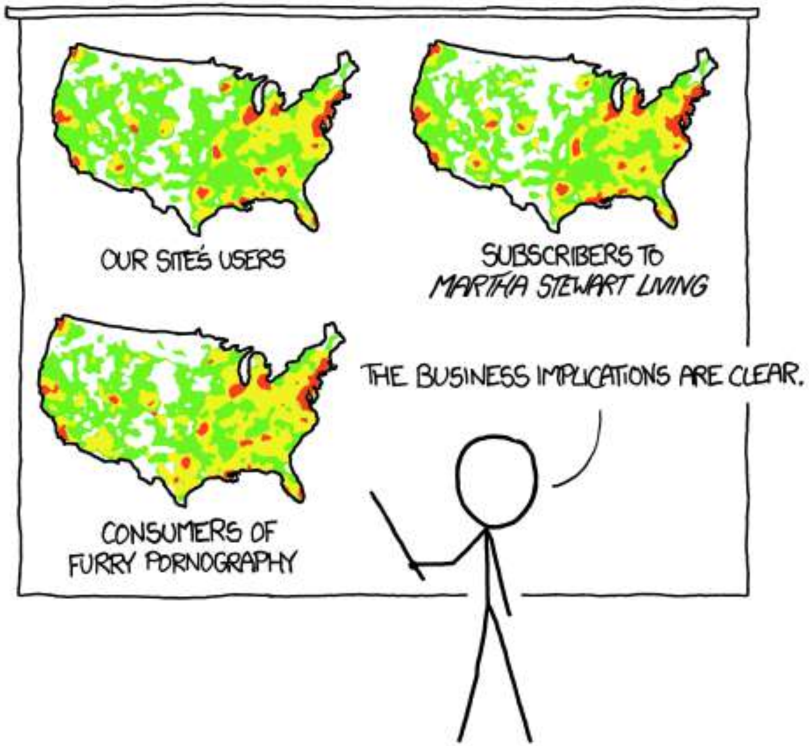
Conclusion

# Geotagged tweets have amazing detail



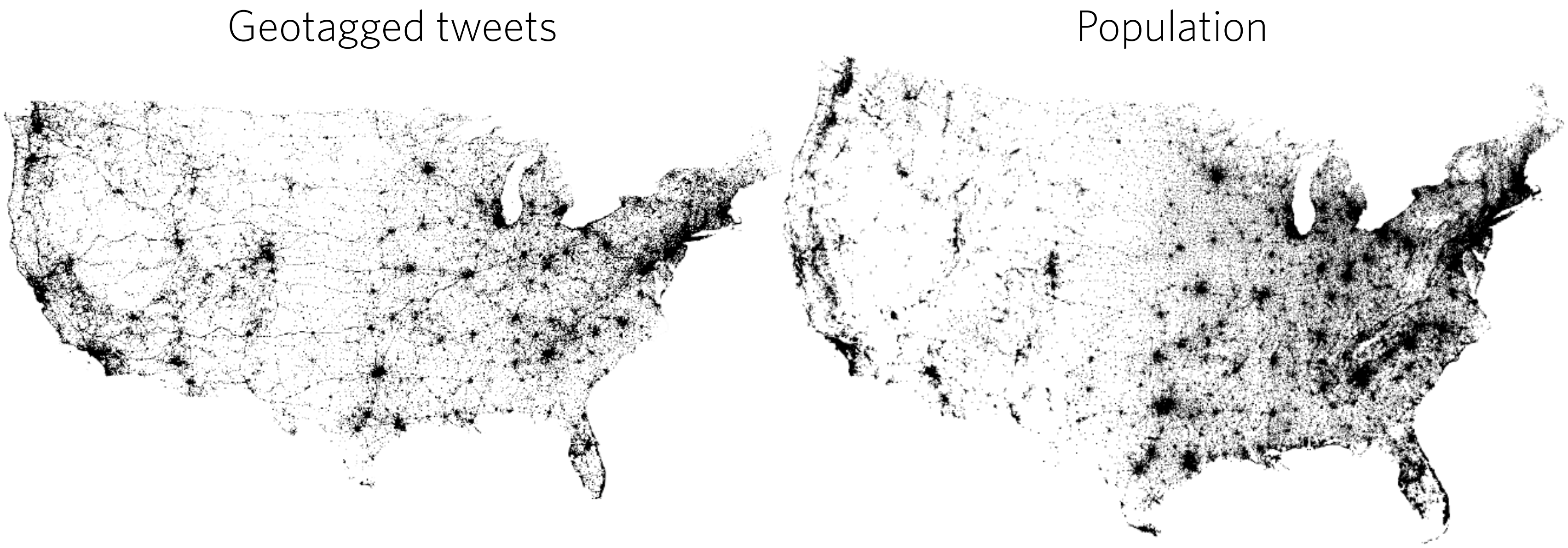


# But maps can be misleading




xkcd (2012). Heatmap. https://xkcd.com/1138/

# How do tweets and population relate?




Adapted from Eric Fischer (2009), Contiguous United States geotag map. https://flic.kr/p/a7WMWS.

Population density in 2010 US Census. Each square represents 1,000 people. Adapted from Geography Division, U.S. Department of Commerce / Economics and Statistics Administration / U.S. Census Bureau, Nighttime Population Distribution Wall Map.

# It matters how well they agree

- Geotagged tweets used to study mobility, urban life, transportation, natural disaster crisis response, public health, and more
- Null hypothesis: users of geotagged tweets are distributed randomly over the US population

# Model users over geographic units

- Users, and noise, proportional to population: $U_i = \alpha P_i + \varepsilon_i P_i$. Take a log transformation,

$$\log U_i = \log \alpha + \log P_i + \varepsilon_i'.$$

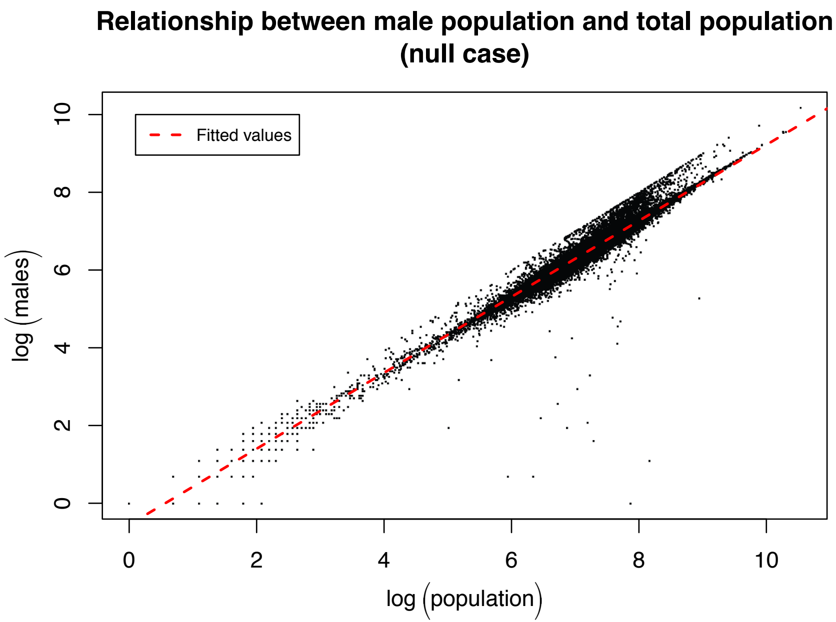
- For linear model

$$\log U_i = \beta_0 + \beta_1 \log P_i + \varepsilon_i',$$

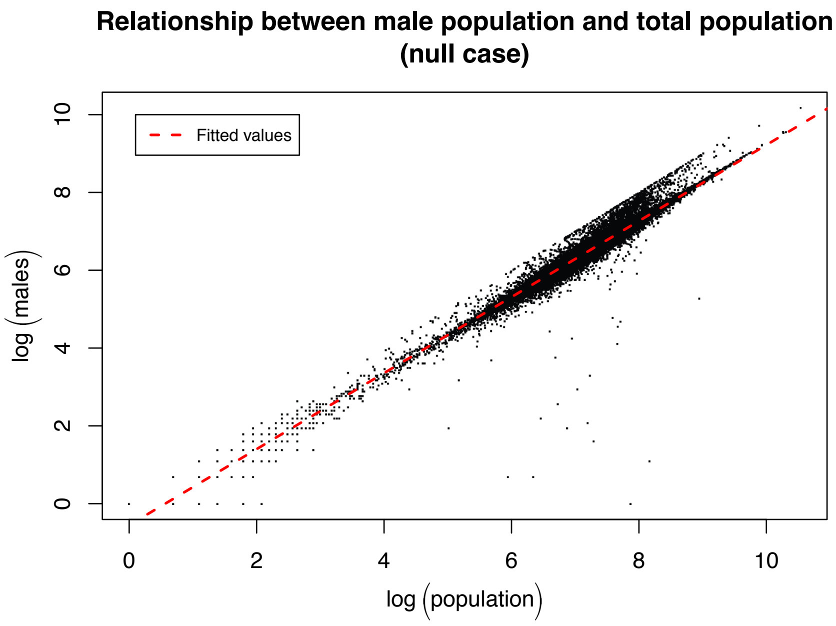
- We get $H_0$: $\beta_1 = 1$.

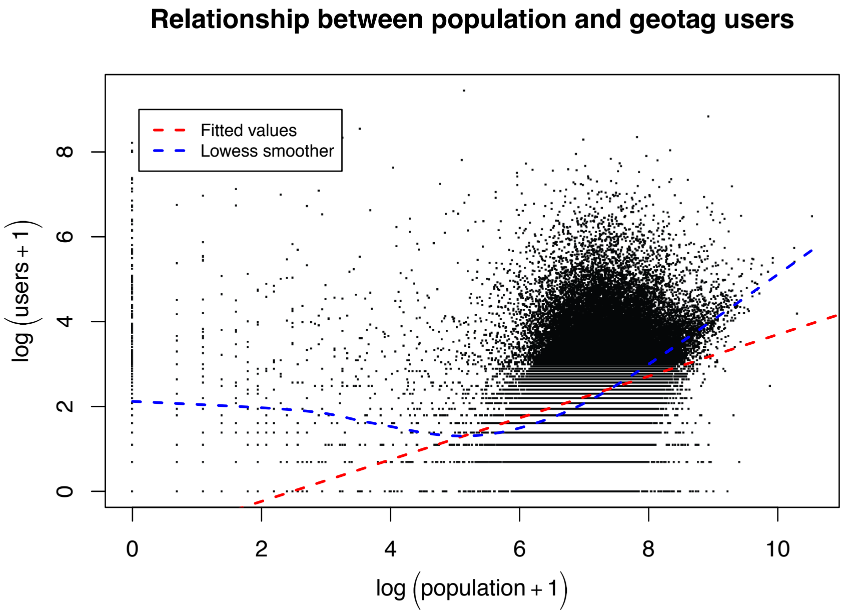
# Distribution of males validates the model





# Geotagged tweets: *not* evenly distributed



# We can identify other differences

- Spatial multivariate modeling reveals specific biases
  - ⬇ Rural, poor, elderly, non-coastal
  - ⬆ Asian, Hispanic, black
- …but these are only the demographics we can access. E.g., harassment of women on Twitter likely discourages geotag use

# Lesson: Geotagged tweets may not generalize

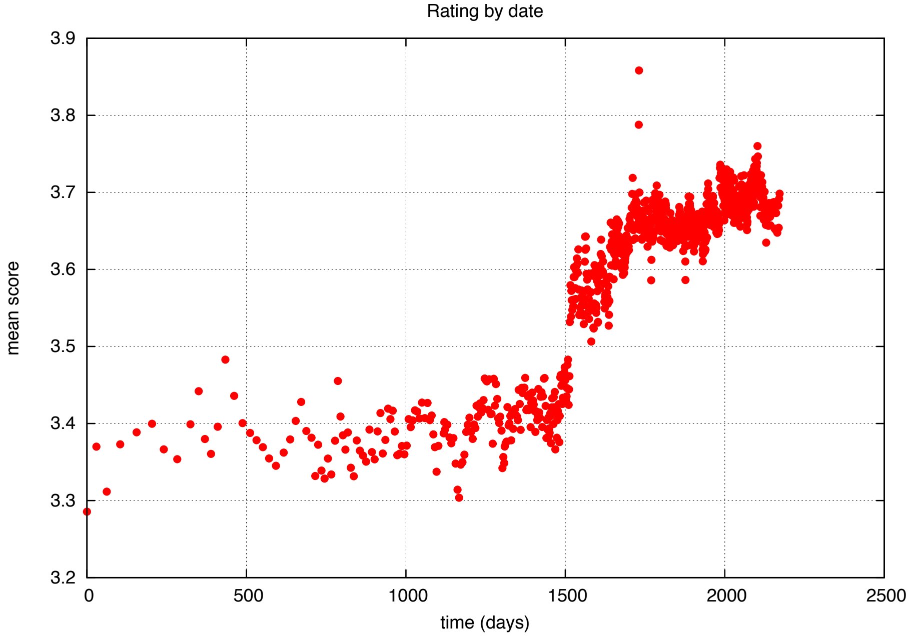
- Don’t use for critical applications without verification!
- Think about other ways to make use of them
- Tasse et al., 2017: “geotags are postcards, not ticket stubs”

Dan Tasse, Zichen Liu, Alex Sciuto, and Jason I. Hong (2017). State of the geotags: Motivations and recent changes. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media* (ICWSM 2017), 250–259.

