Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries

Alexandra Olteanu
Carlos Castillo, Eurecat
Fernando Diaz, Microsoft Research
Emre Kıcıman, Microsoft Research

Social data in digital form, which includes user-generated content, expressed or implicit relationships between people, and behavioral traces, are at the core of many popular applications and platforms, and drive the research agenda of many researchers. The promises of social data are many, including understanding "what the world thinks" about a social issue, brand, product, celebrity, or other entity, as well as enabling better decision making in a variety of fields including public policy, healthcare, and economics. Many academics and practitioners have warned against the naïve usage of social data. There are biases and inaccuracies at the source of the data, but also introduced during processing. There are methodological limitations and pitfalls, as well as ethical boundaries and unexpected consequences that are often overlooked. This survey recognizes that the rigor with which these issues are addressed by different researchers varies across a wide range. We present a framework for identifying a broad range of menaces in the research and practices around social data.

Additional Key Words and Phrases: Social media, user-generated content, behavioral traces, biases, evaluation

1. INTRODUCTION

> *"For your own sanity, you have to remember that not all problems can be solved. Not all problems can be solved, but all problems can be illuminated." –Ursula Franklin*[1]

This survey covers a series of concerns on the usage by researchers of social data for a variety of goals. To set the context, in this section, we describe social data and what is it being used for (§1.1), outline general concerns about its usage as voiced by academics in the past (§1.2), and overview the remainder of the survey (§1.3).

1.1. Social Software and Social Data

Today, companies, researchers, governmental agencies, and nongovernmental organizations all rely in some way or another on *social data*, which we use as a broad umbrella concept for all kind of digital traces produced by or about users, with an emphasis on content explicitly written with the intent of communicating or interacting with others. Social data is used to build or provide products and services, to support decision or policy making, or to characterize human phenomena.

Social data comes from *social software*. Social software provides an intermediary or a focus for a social relationship [Schuler 1994]. It includes a diverse array of *platforms*—from social media and networking (e.g., Twitter, Pinterest, or Facebook), recommendation and Q&A sites (e.g., Booking or Quora), to collaborative sites or search platforms (e.g., Google or Wikipedia); of *purposes*—from finding information [White 2013] to keeping in touch with friends [Lampe et al. 2008]; as well as of *data points* meanings and semantics (e.g., clicks, likes, shares, social links) [Tufekci 2014]. Online social software forms the *social web*, "a class of websites and applications in which user participation is the primary driver of value" [Gruber 2008].

The social web has enabled access to *social traces* at a scale and level of detail, both in breadth and depth, impractical with conventional data collection techniques, such as surveys and user studies [boyd and Crawford 2012; Richardson 2008]. On the social web users search, create, interact and share information on a rich mix of topics including work [Ehrlich and Shami 2010], food [Abbar et al. 2015], health [De Choudhury et al. 2014], relations [Garimella et al. 2014], or weather events [Kıcıman 2012]. While doing so, they leave rich data traces that form what Harford [2014] calls *found data*: "the digital exhaust of web searches, credit card payments and mobiles pinging

---

This is a draft from a work-in-progress. Please write to alexandra@aolteanu.com for up-to-date citation information.

[1] Quoted by M. Meredith in http://bb9.berlinbiennale.de/all-problems-can-be-illuminated-not-all-problems-can-be-solved/.

the nearest phone mast." To highlight the collective and user-driven nature of this data, researchers have coined a variety of terms to refer to it like "human traces", "usage data", or a kind of "wisdom of crowds", among others [Baeza-Yates 2014; Baeza-Yates and Maarek 2012; Dumais et al. 2014].

People volunteer these data for various reasons; these platforms allow them to achieve some goals or receive certain direct benefits. Motivations include communication, friendship maintenance, job seeking, self-presentation or promotion [DiMicco et al. 2008; Joinson 2008; Lampe et al. 2006; Naaman et al. 2010], which are also central to understanding ethical aspects of using this data.

Social data opens unprecedented opportunities to answer significant questions about society, policies, and health. The increased availability of datasets has been recognized as a core reason behind the progress in many areas of computing (e.g., object recognition, crisis informatics, digital health, computational social science) [Crawford and Finn 2014; Torralba and Efros 2011; Tufekci 2014; Yom-Tov 2016]. It is believed to provide insights into both individual-level and large human phenomena, having a plethora of applications and substantial impact [boyd and Crawford 2012; Dumais et al. 2014; Harford 2014; Lazer et al. 2009; Ruths and Pfeffer 2014]. Concomitantly, there is also a growing concern and consensus that while the ever-growing datasets of online social traces offer captivating insights into human phenomena, they are more than just an observational tool.

1.2. A Growing Concern

Social data are used to make inferences about how much you should pay for a product [Hannak et al. 2014], about your likelihood of being a terrorist[2], about your health [Yom-Tov 2016], your employability [Rosenblat et al. 2014], and political views [Cohen and Ruths 2013]. Such inferences are increasingly being used to support decision and policy making, and can have important negative implications [Diakopoulos 2016; O'Neil 2016; Reed and boyd 2016]. Yet, such implications are not always well understood or recognized [O'Neil 2016; Tufekci 2014], and many studies seem to assume that these data, and the frameworks used to handle them, are *adequate*, often *as-is*, for the problem at hand—with little or no scrutiny.

In the light of Google Flu Trends' success [Ginsberg et al. 2009] the provocative essay "The End of Theory" [Anderson 2008] sparked intense debates by saying: "Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves." Yet, while the ability to capture large volumes of data brings along important scientific opportunities [Giles 2012; King 2011; Lazer et al. 2009], size by itself is not enough. Indeed, such claims have been debunked by numerous critics [boyd and Crawford 2012; Giardullo 2015; Harford 2014; Lazer et al. 2014], emphasizing that they ignore, among others, that size alone does not necessarily make the data better [boyd and Crawford 2012] as "there are a lot of small data problems that occur in big data" which "don't disappear because you've got lots of the stuff. They get worse." [Harford 2014].