# Chapter 2: Platform effects

with Jürgen Pfeffer

# Design can cause/change behavior




Average Netflix movie ratings over time. Each point averages 100,000 rating instances.

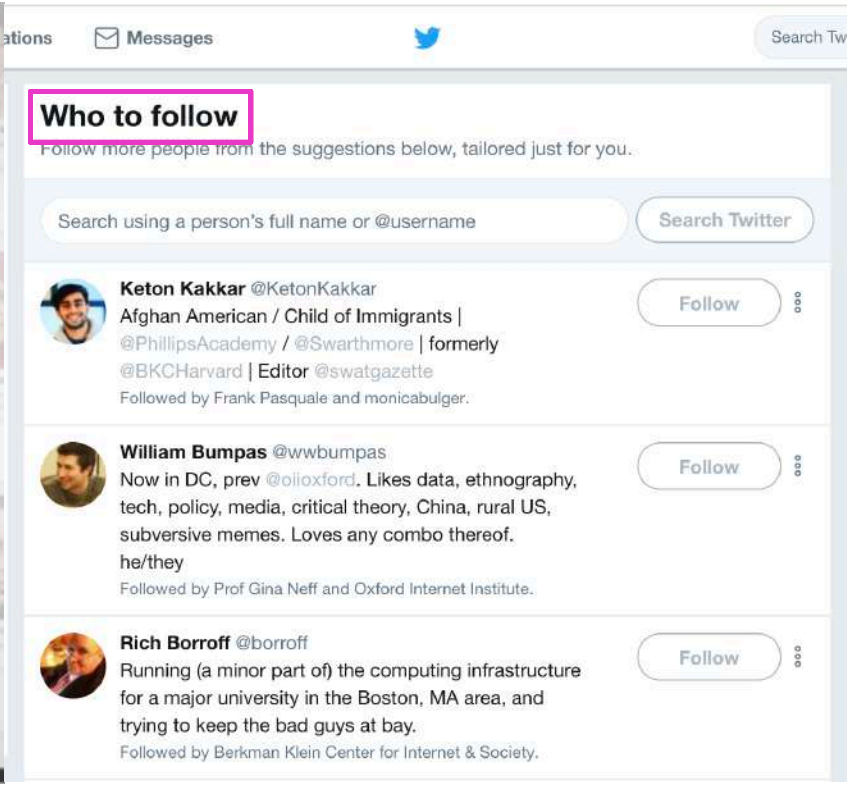
Yehuda Koren (2009). The BellKor solution to the Netflix Grand Prize.

# Social media platforms are businesses

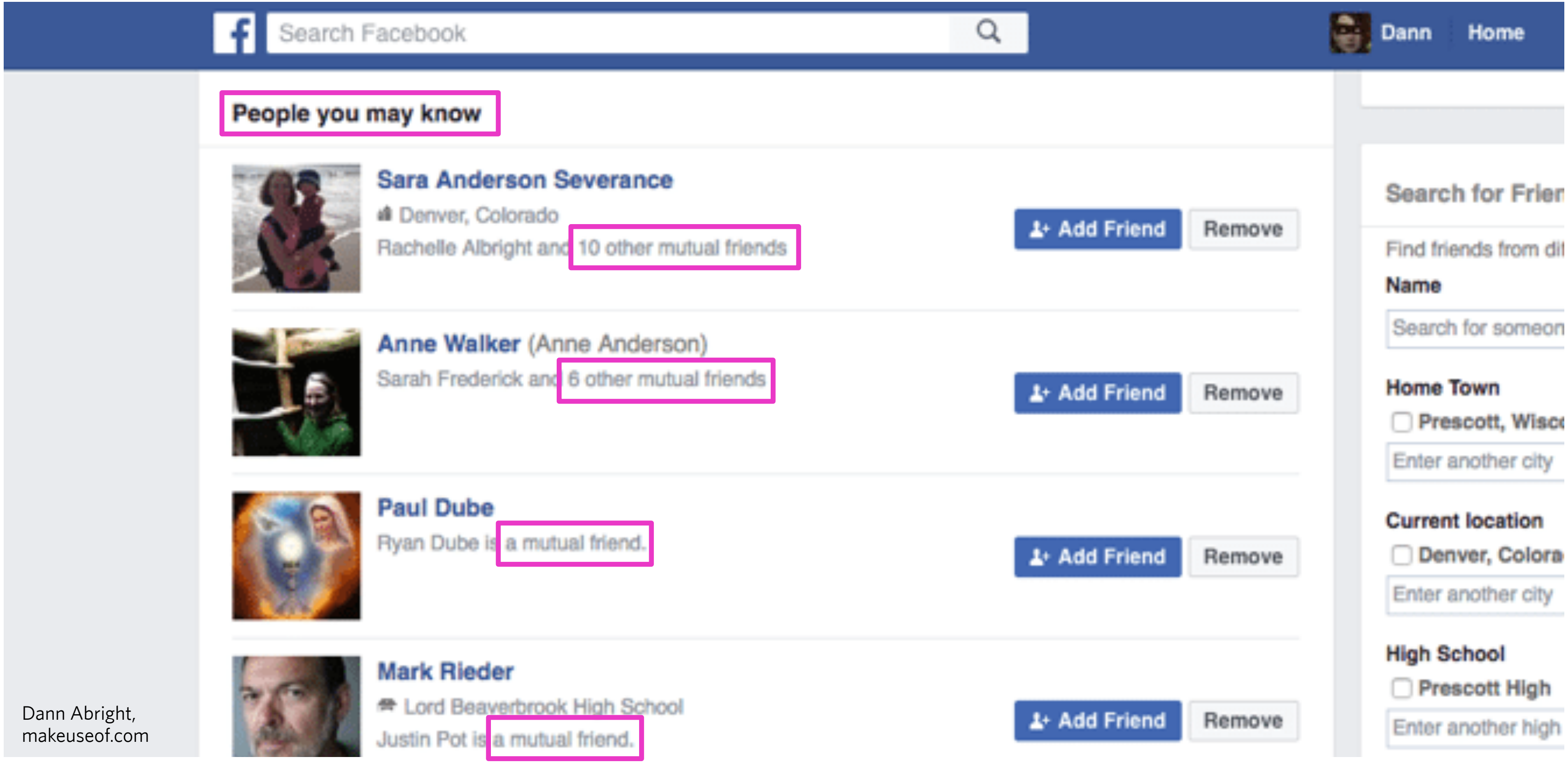

Markets Insider, Business Insider (2018)

- Not neutral utilities or research environments
- Platform engineers try to shape user behavior towards desirable ends

# Sites try to grow their users' networks

# Often through "friend-of-a-friend"



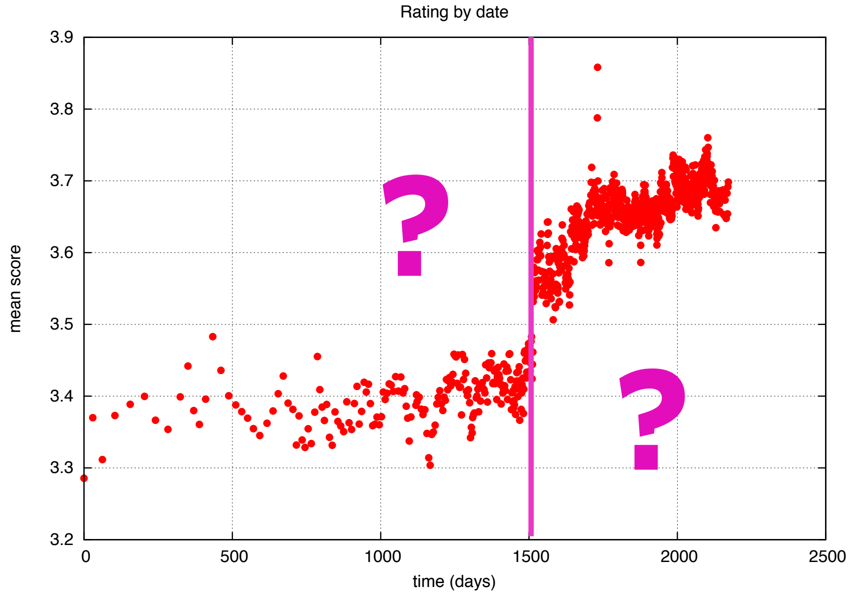

Dann Abright, makeuseof.com

Carnegie Mellon University School of Computer Science



# How do we separate out platform effects?

- When we measure behavior, what are we really measuring? People's behavior, or platform effects?
- How, as outsiders, can we find out?




Average Netflix movie ratings over time. Each point averages 100,000 rating instances.

Carnegie Mellon University
School of Computer Science



# *Data artifacts* can reveal inner workings

The Matrix (1999) “déjà vu” scene

# Data artifacts as natural experiments

- Regression Discontinuity (RD) Design or Interrupted Time Series (ITS) estimate causality

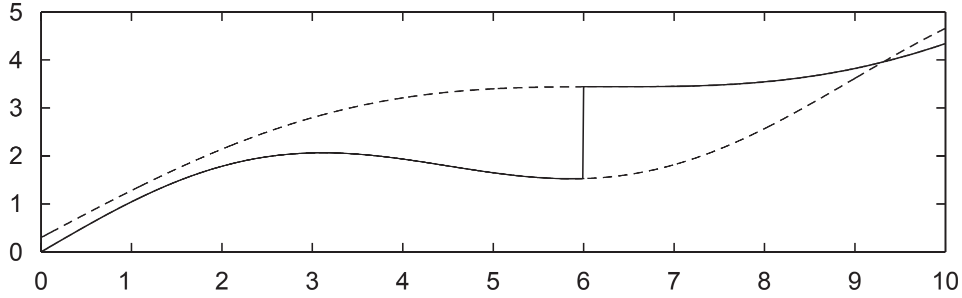



Fig. 2 from Imbens and Lemieux (2008): Potential and observed outcome regression functions.

- The difference between “before” and “after” estimates the *local average treatment effect*

Carnegie Mellon University
School of Computer Science

Introduction

Part I: Critiques

1. Demographic biases

2. Platform effects

3. Sensors and social networks

Part II: Responses

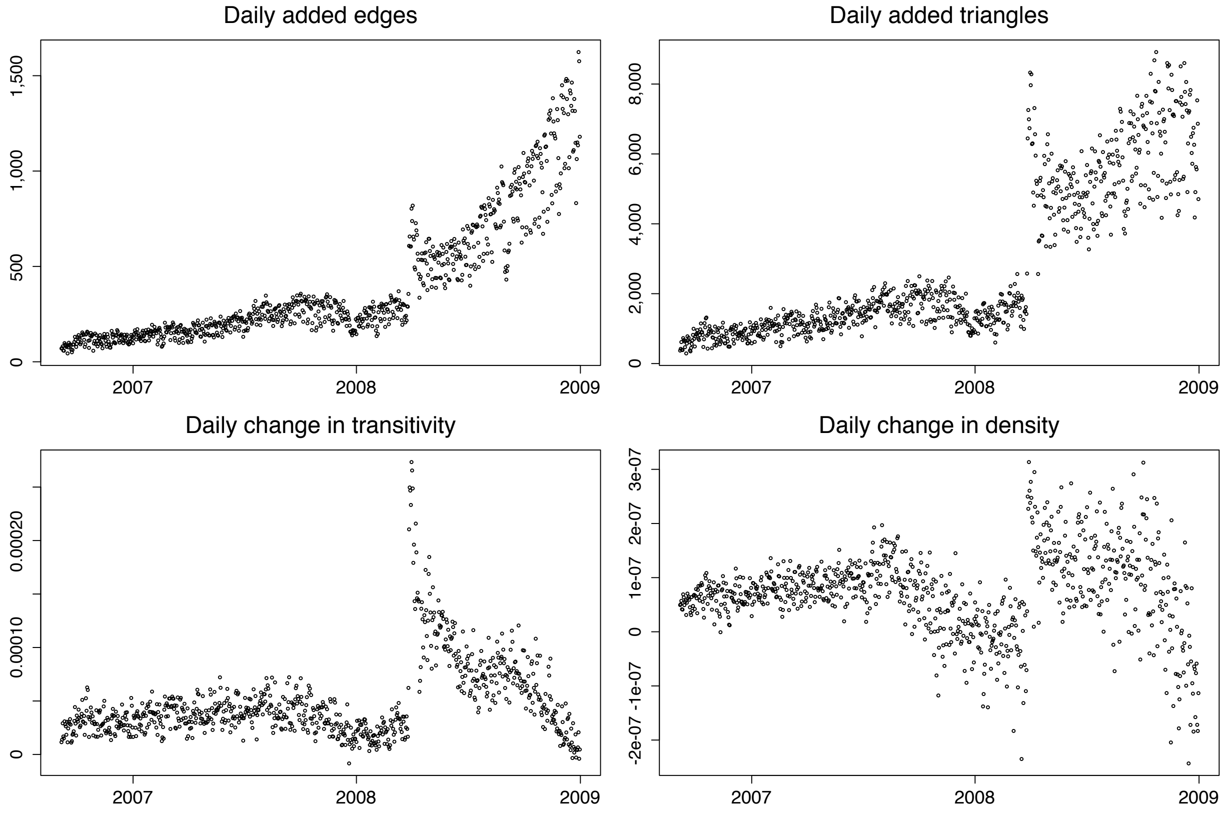
(Thesis)

4. Public health outreach

5. Mobile phone sensors and cohorts

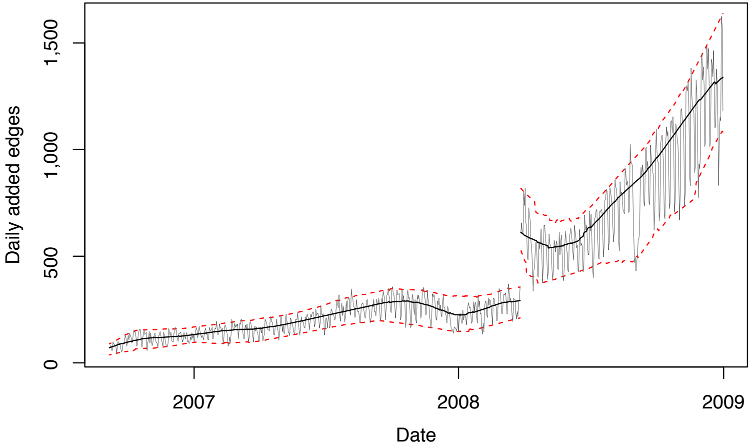
Conclusion

# Case: Facebook's "People You May Know"

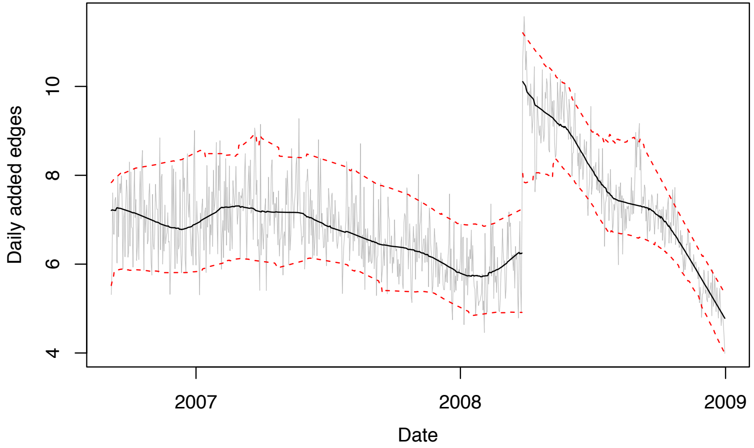


# PYMK changed the Facebook network!

- Facebook links: +300 new edges per day (~200%)




- Triangles: +3.8 triangles per edge (~64%)



# Lesson: Account for platform effects

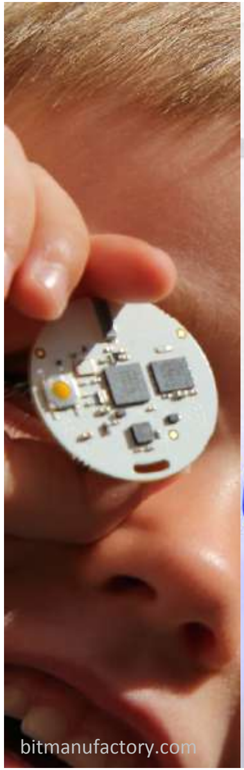
- Decisions made by social media platform engineers are part of what generate data
- How might platform effects change “degrees of separation”? Graph diameter? Small-world properties?
- For both research and applications, consider platform effects

# Chapter 3: Sensors and social networks

with Afsaneh Doryab, Mike Merrill, Jürgen Pfeffer, Anind Dey

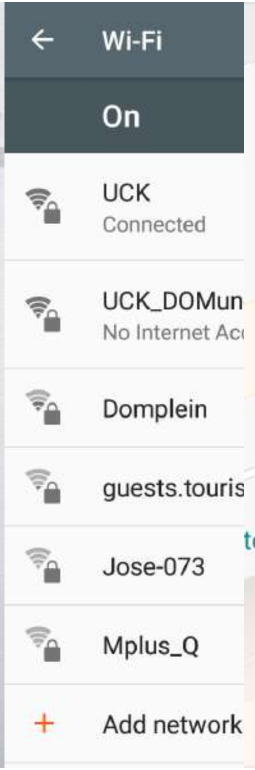
# Relational sensor data

RFID

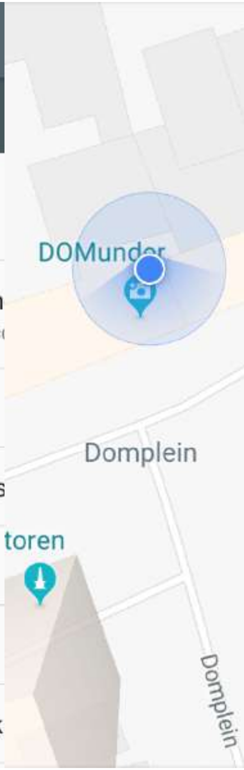



Bluetooth


WiFi

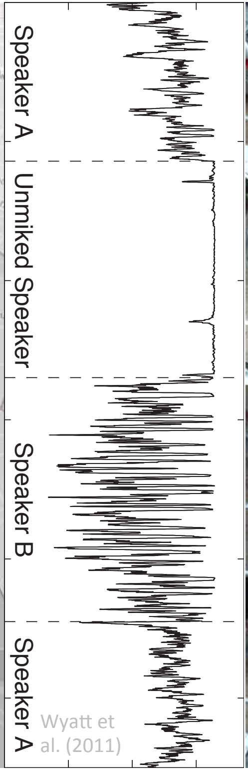



GPS




Cell towers


Audio




Video

# Many sensors + social network studies

| Study | Sensor | Collection |
|---|---|---|
| Sociometric badge | Infrared | 2002, 2007 |
| Reality Mining | Bluetooth | 2004 |
| Social Evolution | Bluetooth | 2008-2009 |
| SocioPatterns | RFID | 2008-2018 |
| Lausanne | Bluetooth | 2009-2010 |
| SocialfMRI | Bluetooth | 2010-2011 |
| Copenhagen Networks Study | Bluetooth, WiFi | 2012-2013 |

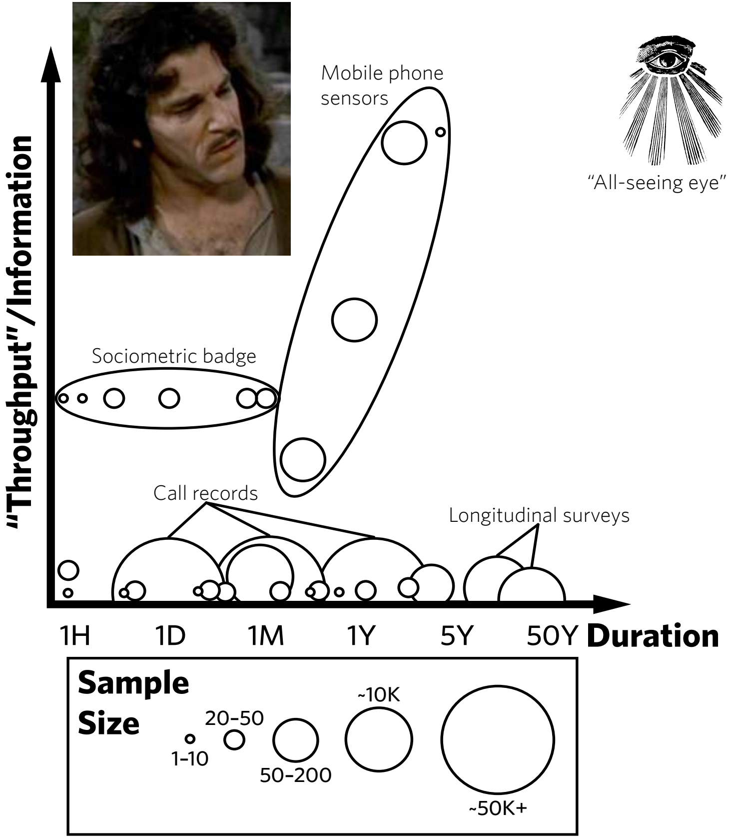
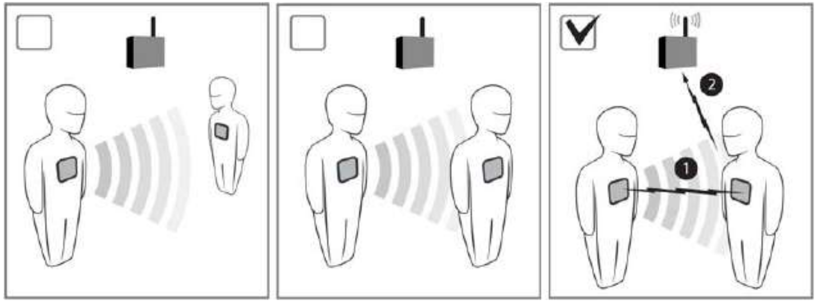



Diagram reproduced from Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing* 7 (6), 643–659.