In fact, regardless of how large or varied social data are, there are also lingering questions about what can be learned from this data about real-world phenomena (online or offline), or even about media or application-specific phenomena—which have *yet* to be rigorously addressed [Barocas and Selbst 2014; boyd and Crawford 2012; Ruths and Pfeffer 2014; Tufekci 2014]: e.g., *How reflective is user behavior within an online medium of the real-world phenomena being studied? How generalizable are the findings from one medium to other media? How representative are the working datasets of the targeted platform data? How robust are our tools to data biases and variations? Are errors evenly distributed across different classes of data?*

Thus, given that these data are increasingly used to drive policies, to shape products and services, and for automated decision making, it becomes imperative to gain a better understanding of the limitations around the use of various social datasets and their consequences, of the cost of different errors across applications and semantic domains, and of the efforts to address them [boyd and Crawford 2012; O'Neil 2016; Wagstaff 2012]. Overlooking such limitations can lead to wrong or

[2]Patrick Tucker: "Refugee or Terrorist? IBM Thinks Its Software Has the Answer". *Defense One*, January 2016. Online: http://www.defenseone.com/technology/2016/01/refugee-or-terrorist-ibm-thinks-its-software-has-answer/125484/.

Table I: General challenges, including generic issues along an idealized data processing pipeline.

General challenges (§3)					
Population biases	Behavioral biases	Content biases	Linking biases	Temp. variations	Redundancy
	Source (§4)	Collecting (§5)	Processing (§6)	Analyzing (§7)	Evaluating (§8)
Data processing pipeline	· Functional biases · Normative biases · External biases · Non-individual/s	· Acquisition · Querying · Filtering	· Cleaning · Enrichment · Aggregation	· Qualitat. analys. · Descript. stats · Inferences · Observ. studies	· Metrics · Interpretation · Disclaimers
Ethical considerations (§9)					
· Respect to individual autonomy		· Beneficence and non-maleficence		· Justice	

inappropriate results [boyd and Crawford 2012; Kıcıman et al. 2014], which could be consequential particularly when used for policy or decision making [Lazer et al. 2014].

At this point, a challenge for both academic researchers and applied data scientists using social data, is that there is not enough agreement on a vocabulary or taxonomy of biases, methodological issues, and pitfalls of this type of research. This survey is intended for researchers and practitioners who want to examine their own work, or that of others, through the lens of these issues.

1.3. Organization of This Survey

This survey aims to consolidate prior calls (including Barocas and Selbst [2014]; boyd and Crawford [2012]; Chou [2015]; Dwork and Mulligan [2013]; Ekbia et al. [2015]; Gillespie [2015]; Kenneally [2015]; Ruths and Pfeffer [2014]; Tufekci [2014]) to carefully scrutinize the use of social data in various disciplines against a variety of possible data and methodological pitfalls. In our analysis, we occasionally go beyond social data and into research that could be placed within social computing or computational social science more generally [Mann 2016; Mason et al. 2014; Oboler et al. 2012]. We recognize that this is an evolving topic, and we aim at providing a strong foundation for it.

We begin the survey by noting that whether a research method is adequate or not depends on the research questions being asked. Hence, we first discuss typical research questions and methods used when analyzing social data (§2). Next, while noting that research using social data, as many other types of research, does not happen in a linear fashion, we describe challenges along an idealized data processing pipeline. We begin with general problems (§3), and then analyze problems at the data source (§4), or introduced during data collection (§5), processing (§6), analysis (§7), or evaluation (§8), as outlined in Table I. Finally, we discuss ethical issues regarding the usage of social data (§9), before wrapping up with an overview of trends and future directions (§10).

2. CONTEXT AND GENERAL FRAMEWORK

Evaluating whether a dataset is biased or a methodology is adequate depends on the context in which research takes place, and fundamentally on the goals of the researcher(s).

2.1. Prototypical Goals of Social Data Analysis

As we have stated, a wide variety of social platforms exists, sometimes overlapping and often complementing each other by supporting different functionalities, purposes or domains of focus. This diversity has enabled researchers to explore the potential benefits of social traces for a variety of domains and applications. We can broadly identify two large classes of research goals:

I. *to understand phenomena specific to social software platforms*, sometimes with the objective of improving them; or
II. *to understand phenomena beyond the social platforms*, seeking to answer questions from sociology, psychology, or other disciplines.

Type I research focuses on questions specific to a particular social platform or to a family of social platforms, typically applying methods from computer science fields such as data mining or

human-computer interaction. This includes, for instance, research on how to maximize the spread of memes in a particular platform, how to make social software more engaging, or how to improve the search engine or recommender system of a platform.

Type II research is about using data from social software platforms to address questions about phenomena that happens outside these platforms. This research may occur in emerging interdisciplinary domains, such as computational social science and digital humanities. Researchers addressing this type of problem may seek to use social data to understand questions relevant to media, governments, non-governmental organizations, and business, or to work on problems from domains such as health care, politics, economics, and education. Sometimes, the research question can be about the impact of social software platforms in these domains, for instance, to try to describe the influence of social media in a political election. In other cases, the research question may lie entirely outside social media itself, for instance, to try to track the evolution of contagious diseases by analyzing symptoms reported online by social media users.

2.2. Prototypical Processing Pipeline for Social Data

While research goals are diverse, social data research involves many common steps, independent of the research problem. For simplicity, to outline these steps we assume that the research question has been defined (§2.1) and the ethical concerns outlined (§9):

Data acquisition and preparation. The first natural step is to gather data collections that satisfy certain characteristics and quality constraints (§3). For this, one would typically identify possible sources of data (§4) from where to extract or sample relevant data through various modes (§5).

Data processing. Next, the acquired data is cleaned, organized and prepared for data analysis (§6). For this, the data points (e.g., users, messages, events) are often represented along a set of features (e.g., n-grams for content, demographic criteria for users), and are (manually or automatically) categorized along various dimensions (e.g., positive or negative messages, active users).

Data analysis. After data is collected, represented along a set of features, and perhaps sliced along a set of classes of interest, various computational and statistical methodologies can be applied. Their selection depends primarily on the specific settings and goals of each problem—such as describe, infer, predict, or extract causal relations—and on the peculiarities of the working datasets (§7).

Evaluation and interpretation. After the selected methods are applied, the results or models they generate are visualized and evaluated to see, e.g., if they prove or disprove a set of relevant hypotheses, and whether results can be generalized beyond a given context (§8). Often, researchers would also attempt to provide an interpretation of the results, which will depend on assumptions made about the meaning of each data point (e.g., whether a social link stems from a friendship relation, or a "like" equates to an endorsement).