# Inconsistent terminology suggests confusion



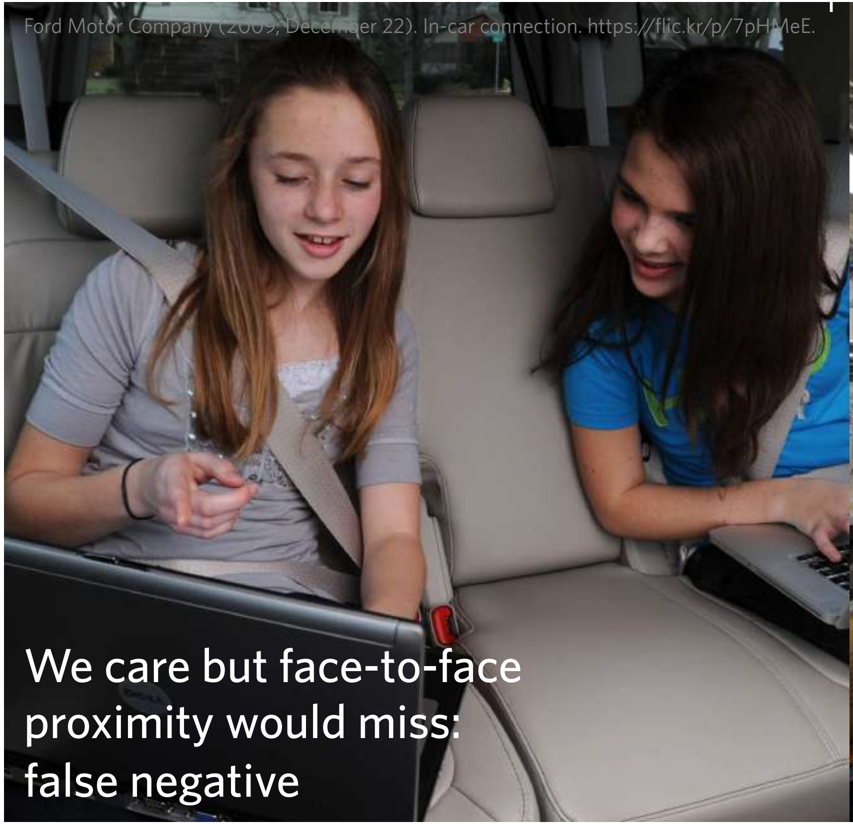

- SocioPatterns (RFID)
  - "Person-to-person interaction"[1]
  - "Face-to-face contacts"[2]
  - "Close-range interactions"[3]
  - "Face-to-face interactions"[4]
  - "Face-to-face proximity"[5]

- Copenhagen Networks Study (Bluetooth)
  - "Proximity data"[6]
  - "Face-to-face interactions"[7]
  - "Close proximity interactions"[8]
  - "Face-to-face contacts"[9]
  - "Physical contacts"[10]

# Back to basics: *Constructs.*

- *Constructs*: basic entities of social science
  - Some constructs are observable, e.g. gender
  - Others are only theoretical, like “verbal ability”
  - Measurements can be a *proxy*
  - Proxies always give errors: need to understand
- Face-to-face interaction: neither the measure nor the construct

# *In-person interaction* is the true construct



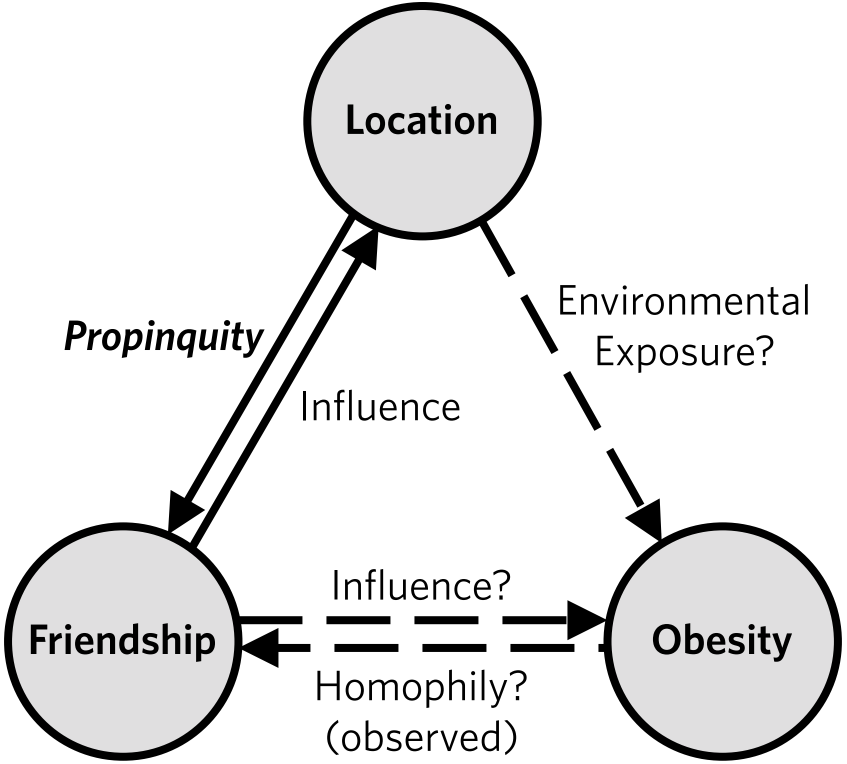

# Interaction is broader than conversation

# Constructs have their own importance

- What we care about?
- Depends on what we want to study/ investigate.
  - Disease transmission? Directional proximity and/ or physical contact.
  - Persuasion? Conversation.
  - Environmental exposure? Proximity.
  - Friendship? Subjective perceptions.

# Theory: Compare multiple phenomena

- Instead of using proximity to measure interaction, or using interaction in place of friendship, *compare*
- In some cases, we want to know which construct is causal



# *Propinquity*: Relates proximity to friendship


Leon Festinger, Kurt W. Back, and Stanley Schachter (1950). *Social pressure in informal groups: A study of human factors in housing*. Stanford University Press.

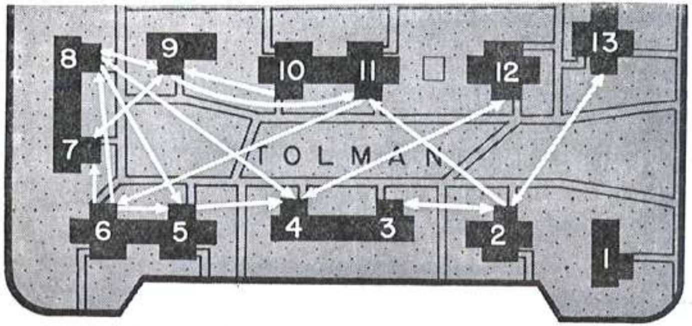



FIG. 9a. Pattern of Sociometric Connections in Tolman Court

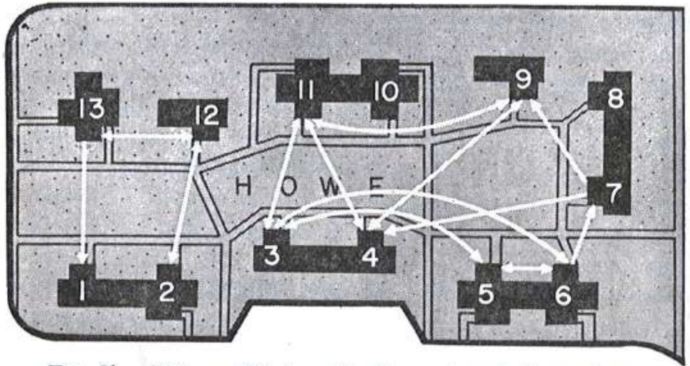
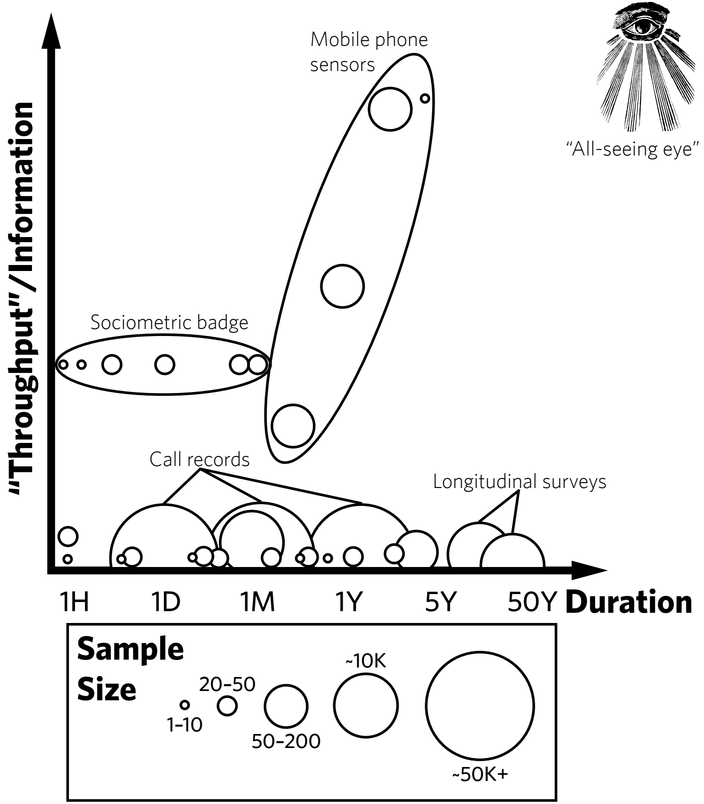



FIG. 9b. Pattern of Sociometric Connections in Howe Court

# Lesson: Identify constructs, establish validity

- Sensors: proximity, not interaction
- To use proximity as a proxy for interaction, first establish validity
- Data sources capture different constructs
- Study relationships between constructs



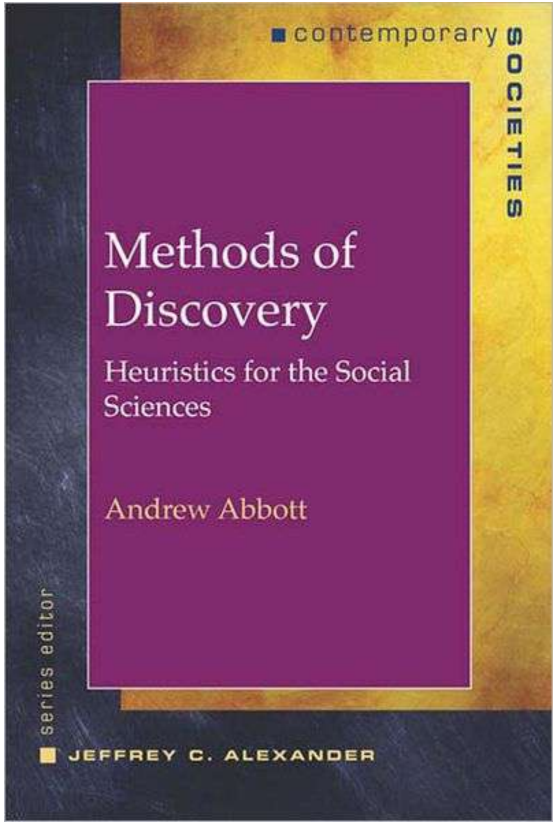
# Part II: Responses

# Central argument: *Shift the scope*

# Analogs to surveys won't work

- Survey data:
  - Sampling strategies and weighting to get representativeness
  - Respondent biases addressed with survey design
- For digital trace data, *such technical approaches will not necessarily work*
  - Corrections for biases and platform effects may remain qualitative
  - Digital trace data may not have an unbiased form: what is a "natural" microblogging platform?

# Instead, *shift the scope*




- Abbott (2004), 3 levels of analysis:
  - Case study analysis, "studying a unique example in great detail"
  - Small-*N* analysis, "seeking similarities and contrasts in a small number of cases"
  - Big-*N* analysis, "emphasizing generalizability by studying large numbers of cases, usually randomly sampled"

# *N* being "big" doesn't have to mean "big-*N*"

- Small-*N* justification:
  - "By making these detailed comparisons, [small-*N* analysis] tries to avoid a standard criticism of single-case analysis—that one can't generalize from a single case—as well as the standard criticism of multicase analysis—that it oversimplifies and changes the meaning of variables by removing them from their context."
- A powerful heuristic in social sciences: *shift the question*
- **Shift the *scope* of trace data from big-*N* to small-*N***

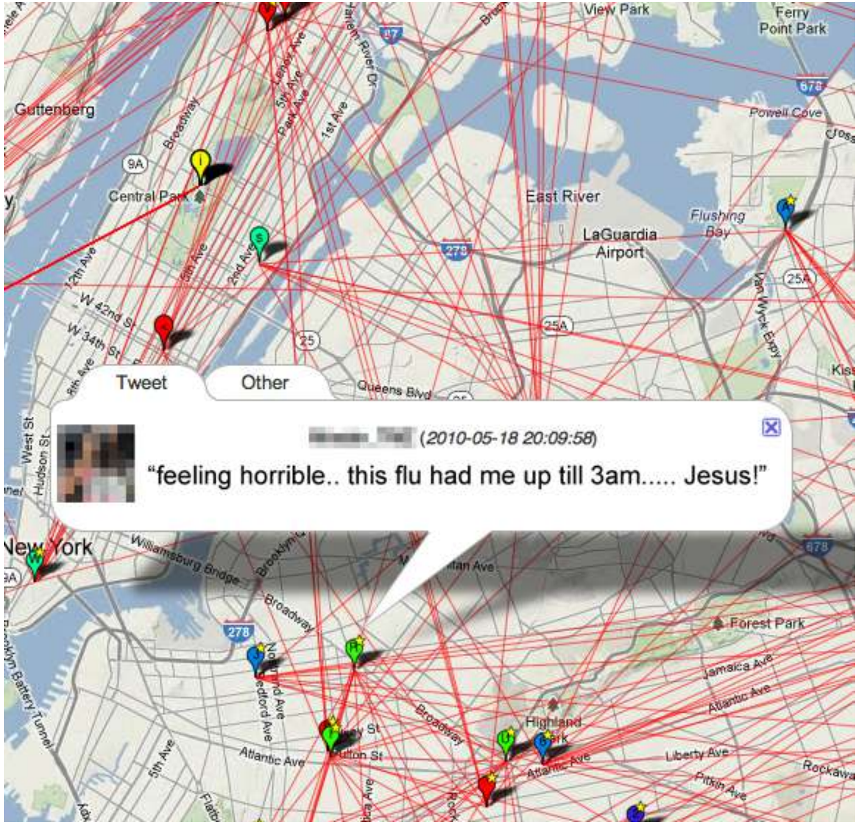
# What will “big-small-*N*” analysis look like?

# Chapter 4: Public health outreach

with Kar-Hai Chu, Jason Colditz, Tabitha Yates, Brian Primack

# Twitter for monitoring is iffy

- Given the biases we know about, should we rely on Twitter for public health monitoring?
- If not, do we give up?




Adam Sadilek, Henry Kautz, and Vincent Silenzio (2012). Modeling spread of disease from social interactions. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (ICWSM-12), 322-329.

# Analogy: Campaigns, not monitoring


- Shift the scope from *measurement* to *outreach*
- Public health campaigns have a long history
- Find the right medium to reach target demographics

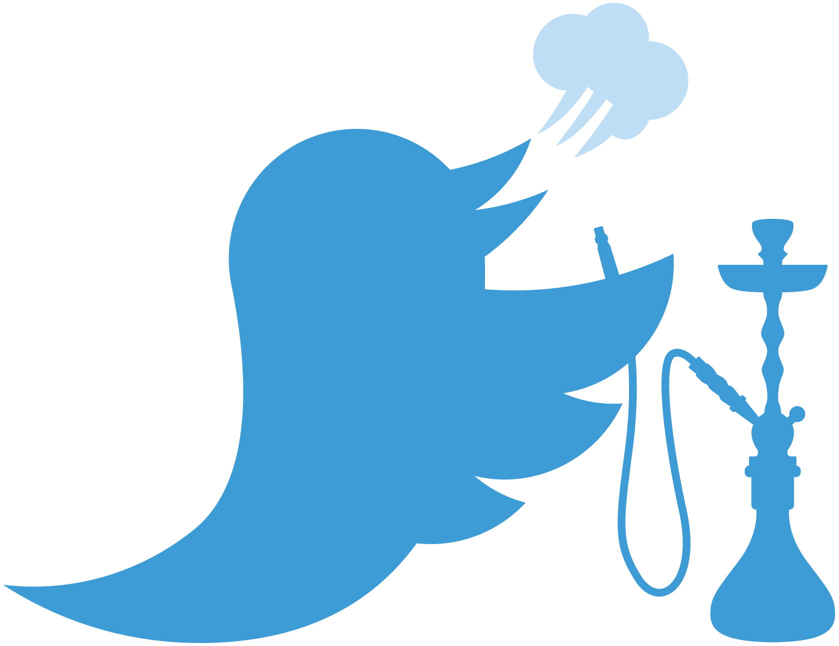
WPA Federal Art Project (1941). LC-DIG-ppmsca-38342 (digital file from original poster). Library of Congress Prints and Photographs Division. http://www.loc.gov/pictures/item/98513584/.

# Case: Hookah smoking

- Cigarette use is down, but hookah use is up, especially among young people
- Misperceptions that hookah is safer than cigarettes; but similar toxins, dependency, cancer risks
- Inform young people via Twitter!


Clickstarpics (2014). "Prince Harry spotted smoking hookah pipe on Abu Dhabi yacht".

# Use social media marketing

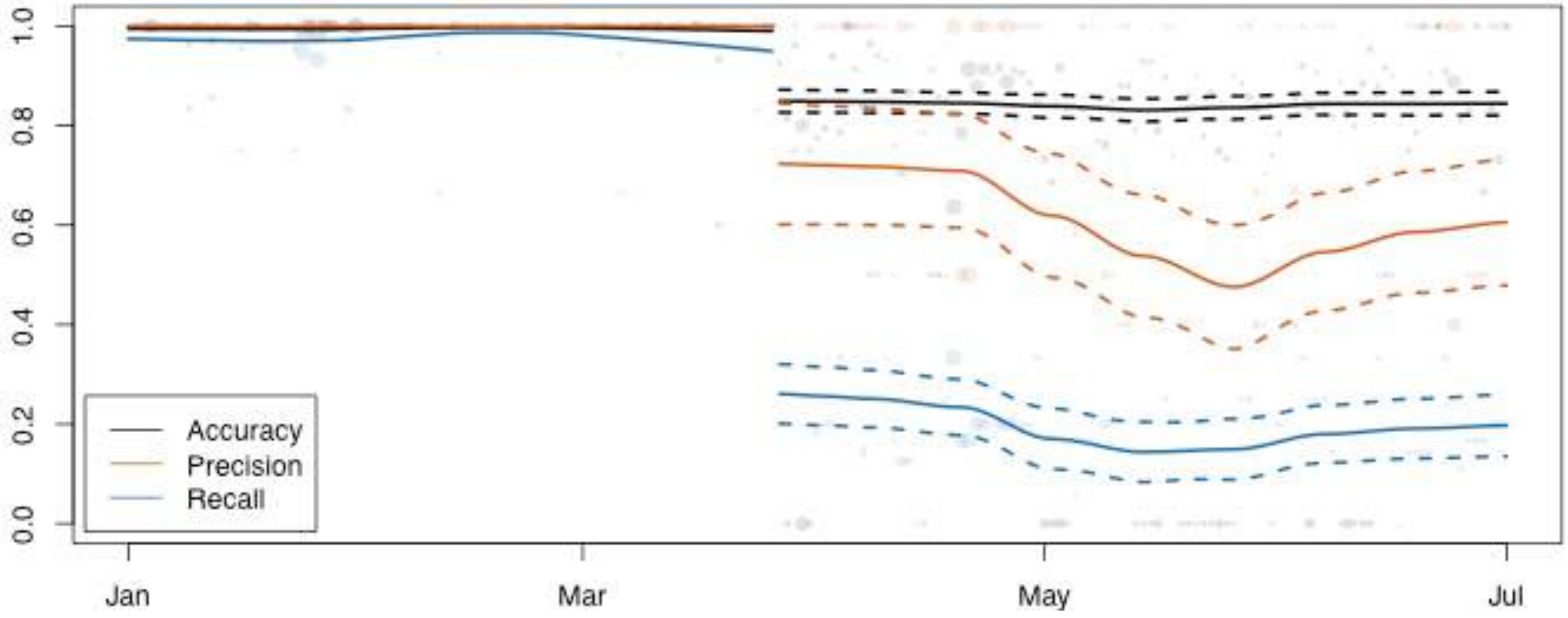


- Twitter is many-to-many, not one-to-many or top-down
- Need to dynamically adjust, interact around trends
- Use social media marketing approaches: first, use machine learning to track hookah expressions on Twitter

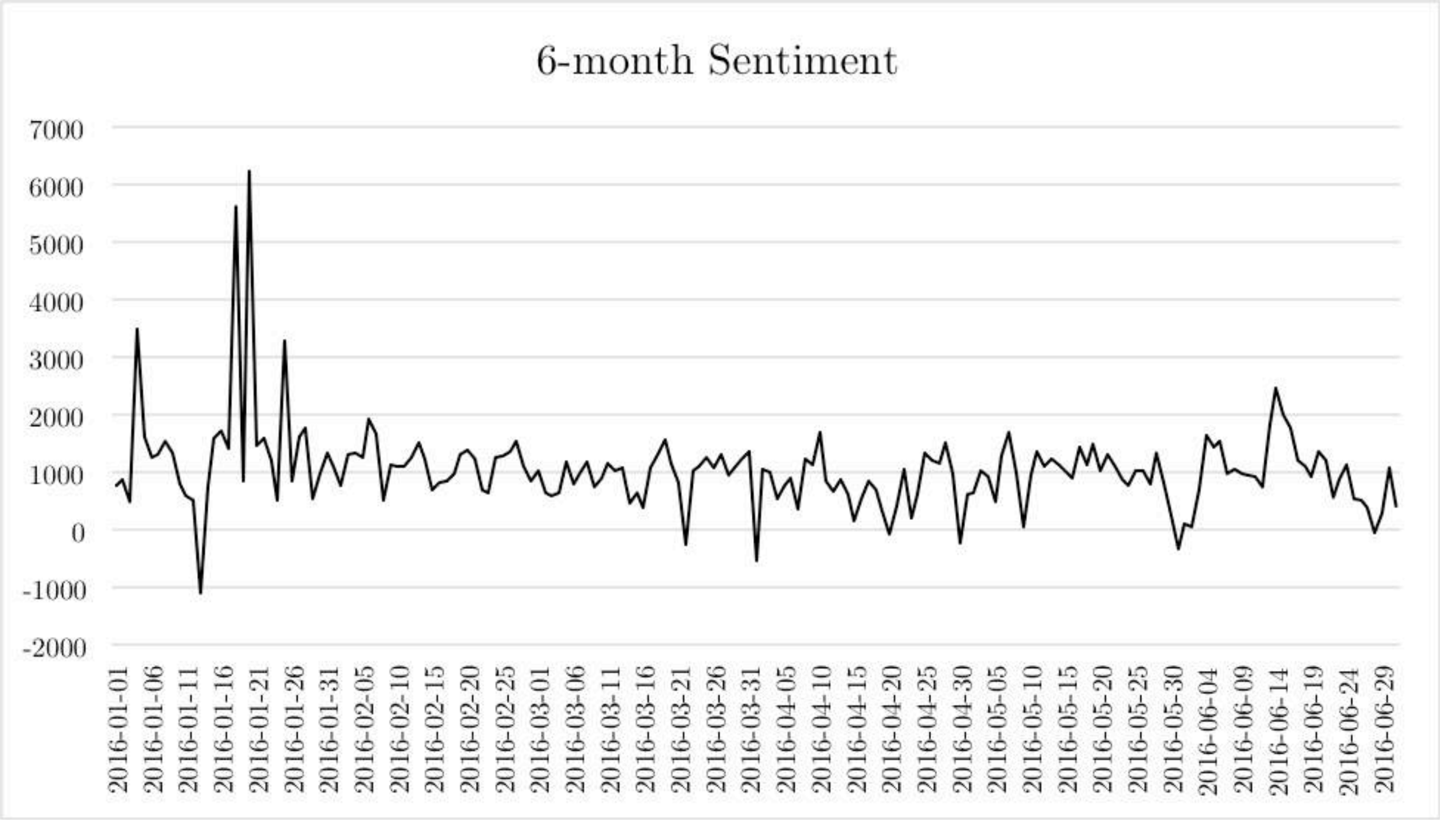
# An appropriate use of machine learning

- There are 560K tweets, too much to code by hand
- Automatic tools won't capture domain knowledge 💨
- Have human coders hand-label 5K tweets to capture domain knowledge, find correlations between words in tweets and labels, apply to rest
- Cannot interpret correlations, but they aren't causal anyway
- All that matters is (properly!) establishing external validity

# Temporal block cross-validation simulates out-of-sample performance



# Use case 1: Scaling up to 500K tweets



# Use case 2: Mixed sentiment discovery

| User | Tweet |
|---|---|
| A | Wednesday about to be lit lmao I need a hookah man |
| A | I don’t want hookah no more dawg lmao |
| B | I wish hookah never existed [URL] |
| B | There’s no hookah so why go [URL] |
| C | FAM be proud of me I havent smoked hookah ALL year |
| C | My ramadan nights bouta consist of me sitting on the porch till 5am skyping and smoking hookah. |

# Lesson: Demonstrating how to use Twitter

- Only a first step: next, will need to try rhetorical and communication strategies
- Set up a system to track relevant activity
- Sampling frame doesn't matter since we are not trying to get findings
- Use machine learning when we don't care how a model works