Of course, these steps may not occur in a neat sequence and orderly fashion, with one step starting when the previous step ended. Additionally, these steps may not capture all the relevant details for all types of studies and they may occur out of order. However, conceptually they are useful to organize the presentation of the various pitfalls that may occur when using social data.

2.3. Validity of Social Data Research

For this discussion on the validity of social data research, let us assume that a researcher is analyzing social data for the purposes of proving or disproving a specific hypothesis. A research challenge, then, is an unaddressed issue within the research design and execution which might put the proof or disproof of the hypothesis into question. Research using social data is often interdisciplinary, and as such, the vocabulary and taxonomies that describe the challenges to research is varied [boyd and Crawford 2012; Howison et al. 2011; Lazer et al. 2014; Ruths and Pfeffer 2014]. Without prejudice, we categorize the various research challenges along the following classes of threats to the validity of research conclusions:

Construct validity or *Do our measurements over our data measure what we think they measure?* [Howison et al. 2011; Lazer 2015; Trochim 2006]. In general, a research hypothesis is stated

as some assertion over a theoretical construct or latent variable (or an assertion over the relationships between theoretical constructs) Construct validity asks whether a specific measurement (a calculation over a given dataset) actually measures the construct referred to in the hypothesis. However, while construct validity—and instability due to e.g., platform algorithms or usage conventions that are either cultivated or organically emerge—has been recognized as important to understand the reliability of measurements [Lazer 2015], only a few works using social data address it (see, e.g., Gilbert and Karahalios [2010]; Kramer [2010]).

For example, if a hypothesis states that "self-esteem" increases with age, research tracking self-esteem over time from social media must ask whether its assessment of self-esteem from text is actually measuring "self-esteem" versus other related or unrelated constructs. For example, are the observed behaviors (such as words used or frequency of posting) driven primarily by self-esteem as opposed to due to community norms, variations in system functionality, or other individual aspects.

Internal validity or *Does our analysis correctly lead from the measurements to the conclusions of the study?* Internal validity focuses on the analysis and assumptions of the data [Howison et al. 2011]. This survey covers subtle errors of this kind, such as biases that can be introduced through data cleaning procedures, the use of machine learned classifiers, mistaken assumptions about data distributions, and other inadvertent biases introduced through common analyses of social media.

For example, an analysis of whether self-esteem increases with age may not be internally valid if data cleaning accidentally removes messages expressing confidence; or if machine learned classifiers were inadvertently trained to recognize self-esteem only in younger people. Of course, while we do not dwell on them, researchers should also be aware of more blatant logical errors—e.g., comparing the self-esteem of today's younger population to the self-esteem of today's older population would be consistent with but would not actually prove that self-esteem increases with age.

External validity or *To what extent can research findings be generalized to other situations?* External validity focuses on ways in which the experiment and the analysis could not represent the broader population or situation [Trochim 2006]. For example, effects observed on one social media platform may manifest differently on another platform due to platform differences, differing community or cultural norms [Lazer et al. 2014; Malik and Pfeffer 2016; Wijnhoven and Bloemen 2014]. This concept includes what is sometimes called *ecological validity*, which captures the extent to which an artificial situation (constrained social media platform) properly reflect a broader real-world phenomenon [Ruths and Pfeffer 2014]; as well as *temporal validity*, which captures the extent to which constructs change over time [Howison et al. 2011] and invalidate conclusions (e.g., why people connect on social networks and why topics trend vary over time [Celik et al. 2011]).

For example, even after we conclude a successful study of self-esteem in a longitudinal social media dataset, its findings may not generalize to a broader setting because of worries that the kinds of people who self-select into a particular platform are not representative of the broader setting; or that the behaviors they express online may not be representative of their behaviors in other settings.

Each of these criteria is complex to define and complex to evaluate, and are also general to many types of research beyond social data analyses; the interested reader can consult Howison et al. [2011]; Lazer [2015]; Trochim [2006]. The specific challenges of each research program are determined by the particular objectives and research questions researchers are trying to answer. For instance, a study seeking to improve the ordering of photos shown to users in one photo sharing site may not need to be valid for other photo sharing sites. In contrast, a study seeking to uncover how public health issues in a country are reflected in social media may aspire to ensure that the results are independent of the particular social media sites selected for the study. While these challenges have been determined to be important, they are also believed to have been neglected [Crawford and Finn 2014; Gayo Avello et al. 2011; Ruths and Pfeffer 2014; Tufekci 2014].

3. GENERAL CHALLENGES

This section deals with general challenges of research done on social data. They include population biases (§3.2), behavioral biases (§3.3), content biases (§3.4), linking biases (§3.5), temporal

variations (§3.6), and redundancy (§3.7). The challenges refer to concepts of *data quality*, *representativeness*, *sparsity*, *noise*, and *bias*, that we describe next.

3.1. Concepts

Data quality. The impact of each type of data quality issues varies with the analysis task. In general, the quality of a dataset bounds the type of questions that can be answered with it and whether these answers are valid or not, regardless of the methods being used. For social datasets, researchers have often little leverage to control data quality as they gather data from platforms outside their control. In this context, a good understanding of what data quality aspects are important, what factors that may impact them, and what are some of their main ramifications is particularly important.

Representativeness. Sampling is so prevalent that we rarely question it. Sampling means taking a small portion of elements (a "sample") from a larger population, typically to learn something about the larger population. The key question in sampling is whether the sample is *representative* or not of the larger population.

Coming back to the classification presented in (§2.1), this affects research questions of type I (internal to a platform) that may need to focus on certain subgroups of users. Yet, identifying such groups is not trivial, as the available data may not capture all relevant properties of users. Research questions of type II (about external human phenomena) are further complicated. Often, they come along with a definition of a target population of interest [Ruths and Pfeffer 2014] such as estimating the political preferences of young female voters or of citizens with a college degree.

We have several ways in which to express our representativeness objective: (a) representativeness with respect to some *offline population*, (e.g., women who play online games); (b) representativeness with respect to some notion of *relevant content* (e.g., content related to sports); (c) representativeness with respect to some *behavior* (e.g., people expressing happiness).

Sparsity. Social data suffers from sparsity—many measures follow a power low distribution, which makes them easy to mine on the "head," (i.e., in relation to frequent elements or phenomena) but difficult on the long-tail (i.e., rare elements or phenomena) [Baeza-Yates 2013]. This phenomenon can be further exacerbated by platform functionality design (e.g., limiting or not the size of users' posts) [Saif et al. 2012], particularly affecting data retrieval tasks [Naveed et al. 2011].