# Chapter 5: Mobile phone sensors and cohorts

with Afsaneh Doryab, MikeMerrill, Anind Dey

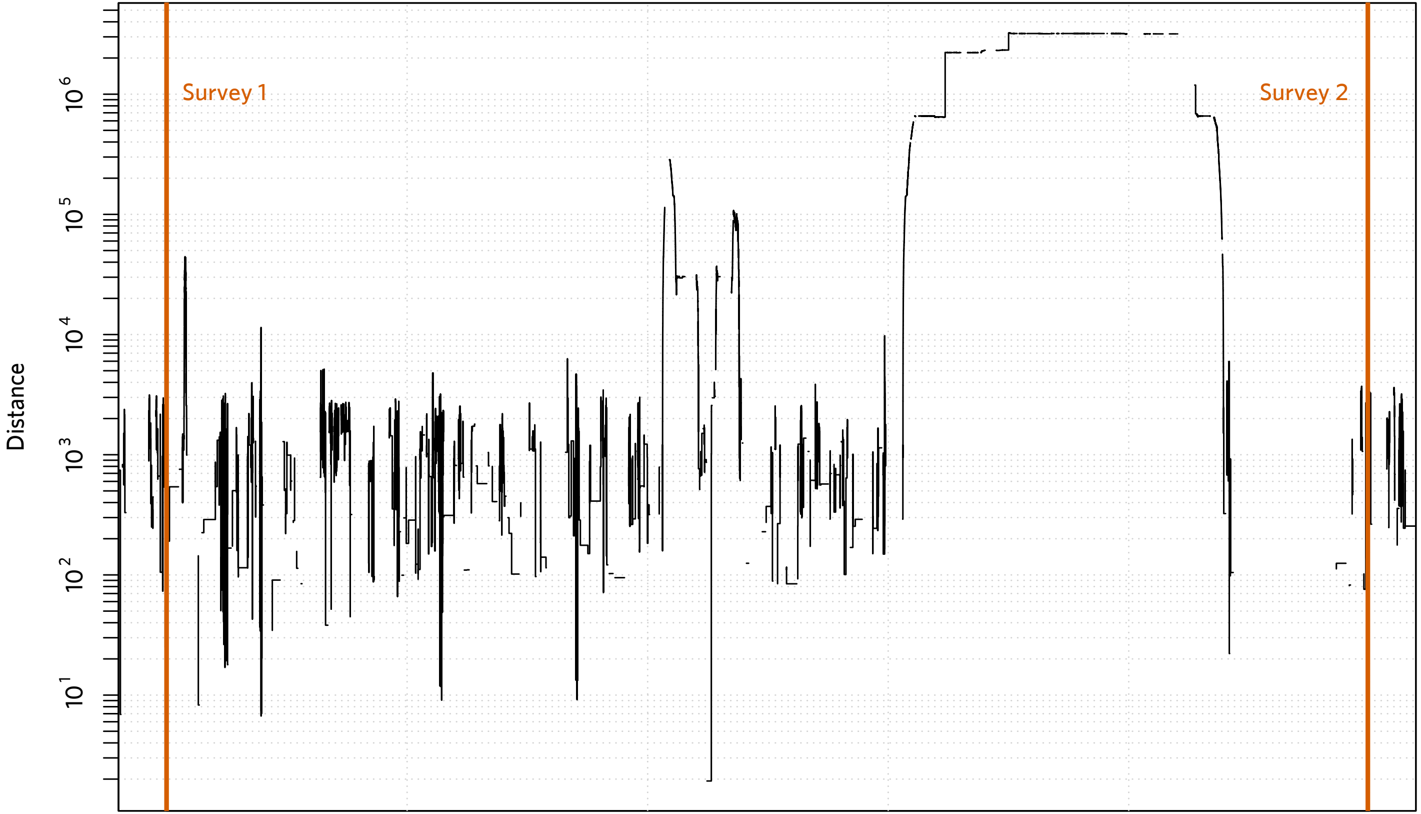
# Shifting the scope to a cohort

- How can we use large-scale trace data not (just) opportunistically, but purposively?
- Sensors provide opportunities for careful study design, relating different constructs
- Combine survey data (self-reported friendships) and a *cohort* boundary with mobile phone sensor tracking

Carnegie Mellon University School of Computer Science



# Core problem: *Different resolutions*



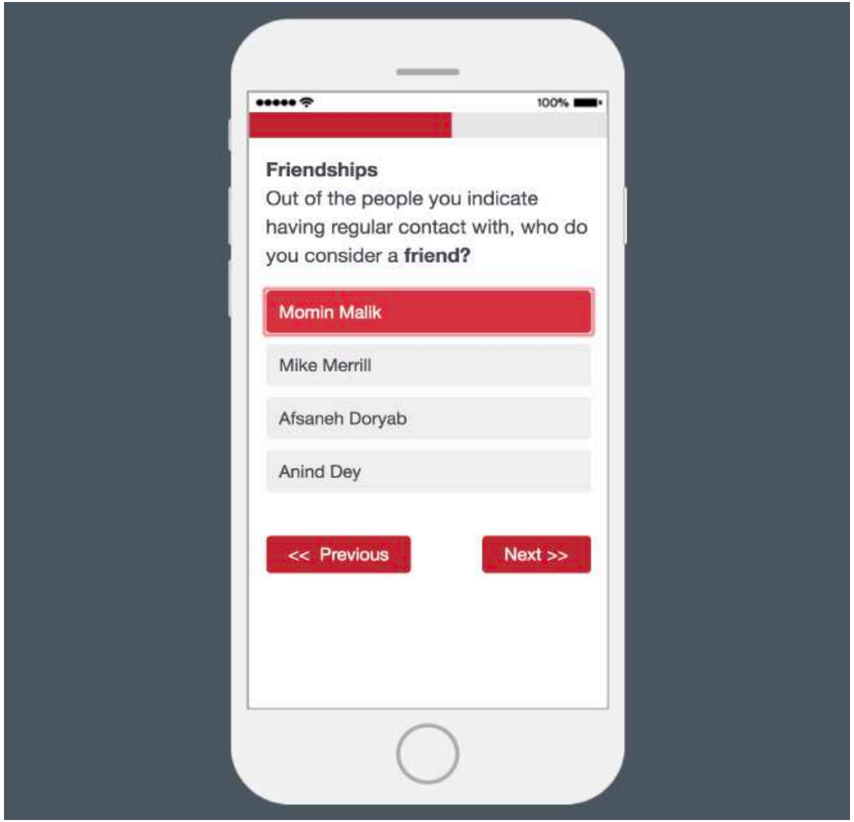
# How does friendship relate to proximity?

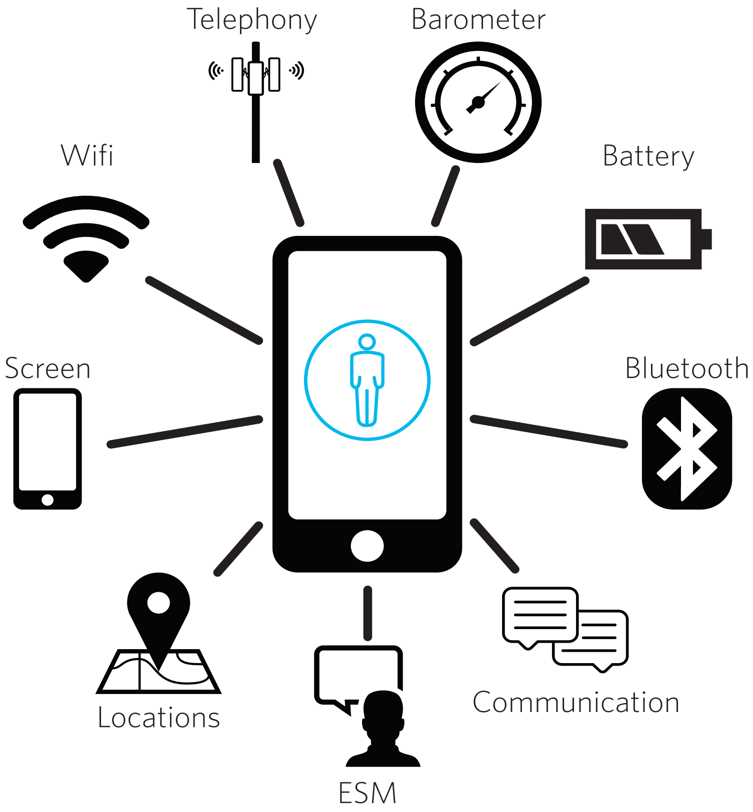
- People who are proximate become friends
- But also, friends spend time together
- Friendship and proximity *co-evolve*
- Compare proximity (via "location", WiFi) to longitudinal sociometric choice (friendship self-report)
- Use a fraternity cohort to get a good *boundary specification*, like in the "Newcomb-Nordlie fraternity" study. We recruited 66% of a fraternity of 70 men

Theodore Mead Newcomb (1961). The acquaintance process. Holt, Reinhard & Winston.
Peter G. Nordlie (1958). A longitudinal study of interpersonal attraction in a natural group setting. PhD thesis, University of Michigan.

# Data: Surveys + mobile phone tracking





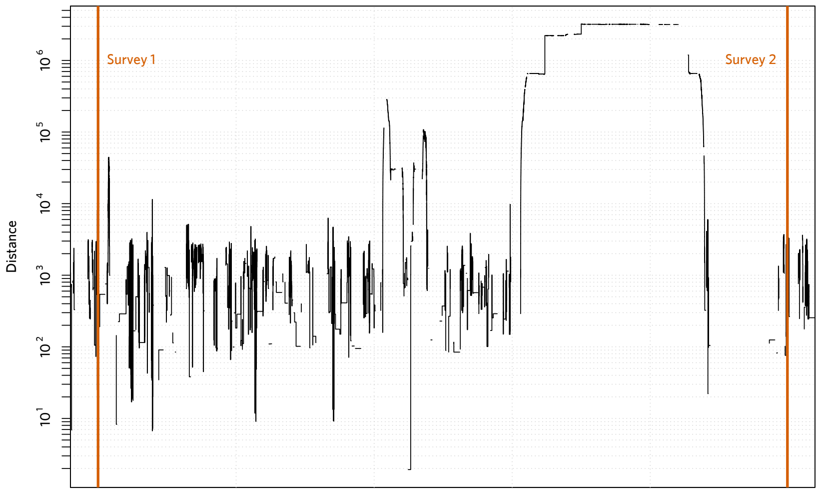



Carnegie Mellon University School of Computer Science



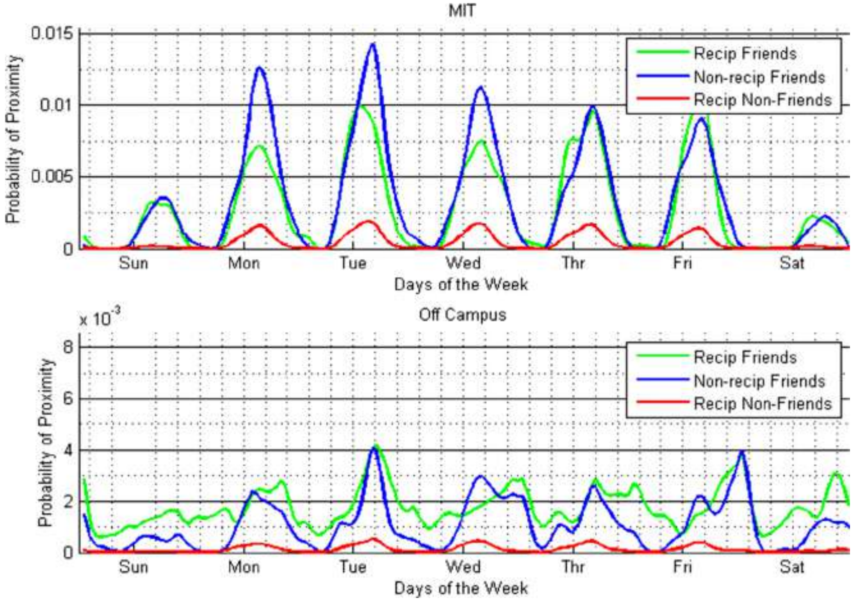
# Use machine learning to find a signal

- How do we address different in resolution?
- We don't *a priori* know how to summarize proximity
- Time of day? Span? Latency?