Noise. Social data suffers from noise (e.g., content that is not reliable or credible, content that is incomplete or corrupted, typos, infrequent terms, stop words) [boyd and Crawford 2012; Naveed et al. 2011]. The distinction between what represents noise and what signal is often unclear, depending in subtle ways also on the question being asked [Salganik 2017]. As Baeza-Yates [2013] points out, often, the problem is not finding more data, but the *right data*—e.g. if adding more data also increases the level of noise, the quality and reliability of the results may deteriorate.

Bias. The representativeness of a sample can be compromised if it has biases. Biases can be quite pervasive in some fields. For instance, looking at biases in data collections used for object recognition research, Torralba and Efros [2011] found that datasets can be uniquely identified due to built-in biases, even when the dataset assembling goals were similar; consequently, one may identify the dataset a specific data entry comes from. Though the goals might often differ, social datasets are not foreign to such biases either: they also appear to exhibit built-in biases due to how the datasets are assembled [González-Bailón et al. 2014a; Olteanu et al. 2014a], along with biases and quality issues peculiar to this type of data (e.g., behavioral biases due to community norms).

Data biases[3] and other data quality issues may arise from how the data is collected, processed or created, and may manifest in the quality and the type of content or users captured by the working datasets. Thus, they can be classified per many criteria, including *what are the observed effects* (e.g., systematic differences among demographic groups) and *what factors trigger them* (e.g., mechanisms

[3]By data bias we mean a systematic distortion in the data. This is often measured by contrasting a working data sample with reference samples drawn from different sources or contexts, as obtaining an absolute data sample is often unfeasible (as we will see in §5). Thus, in the context of social data, bias is often a relative concept [Mowshowitz and Kawaguchi 2005].

on a platform). In the rest of the section, we discuss data biases from the perspective of how the distortions, regardless of their underlying causes, manifest in social data.

3.2. Population Biases

Def.: **Population biases**. Differences in demographics or other user characteristics between a population of users represented in a dataset or platform and a target population.

Population biases affect the representativeness of a sample and the ecological/external validity of research. They are problematic for research of type II (§2.1), in which conclusions about society in general are sought from the data. For instance, the relationship between a studied population (e.g., adults on Twitter declaring to live in the UK) and a target population (e.g., all adults living in the UK) is often unknown. Population biases typically refer to the extent to which data collected from a specific social platform differs from a specific target population. Yearly surveys from the Pew Research Center[4] of social media usage show that the demographic composition of the major social platforms consistently differ both with respect to each other, as well as with respect to the offline or Internet population [Duggan 2015; Duggan et al. 2015]. Further, differences between a platform's users and target populations may be exacerbated within narrower contexts, such as in the case of studies focused on certain user traits or application domains. Such observations are also supported by academic studies [Hargittai 2007; Mislove et al. 2011; Ruths and Pfeffer 2014].

Next, we cover a few main themes under which prior work has described population biases:

– Different user demographics tend to be drawn to different social platforms, or social platforms users do not represent society. Prior surveys and studies on the use of social platforms show gaps in gender representation across platforms [Anderson 2015], but also along race, ethnicity, and parental educational background [Hargittai 2007]. For instance, Mislove et al. [2011] finds that Twitter users significantly over-represent men and the population of regions that are densely populated, while women are over-represented on Pinterest [Ottoni et al. 2013]. Yet, Archambault and Grudin [2012] study of social media use among a tech giant employees shows that though growth in use and acceptance across social media platforms is not uniform, privacy and other user concerns that deter some from adopting them level off over time.

– Users of different demographics or traits use social platforms mechanisms differently. Users of different countries tend to use Twitter differently—Germans tend to use hashtags more often (suggesting a focus on information sharing), while Koreans tend to reply more often to each other (suggesting conversational purposes) [Hong et al. 2011]. Other example may be a question-answering site in which the culture encourages hostile corrections, driving users to remain "unregistered and passive."[5] Thus, studies assuming a certain usage may miss-represent certain groups of users.

– Proxies for personal traits or demographic criteria vary in reliability, and most users do not self-label along these axes. For example, a study interested in the opinion of young college graduates about a new law may rely on a proxy population: those reporting on a social platform to be alumni of a given set of universities. This choice can end up being an important source of bias [Ruths and Pfeffer 2014]. In the context of predicting users' political orientations, researchers have shown that the choice of the proxy population drastically influences the performance of various prediction models [Cohen and Ruths 2013]. In other words, it highlights that the quality of the inference of such latent attributes varies across a dataset demographics, the study showing that existing classifiers tend to do much better on politicians than on "normal" users (with only a few political posts) producing misleading estimates of users' political orientations on Twitter [Cohen and Ruths 2013].

[4] Pew Research Center: http://www.pewinternet.org/

[5] Tessa Harmon: "Stack Overflow's Developer Survey Analysis Hurts Women." *Medium*, April 2019. Online: https://medium.com/@glitterwitch/stack-overflow-s-developer-survey-analysis-hurts-women-ec4d568e2352

3.3. Behavioral Biases

Def.: **Behavioral biases**. Differences in user behavior across platforms or contexts, or across users represented in different datasets.

Behavioral biases affect the ecological/external validity of research, as they may condition the results of a study on the platform and context chosen. They affect both types I and II of research (§2.1), and are not entirely dependent on population biases.

For instance, Teevan et al. [2011] found web and microblogging search to capture distinct use cases: queries on Twitter are shorter and more popular, focusing more on temporally relevant information and people, while web queries tend to change and develop as users learn more about a topic. Observing the interplay between what people search and share about health on social media, De Choudhury et al. [2014] found information seeking and sharing practices to be dependent on the condition type like being a serious condition or not. Leskovec et al. [2009] compared news media with weblogs, finding a few hours lag between the attention peak of a meme (short phrase) in news media versus weblogs. Other studies also look at the similarities and differences among various social platforms w.r.t. adoption patterns [Kwon et al. 2014], user personalities [Hughes et al. 2012], news spreading [Lerman and Ghosh 2010], geographical and socioeconomic patterns [Li et al. 2013], shared content [Ottoni et al. 2014], and behaviors [Lim et al. 2015].

We analyze separately behavioral biases affecting the generation of content by users (§3.4) and those affecting linking patterns between users (§3.5). Three other classes of behavioral biases are those involving the interaction among users, the interaction between users and content, and the biases that cause users to be included or not into a study population.

– Interaction biases affect how users interact with each other. There are differences in how people communicate on social platforms that are influenced by the relationships they share, and by the features of the platform itself. Wilson et al. [2009] showed that the interaction patterns do not follow explicitly created social links, being significantly sparser, with 20% of friends accounting for 80% of interactions. Backstrom et al. [2011] found differences in how users balance their attention within their social network with some users being more focused than others, which also depends on their demographics (e.g., female users tend to be more focused towards their top friends); while others indicate that how users interact depends on the type of relation they share [Burke et al. 2013] and on shared characteristics (i.e., homophily) [McPherson et al. 2001; Wu et al. 2011].

– Content consumption biases affect how users interact with content, due to variations in interests or types of sought content. User consumption behavior is correlated with their demographic attributes and other personal characteristics: Goel et al. [2012] shows how page views vary across demographics (e.g. age, gender, race, income), while Kosinski et al. [2013] links "likes" on Facebook with personal traits. Given that users tend to consume more content from like-minded people, such consumption biases are linked to the creation of filter bubbles [Nikolov et al. 2015].

– The nature of users' tasks influences the traces they leave. Studying web search behavior, Silvestri [2010] found that it varies across semantic domains, and Aula et al. [2010] that it changes as the task at hand becomes more difficult. Further, Bennett et al. [2015] show that many such tasks involve searching on behalf of others and the traces users leave depend on the beneficiary's "persona"; while Oeldorf-Hirsch et al. [2014] found the type of user needs to impact what platform was used for information seeking tasks.

– Misreports and self-selection may occur due to behavioral biases; studies relying on self-reports on a certain aspect may be biased due to *what* users chose to report, *when* they chose to do it, and *how* they chose to do it. Gong et al. [2015, 2016] show that many users remain silent despite their interest in a given topic. This could be in part due to some users practicing self-censorship by either deciding not to share a post (often in the last minute) [Das and Kramer 2013] or by deleting content right after posting it [Wang et al. 2011]. In addition, users have a higher propensity to talk about their extreme or positive experiences, rather than about their negative or common experiences [Guerra et al. 2014; Kıcıman 2012].

Apart from "missing" reports, inaccurate self-reports can also bias social datasets—termed as *response bias*. Zhang et al. [2013] found that about 75% of Foursquare check-ins do not match users' real mobility; misreports that Wang et al. [2016] found to be motivated by competitive elements employed by social platforms such as badges. Discrimination can also occur as an artifact of such reporting biases *due to economic inequality* if they are overlooked [Barocas 2014; Crawford 2013] (see also the "digital divide" on §9.4).

3.4. Content Production Biases

Def.: **Content production biases**. Behavioral biases that are expressed as lexical, syntactic, semantic, and structural differences in the contents generated by users.

Content production bias is a behavioral bias that has raised concerns and received significant attention as they affect several popular tasks such as user classification, trending topics or content filtering [Cohen and Ruths 2013; Nguyen et al. 2016; Olteanu et al. 2014a], and may also impact user exposure to a variety of information types [Nikolov et al. 2015].

– The use of language(s) varies across and within countries and populations. By mapping the use of languages across countries, Mocanu et al. [2013] observed seasonal variations in the linguistic composition of each country, as well as between geographical areas at different granularity scales, even at the level of city neighborhoods. Rao et al. [2010] offer insights into distinctive language-usage variations across gender, age, regional origin, and political orientation on Twitter.

– Contextual factors impact how users talk. The use of language is even shaped by the relations among users, Burke et al. [2013] showing how mothers and fathers use language differently when they speak with their daughters and sons, and vice versa. Further, Schwartz et al. [2015] show that the temporal orientation of messages (emphasizing on the past, present, or future) may be swayed by factors like openness to experience, number of friends, satisfaction with life, or depression.

– Different populations have different propensities to talk about certain topics. For instance, by selecting political tweets during 2012 US election, Diaz et al. [2016] noticed a user population biased towards Washington, DC; while Olteanu et al. [2016] found African-Americans to be more likely to use on Twitter the #BlackLivesMatter hashtag (related to a large movement on racial equality).

– Content from popular or "expert" users differs from regular users' content. For instance, Bhattacharya et al. [2014] found that on Twitter "expert" users tend to create content mainly on their topic of expertise, while Zafar et al. [2015] show how focusing the sampling of content on "expert" users (e.g., whose who appear across topic-specific lists) biases the resulting sample towards more trustworthy and high-quality content.

3.5. Linking Biases

Def.: **Linking biases**. Behavioral biases that are expressed as differences in the attributes of networks obtained from user connections, interactions or activity.

Misrepresenting both the population or the behavior on a platform may also lead to disparities in the attributes of various types of networks that can be constructed from social data (e.g., interaction, interest or social networks). For instance, some user behaviors appear to be associated with various social network attributes; Kıcıman [2010] found differences in behavior (as indicated by language models and pronoun usage) that correlate with users' follower count; while Dong et al. [2016] found age-specific degrees of separation, with younger people being better connected than older generations. Geography has also been linked to the properties of such online networks [Poblete et al. 2011], with the likelihood of a social link decreasing with the distance among users, which has consequences on information diffusion [Volkovich et al. 2012]. Conversely, the network structure can also distort users' observations about their peers, systematically biasing social perceptions and resulting in the overestimation of the prevalence of certain attributes within a population, or of

the popularity of an event [Lerman et al. 2016]. Such biases can also occur due to how the data is sampled, impacting both the properties of the interaction network among users, as well as user attributes and position in the network [De Choudhury et al. 2010; González-Bailón et al. 2014b].

3.6. Temporal Variations

Def.: **Temporal variations**. Differences in populations or behaviors over time.

Temporal variations affect the temporal validity and both the internal and ecological/external validity of research. They are problematic for both type I and type II research (§2.1), as they may affect the ability to understand the generalizability of observations over time (e.g., what factors vary and how they confound with the current patterns in the data). If overall a platform or offline context are not stable, it may be impractical to disentangle the effects due to a specific variable of interest from variations in other possible confounding factors. At the level of the system that collects digital traces from users, Salganik [2017] points to three types of variations (which he calls *drifts*): variations in who is using the system, in how the system is used, and in the system affordances.