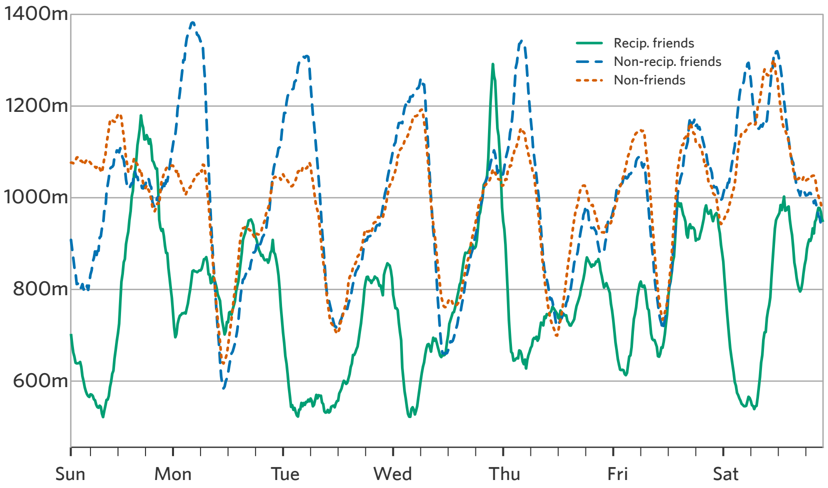


# Aggregation can be misleading

“Probability of proximity” (Reality Mining)

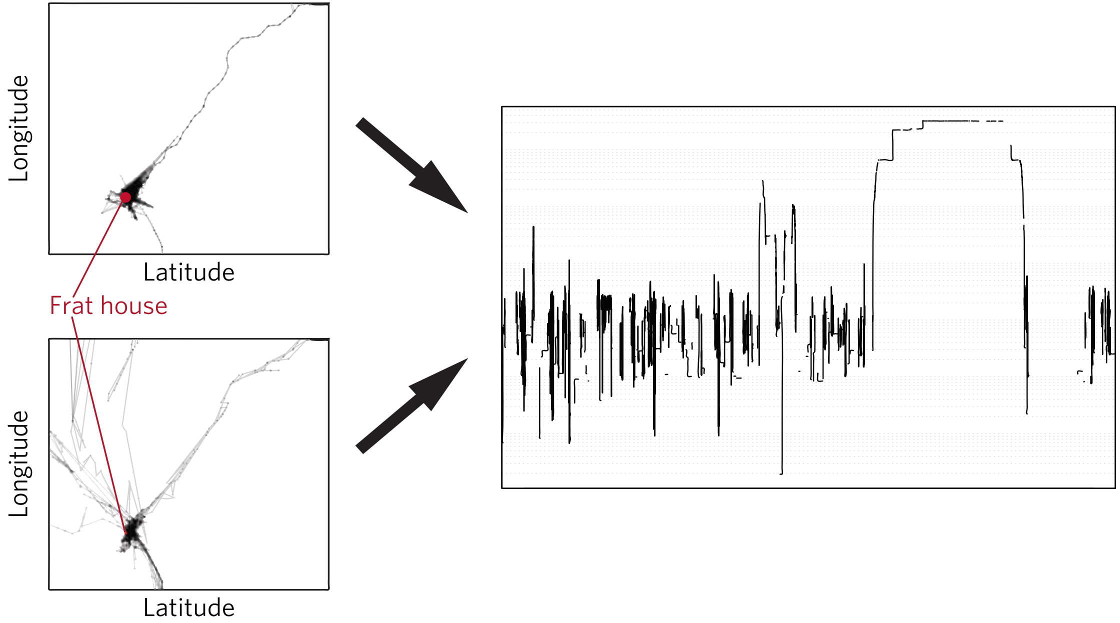



Median pairwise distance (my study)



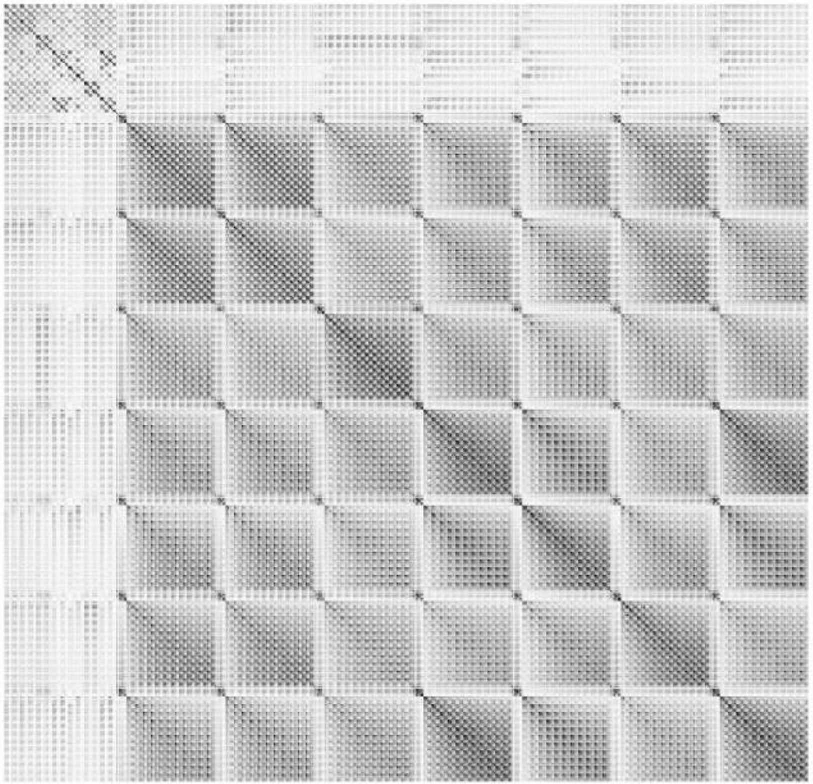

Nathan Eagle, Alex Pentland, and David Lazer (2009). Inferring friendship network structure by using mobile phone data. *PNAS* 106 (36), 15274–15278. doi: 10.1073/pnas.0900282106.

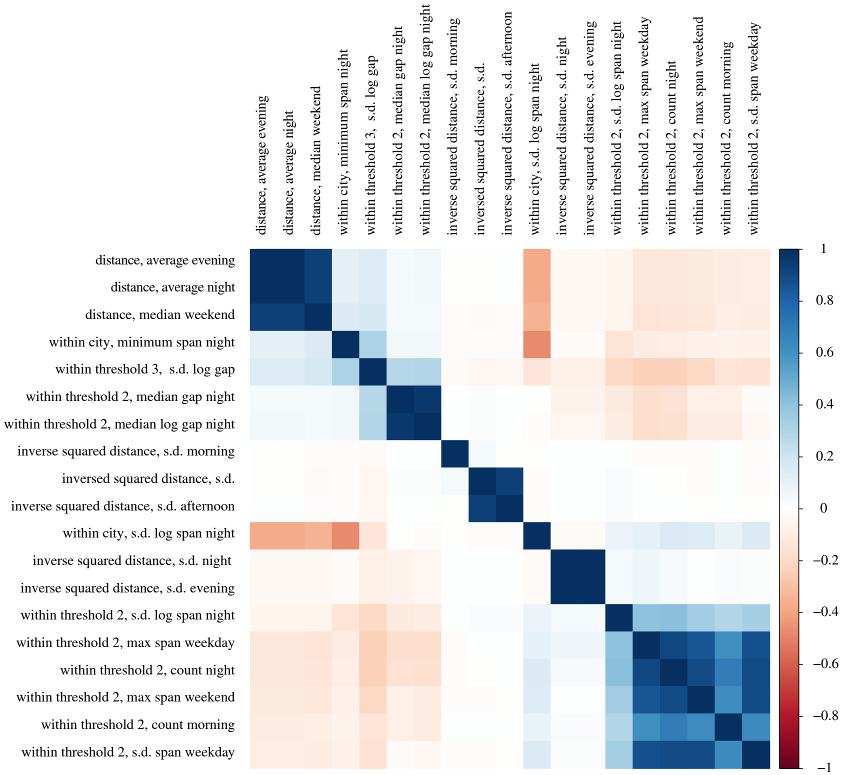
# Data processing and “feature extraction”




| | | | | |
|---|---|---|---|---|
| 0.086 | 0.281 | 0.0793 | 0.079 | |
| 0.005 | 0.073 | 0.0054 | 0.005 | |
| 0.057 | 0.234 | 0.0547 | 0.054 | ⋯ |
| 0.007 | 0.086 | 0.0074 | 0.007 | |
| 0.071 | 0.258 | 0.0669 | 0.066 | |
| 0.024 | 0.154 | 0.0238 | 0.023 | |
| | | ⋮ | | |

# For redundancy, use *feature selection*





# About 30% match, evening/night features

| | Distribution | Summary Statistic | Timeframe |
|---|---|---|---|
| 1 | Distance | Mean | Evening |
| 2 | Distance | Mean | Night |
| 3 | Distance | Median | Weekend |
| 4 | Within city | Minimum span | Night |
| 5 | Within threshold 3 | Log gap | All |
| 6 | Within threshold 2 | Median gap | Night |
| 7 | Within threshold 2 | Median log gap | Night |
| 8 | Inverse squared distance | Standard deviation | Morning |
| 9 | Inverse squared distance | Standard deviation | All |
| 10 | Inverse squared distance | Standard deviation | Afternoon |
| 11 | Within city | SD log span | Night |
| 12 | Inverse squared distance | Standard deviation | Night |
| 13 | Inverse squared distance | Standard deviation | Evening |
| 14 | Within threshold 2 | SD log span | Night |
| 15 | Within threshold 2 | Max span | Night |
| 16 | Within threshold 2 | Count | Night |
| 17 | Within threshold 2 | Max span | Weekend |
| 18 | Within threshold 2 | Count | Morning |
| 19 | Within threshold 2 | SD span | Weekday |

- Best performance: Matthews Correlation Coefficient/Pearson's $\phi = 0.3$
- This approach gives a principled way to characterize how friendship and proximity relate

# Lesson: How to approach sensors

- Build on established social scientific study designs, survey instruments
- Combine types of measurement to compare
- Reduce the sensor data in principled ways
- Finding: spans and variances of inverse squared distance, on evenings and nights, is most correlated with friendship

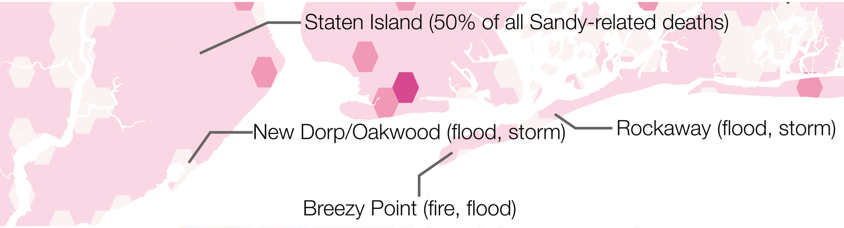
# Conclusion

# Identify bias, and shift scope

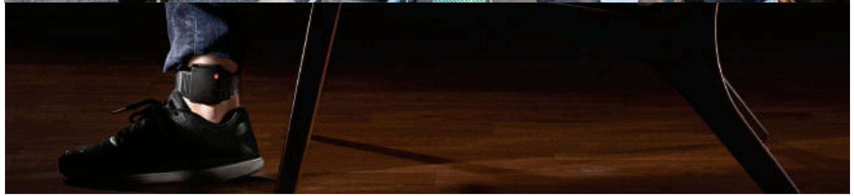
- In Part I (Critiques), I identified biases:
  - Population bias exists, will give unrepresentative results
  - Platform effects change what we think we are studying
  - Sensors measure proximity, not interaction, or friendship
- I claimed that by shifting the scope, we can find new, valid uses of digital trace data
- In Part II (Responses), I demonstrate two shifted scopes:
  - Use Twitter for public health engagement, not public health monitoring
  - Use sensors to study the interplay between proximity and friendship, not as a replacement for studying friendship

# Implications for usage?

# Returning to examples from intro




| Goal: | Predict, Monitor, and Prevent Risk In/Around Protests |
|---|---|
| Anticipated Activity: | Protests, Riots, Looting |
| Overt Threats: | Unions, Activist Groups, Etc. |
| Locations: | Schools, Public Spaces, Malls, High-Rent Districts |
| Actions Taken: | During Event(s), Post-Event |