Thus, the temporal parameters including the boundary specifications of a dataset should not be overlooked. For instance, when tracking social signals at an aggregate level, there are general variations regarding when and for how long users focus on certain topics that may be triggered by current trends, seasonality or periodicity in activities, surprise factors, or even noise [Radinsky et al. 2012]. How one aggregates and truncates datasets along the temporal axis impacts what types of patterns we can observe and what types of research questions we can answer, which we detail next:

– Populations, behaviors, and systems change over time, as neither a platform population (or their behavior), nor its' suite of affordances is static, but they evolve in time, exhibiting significant temporal dynamics [Lampe et al. 2008; Radinsky et al. 2012; Salganik 2017]. Studies on Facebook [Lampe et al. 2008] and Twitter [Liu et al. 2014] show that users' demographic composition and the way in which the platforms are used changed over time. Even the demographic composition and the level of participation for users posting on a topic (e.g., elections) is often non-stationary over time [Diaz et al. 2016; Guerra et al. 2014]. For instance, people may use a hashtag at the beginning of an event, then continue discussing the event without the hashtag [Tufekci 2014]. There are also complex interplays between behavioral trends on a platform (e.g., trends in the use of language) and the online communities' makeup and users' lifecycles [Danescu-Niculescu-Mizil et al. 2013]. Such trends can emerge organically (e.g., through changes in the user base), or can be driven by platform goals (e.g., new features rolled out).

– There are seasonal and periodic phenomena that can further trigger systematic or periodic variations in how a platform is used and by whom [Grinberg et al. 2013; Radinsky et al. 2012; Scheffler and Kyba 2016]. Studying neighborhood boundaries inferred from geo-located tweets, Kıcıman et al. [2014] found that conditioning the analysis on different temporal contexts (day vs. night, or weekday vs. weekend) changed the shapes of the inferred neighborhoods. Grinberg et al. [2013] found Foursquare check-ins to exhibit clear weekly patterns for several real-world categories (patterns similar to Twitter posts on corresponding topics), while Golder and Macy [2011] observed a relationship between the sentiment of tweets and cycles of sleep and seasonality—with such periodic patterns being also observed for query logs [Vlachos et al. 2004] and editorial activity on Wikipedia [Yasseri et al. 2012].

– Sudden-onset phenomena affect populations, behaviors, and platforms; including suddenly emerging patterns in the data (e.g., a spike or drop in activity) due to external events (e.g., sudden onset crisis like an earthquake or accident) or platform changes. Malik and Pfeffer [2016] show how introducing a new feature may result in a sudden jump in the activity on a platform, while real-world events may result in activity peak, which typically define temporally their corresponding social datasets (e.g., Crawford and Finn [2014] discuss the case of crisis situations).

– The time granularity can be too fine-grained to observe long-term phenomena; such as maintaining relatively constant patterns or evolving over long periods [Crawford and Finn 2014; Richard-

son 2008]. For instance, while social datasets corresponding to real-world events tend to be defined around activity peaks, distinct events may have distinct temporal fingerprints that such datasets may miss (e.g., disasters may have longer term effects than sport events). In addition, the temporal fingerprints of protracted situations such as wars may be characterized by multiple peaks. Further, Fourney et al. [2015]; Richardson [2008] observed that long-term search logs (as opposed to short-term, within session search information) provide richer insights into the evolution of users' interests, needs, or how experiences unfold over time (e.g., pregnancy or career evolution).

– The time granularity can be too coarse-grained to observe short-lived phenomena. This is important when tracking how some experience unfolds for a user, short-lived effects, or smaller phenomena at the granularity of, e.g., hours of minutes. Fourney et al. [2015] highlight shifting needs and experiences for pregnant users indicating that how one would align and truncate users' timelines may influence what type of patterns they capture. Similarly, by exploring the temporal evolution of outcomes of a variety of personal experiences, Olteanu et al. [2017] also show that they follow different temporal patterns. Further, some of the patterns and correlations observed in social data may be evolving or only short-lived [Starnini et al. 2016].

– Datasets decay and lose utility over time—an important effect of temporal variations, not only while a dataset is collected, but also after. It has been linked to the deletion of content (by the users or by the platform [Gillespie 2015; Liu et al. 2014]), as well as to the restrictive nature of the social platforms APIs' terms of service (that prevent sharing of the datasets as they were collected). This renders datasets unusable over time as it makes them impractical to reconstruct, leaving important holes—dubbed as the "Swiss Cheese" decay of social datasets [Bagdouri and Oard 2015]. Maddock et al. [2015] found that from tweets collected during the Boston Bombings in 2014, more than 13% were unavailable later; while Almuhimedi et al. [2013] found that about 2.4% of tweets posted during one week in 2013 by a fixed set of about 300 million users were later deleted (with about half of users deleting at least one tweet from the same period), and Bagdouri and Oard [2015] found that from tweets in Arabic posted in October and December 2014 about 2.3%–3.6% were later deleted. There are several mechanisms rendering a message unavailable later [Liu et al. 2014]: the message was explicitly deleted by the user; the user switched their account to "protected" or private (thereby making their messages only available to other approved users); the user's account was suspended by the platform; or the user deactivated their entire account. Yet, often, it may be unclear why certain posts were removed, and hard to gauge how such deletions may impact the analysis results.

3.7. Redundancy

> Def.: **Redundancy**. Single data items that appear in the data in multiple copies, which can be equal to each other (duplicates), or approximately equal to each other (near duplicates).

Redundancy, when unaccounted for, may affect both the internal and ecological/external validity of research, being problematic for both type I and type II research (§2.1). It may negatively impact the utility of tools [Radlinski et al. 2011], and distort the quantification of phenomena in the data. Lexical (e.g., duplicates, re-tweets, re-shared content) and semantic (e.g., near-duplicates or same meaning, but written differently) redundancy often constitutes a significant-fraction of content [Baeza-Yates 2013], and may occur both within and across social datasets.

There are several sources of content redundancy including the same entity posting from multiple accounts or on multiple platforms, multiple users posting from the same account, or multiple accounts posting the same content—without necessarily referring to plagiarism, but cases such as tweeting quotes suggested from certain websites like news websites, or other types of content sharing such as re-tweeting. However, it is also important to note that redundancy can also be a signal by itself, e.g., if a message is reposted many times, it may also be a signal of importance.

4. ISSUES AT THE DATA SOURCE OR ORIGIN

The "garbage in – garbage out" principle in computing implies that a system that receives the wrong data often gets the wrong conclusions. This means that attention should be paid at the source of social data being used. The behaviors we observe in online social platforms can hardly be considered as "naturally occurring" [Salganik 2017]. Instead, they are determined by platform capabilities and engineered towards certain goals [Gillespie 2015; Tufekci 2014].

In this section, we first overview biases due to platform design and affordances (§4.1) and due to behavioral norms emerging on each platform (§4.2). Then, we examine factors that are external to the social platform, but which may influence user behavior, including their likelihood to use the platform (§4.3). Finally, we briefly cover other common characteristics of social data at its origin, including the presence of non-individual accounts (§4.4).

4.1. Functional Biases

> Def.: **Functional biases**. Biases that are a result of platform-specific mechanisms or affordances, that is, the possible actions within each system or environment.

Functional biases affect the ecological/external validity of research, and are problematic for type II research (§2.1). They make conclusions from studies difficult to generalize or transfer, as each medium or platform exhibit its own structural biases [Tufekci 2014], which can lead to platform-specific phenomena [Ruths and Pfeffer 2014]. However, due to the limits in the availability of social data, research to date has been concentrated around a handful of social platforms–typically the ones providing programmatic access to large volumes of up-to-date content. Particularly, Twitter has emerged as the model organism of social media research [Bruns 2013; Tufekci 2014].

The consequences of this bias towards certain platforms are often overlooked. Each platform carries its own suite of affordances—the set of actions that can be performed or that are encouraged, and the set of actions that are not supported or are hard to perform [Tufekci 2014]. These affordances are typically driven by the product goals of these platforms [Salganik 2017], including commercial and engineering considerations; and may shape behavioral norms emerging around each platform (as we shall see in §4.2). Further, each platform uses dynamic, proprietary, often undocumented platform-specific algorithms to organize and promote content (or users), affecting with what content and with whom users are more likely to interact on each platform. Such functional peculiarities may introduce population (§3.2) and behavioral biases (§3.3) by influencing what user demographics are likely to be drawn to each platform (resulting in the *misrepresentation of population*), and the kind of actions they are likely to perform (resulting in the *misrepresentation of behavior*) [Harford 2014; Ruths and Pfeffer 2014; Salganik 2017; Tufekci 2014].

Next, we discuss in detail three main related themes:

– Platform-specific design and features shape user behavior. Examples include observations of how introducing a new feature, or changing an existing feature on a platform, impacts usage patterns. Facebook observed that decreasing the size of the "reply" window to a posting, resulted in users sending shorter replies, faster, and more frequently.[6] Also on Facebook, Malik and Pfeffer [2016] showed that the introduction of the "People You May Know" feature elicited a significant increase in the creation of friendship ties. On Twitter, Pavalanathan and Eisenstein [2016] found that the introduction of emoji has resulted in a decrease in the usage of emoticons.

Other ways of observing the effect of platform design on user behavior are cross-platform studies, or studies of the elements that determine users joining or leaving a platform. Newell et al. [2016a] studied the differences among the book retailer Amazon and the social network for book readers Goodreads. Both platforms allow book reviewing and rating, yet differ in the content of the reviews, the ratings, and how reviews are promoted—being influenced by the design of the review feedback feature. Osborne and Dredze [2014] found that although distinct social media platforms (Twitter,

[6] From Facebook's Joel Seligstein, ICWSM'11 keynote, available at: http://videolectures.net/icwsm2011_seligstein_trends/

Facebook, Google+) provide a similar coverage of major events, there are clear differences in the content shared on these platforms. Since Twitter imposes a maximum posting length of 140 characters, while other platforms have larger maxima; the average posting length on Google+ (resp., Facebook) is about 5 times (resp., 10 times) larger than on Twitter. Users seem aware of the differences in features across platforms, highlighting several features as important in drawing them (or not) to a platform such as interface aesthetics, voting functionality, community size, as well as the diversity, regency and quality of the content available on the platform [Newell et al. 2016b].

– Algorithms used for organizing and ranking content influence user behavior. The engagement of users with content, and with other users, is affected by *what* information is provided to them, *when* it is provided, and *how* it is provided—which are often algorithmically determined. The effect of algorithms on data has been dubbed "algorithmic confounding" [Salganik 2017].

A typical example is a ranked list of content (e.g., search results) that "buries" content found in the lower positions, due to click and sharing bias or users perceiving content ranked higher as more trustworthy [Hargittai et al. 2010]. Such features may provide an advantage to, e.g, certain ideological or opinion groups [Liao et al. 2016]. Personalized rankings further exacerbate these issues, Hannak et al. [2013] observing that, on average, about 12% of Google search results exhibit differences due to personalization. Another example are recommendation algorithms that promote products that users may like, influencing their consumption patterns [Konstan et al. 2012]. News feeds depicting recent and relevant activities by contacts—a popular feature on social media platforms—are also a type of content recommendation. For instance, Facebook's algorithmically constructed news feed may influence the ideological diversity of content that users are exposed to [Bakshy et al. 2015]. This has important societal implications as it can lead to less diverse exposure to content, or to being less exposed to content that challenges one's views [Resnick et al. 2013].

– The way in which contents are presented influences user behavior. How different aspects of a data item are organized and emphasized, or the way in which various data items are represented, also impacts how users interact with it and interpret it across platforms. For instance, Miller et al. [2016] show that variations in the way emojis are displayed across different smartphone platforms can lead to confusion among users, as distinct renderings of the same emoji are sometimes interpreted as having different meanings and/or emotional valence. Schwarz and Morris [2011]; Yamamoto and Tanaka [2011] show that augmenting search results and webpages with additional information, or making existing information more salient, can adjust users' perception of their reliability. Finally, Chang et al. [2016] show that user interface changes can influence information disclosure behavior and, more generally, the norms users adopt. Norms are the topic we discuss next.

4.2. Normative Biases

Def.: **Normative biases**. Biases that are a result of written norms or expectations about unwritten norms describing acceptable patterns of behavior on a given platform.