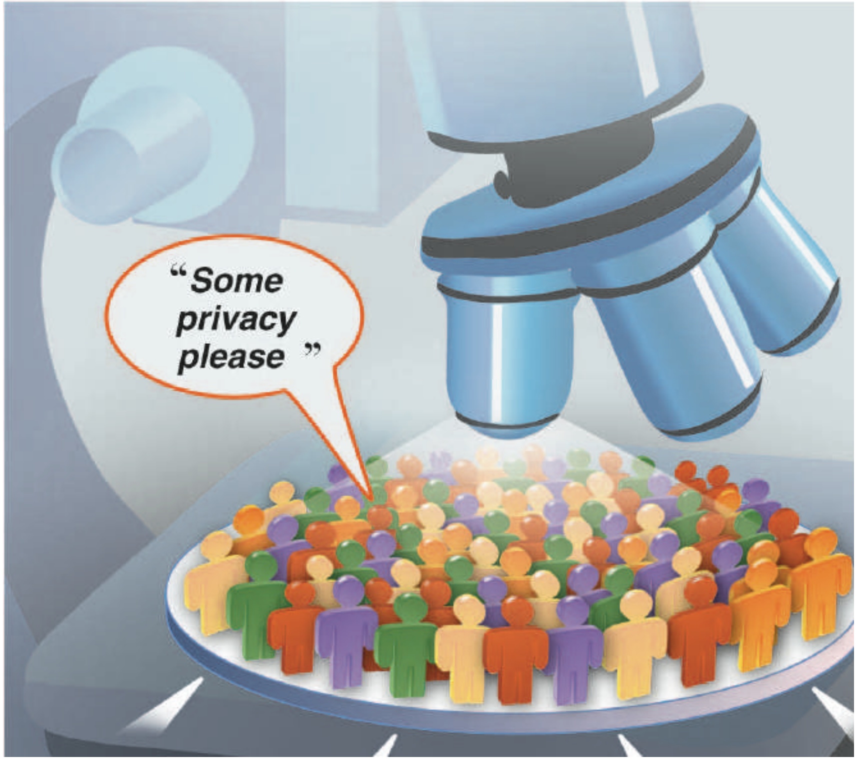
- →Find ways to correct the signal
- →Work on legal protections
- →Build support tools, oppose punishment
- →Study false positives/negatives

# Parting thought: On measurement and the development of science

# A “microscope” for social science?

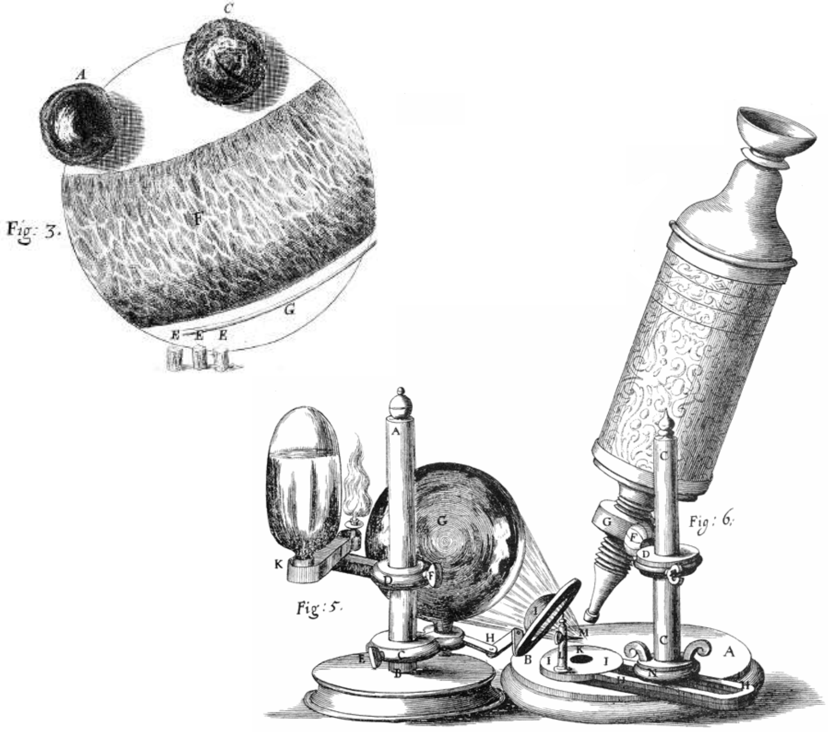


“Disciplines are revolutionized by the development of novel tools: the telescope for astronomers, the **microscope for biologists**, the particle accelerator for physicists, and brain imaging for cognitive psychologists. **Social media provide a high-powered lens into the details of human behavior and social interaction** that may prove to be equally transformative.”

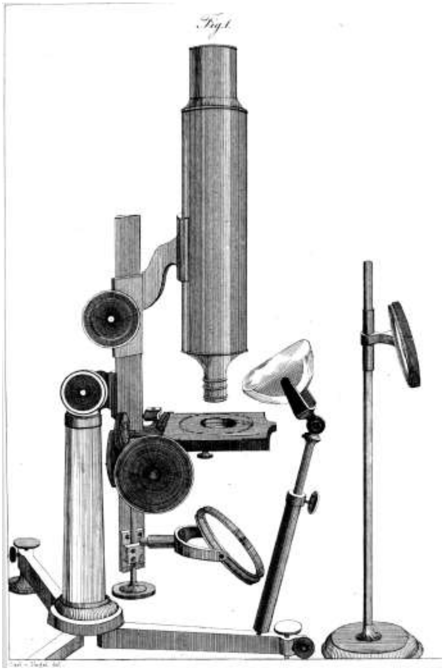


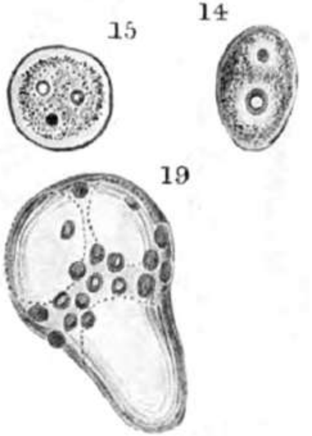

Scott Golder and Michael Macy (2012). Social science with social media. *ASA footnotes* 40(1). Gary King (2011). Ensuring the data-rich future of the social sciences. *Science* 331, 719–721.

# Cells *described* in 1665; *cell theory* in 1830s!




Robert Hooke (1665). *Micrographia: or some phyſiological deſcriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon*.





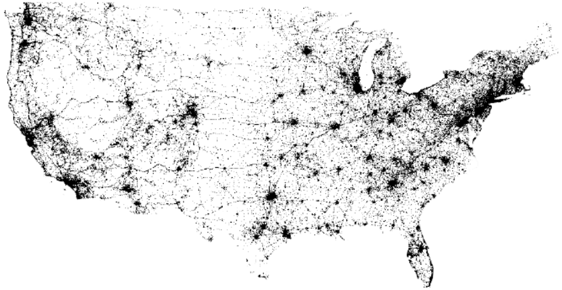

Theodor Schwann (1839). Mikroskopische Untersuchungern uber die Uebereinstimmung in der Stuktur und dem wachsthum der Thiere und Pfanzen. https://wellcomecollection.org/works/mjpkz6zb.
Joseph Berres (1837). Anatomie der mikroskopischen Gebilde des menschlichen Körpers.

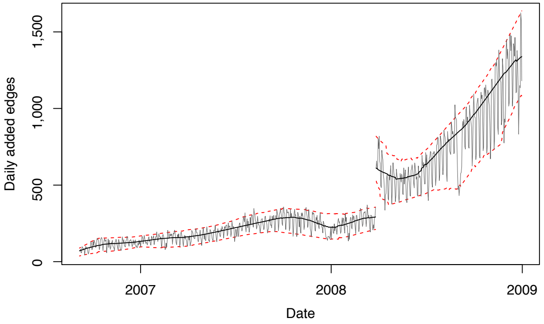
# Tools are not self-contained or sufficient

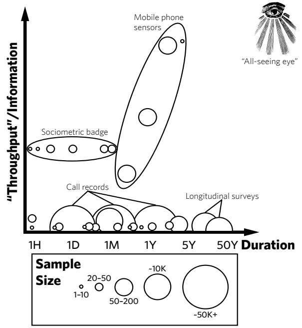
- As we understand more, we improve the tool
- We may need to manipulate the phenomenon to make it visible to the tool
- Hopefully digital trace data won't take 130 years to lead to new theory…
- By understanding biases in digital trace data, we can go beyond its current limits.

Carnegie Mellon University
School of Computer Science

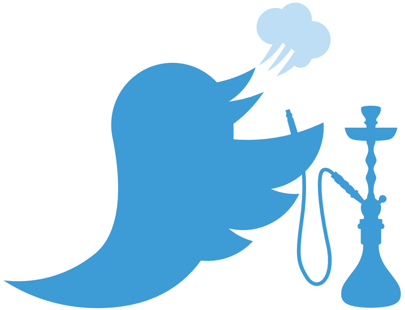


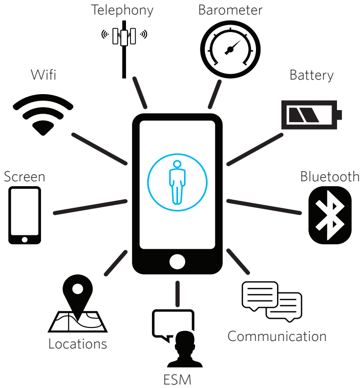





# Thank you!





# Endnotes



Slide 36

1. Danny Wyatt, Tanzeem Choudhury, Jeff Bilmes, and James A. Kitts (2011). Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science. *ACM Transactions on Intelligent System Technologies* 2 (1), 7:1–7:41. doi: 10.1145/1889681.1889688.
2. M. S. Ryoo and J. K. Aggarwal (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. *Proceedings of the IEEE 12th International Conference on Computer Vision* (ICCV), 1593–1600.

Slide 38

1. Vedran Sekara and Sune Lehmann (2014). “The strength of friendship ties in proximity sensor data”. *PLOS ONE* 9 (7), 1–8. doi: 10.1371/journal.pone.0100915.
2. Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann (2014). “Measuring large-scale social networks with high resolution”. *PLOS ONE* 9 (4), 1–24. doi: 10.1371/journal.pone.0095978.
3. Stopczynski, Arkadiusz, Piotr Sapiezynski, Alex Pentland, and Sune Lehmann (2015). “Temporal fidelity in dynamic social networks”. *The European Physical Journal B* 88 (249). doi: 10.1140/epjb/e2015- 60549-7.
4. Anders Mollgaard, Ingo Zettler, Jesper Dammeyer, Mogens H. Jensen, Sune Lehmann, and Joachim Mathiesen (2016). “Measure of node similarity in multilayer networks”. *PLOS ONE* 11 (6), 1–10. doi: 10.1371/journal.pone.0157436.
5. Enys Mones, Arkadiusz Stopczynski, and Sune Lehmann (2017). “Contact activity and dynamics of the social core”. *EPJ Data Science* 6 (1). doi: 10.1140/epjds/s13688-017-0103-y.
6. Ciro Cattuto, Wouter van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani (2010). “Dynamics of person-to-person interactions from distributed RFID sensor networks”. *PLOS ONE* 5 (7), e11596. doi: 10.1371/journal.pone.0011596.
7. Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Caterina Rizzo, Alberto Eugenio Tozzi, and Wouter van den Broeck (2012). “Wearable sensor networks for measuring face-to-face contact patterns in healthcare settings”. *Revised Selected Papers from the Third International Conference on Electronic Healthcare* (eHealth 2010), 192–195. doi: 10.1007/978-3-642-23635-8_24.
8. Ciro Cattuto, Marco Quaggiotto, André Panisson, and Alex Averbuch (2013). “Time-varying social networks in a graph database: A Neo4J use case”. *Proceedings of the First International Workshop on Graph Data Management Experiences and Systems* (GRADES ’13), 11:1–11:6. doi: 10.1145/2484425.2484442.
9. Alain Barrat, Ciro Cattuto, Vittoria Colizza, Francesco Gesualdo, Lorenzo Isella, Elisabetta Pandolfi, Jean- François Pinton, Lucilla Ravà, Caterina Rizzo, Mariateresa Romano, Juliette Stehlé, Alberto Eugenio Tozzi, and Wouter van den Broeck (2013). “Empirical temporal networks of face-to-face human interactions”. *The European Physical Journal Special Topics* 222 (6), 1295–1309. doi: 10.1140/epjst/ e2013-01927-7.
10. Alain Barrat, Ciro Cattuto, Alberto Eugenio Tozzi, Philippe Vanhems, and Nicolas Voirin (2014). “Measuring contact patterns with wearable sensors: Methods, data characteristics and applications to data-driven simulations of infectious diseases”. I*linical Microbiology and Infection* 20 (1), 10–16. doi: 10.1111/1469-0691.12472.