In addition to the set of affordances that each social platform allows, each platform is also characterized by certain behavioral norms. These norms usually take the form of expectations about what constitutes an acceptable use of each platform, which are shaped by varied factors including the value proposition of each platform, or the composition of their user base [boyd and Ellison 2007; Newell et al. 2016b; Ruths and Pfeffer 2014].

– Norms vary across platforms, communities, and contexts. Norms are shaped by user communities, but also by other elements, such as design choices, explicit terms of use, moderation policies, and moderator activities. In general, norms developing around a social platform may have an important role in shaping the datasets emerging from it in ways that are platform-specific [Tufekci 2014]. Users may exhibit different behavior on different platforms [Skeels and Grudin 2009]: e.g., they may find acceptable to share family photos on Facebook, but not on LinkedIn [Van Dijck 2013].[7]

[7]LinkedIn is a social networking site oriented to professional usage, https://linkedin.com/

Platform-specific norms can emerge in subtle ways. Cheng et al. [2014, 2015] show that negative feedback makes trolling behavior worse, with the quality of posts dropping after a negative evaluation; further, users who receive negative votes post more frequently, being more likely to downvote others. Platform-specific norms may also change over time due to, e.g., changing demographics or population shifts [McLaughlin and Vitak 2012; Ruths and Pfeffer 2014]. Norms are also sensitive to context, as the meaning of the same action or mechanism may change under different circumstances [boyd et al. 2010; Freelon 2014]. Tufekci [2014] discusses how the meaning of retweets or likes can "range from affirmation to denunciation to sarcasm to approval to disgust."

– The awareness of being observed by others impacts user behavior; which we dub as "online" Hawthorne effect[8], referring to "others" such as platform administrators, platform users, or researchers. In general, users are more likely to share unpopular opinions or to make sensitive and personal disclosures in private, anonymous spaces, than in public ones [Bernstein et al. 2011; Schoenebeck 2013; Shelton et al. 2015]. Contrasting users that disclose their name with those that remain anonymous on Twitter, Peddinti et al. [2014] finds anonymous users to generally be less inhibited to be active participants. This effect may further be amplified by users' privacy concerns, with Lindqvist et al. [2011] observing people checking-in more often in public locations (e.g., bars and restaurants) than in private ones (e.g., a doctor's office or home).

Differences have been observed in the type of traces users leave when trying to answer a question by searching on a search engine, compared to asking through social media [Oeldorf-Hirsch et al. 2014]. In social media, user posts are often shared publicly or with social connections, and this influences what users feel comfortable to share. Indeed, the survey and analysis by Beasley and Mason [2015] suggests a "positivity bias" in social media posts—with users being less likely to use negative words, feeling uncomfortable to talk about negative elements in public.

– Social conformity and "herding" happen in social platforms; and such behavioral traits may end up shaping user behavior as well. For instance, Preist et al. [2014] discuss how the use of competitive elements like point scoring or leaderboards may result in a normalization of behavior as users may emulate others in the community. Michael and Otterbacher [2014]; Muchnik et al. [2013] indicate that prior ratings and reviews introduce significant bias in individual rating behavior and writing style, creating a *herding effect*. Similarly, Sukumaran et al. [2011] observed that users conform to informal standards of comment length, time taken to write a comment, number of aspects covered, that are set by others when writing comments on news websites.

Further, Jackson [2016] discuss the effect of the "friendship paradox" in distorting people's perceptions of the norms in a network—and, thus, their behavior—when popular individuals behave differently compared to average individuals. In addition, users also tend to conform with reference points either available on the platform or inherited from society (dubbed as the *anchoring effect*); Pal and Counts [2011] show that user names influence the perception of the quality for the content they produce, with male or organization names resulting in a positive shift in perceptions.

4.3. External Sources of Bias

> Def.: **External biases**. Biases resulting from factors outside the social platform, including considerations of socioeconomic status, education, social pressure, privacy concerns, topical interests, language, personality, and culture.

Social platforms are not closed systems. They are open to external influences that affect the makeup of the user populations enticed to each platform, as well as their interests and activities. The demographic makeup affects the language, the topics, and the viewpoints observed on a platform [Flekova et al. 2016; Preoţiuc-Pietro et al. 2015]. Usually, as the social context changes, users may also change their behavior on a platform, which may in turn render the observations from past cross-sectional studies harder to compare with a current situation. Intrinsic characteristics of various external events that are reflected on social media also alter the composition of the datasets obtained in

[8]Hawthorne effect. Wikipedia https://en.wikipedia.org/wiki/Hawthorne_effect, accessed Dec. 2016.

relationship to each event [Olteanu et al. 2015]. In general, external factors may impact the various quality dimensions of social datasets like coverage or representativeness, yet they can be subtle and easier to overlook. Thus, depending on the problem at hand, the targeted context, and the semantic domain being studied, may affect the reliability of observations drawn from social datasets [Kıcıman 2012; Olteanu et al. 2015; Silvestri 2010]. We further cover several types of extraneous factors including social and cultural context, external events, semantic domains and sources.

– Cultural elements and social contexts are reflected in social datasets. The effect of a particular culture is typically demonstrated through transversal studies comparing a platform's usage across countries. As noted earlier, Hong et al. [2011] observed that, on Twitter, German-speaking users are more likely to use hashtags and include URLs in their messages, while Indonesian and Korean-speaking users seem to engage more in social behaviors, their messages being more likely to be replies or contain mentions of other users. Similarly, Yang et al. [2011] found the country from where users interact with a platform to be a key factor in predicting their questioning and answering behavior on social networking sites. Such factors even seem to explain biases observed in geo-spatial social datasets such as OpenStreetMap [Quattrone et al. 2015]. Their understanding is considered important for guiding the design of cross-cultural tools [Hong et al. 2011; Johnson and Hecht 2015; Yang et al. 2011], and discerning socio-economic phenomena [Garcia-Gavilanes et al. 2013].

In addition to cultural peculiarities, the social context of users in a broader sense (including socio-economic or demographic factors) also plays a role in users' behaviors and interactions. For instance, the way in which users are perceived affects their interaction patterns (e.g., the amount of shared content or the number of followers), as well as their visibility on the platform (e.g., how often they are followed, added to lists, or retweeted) [Nilizadeh et al. 2016; Terrell et al. 2016]. Nilizadeh et al. [2016] shows that "for users in lower quartiles of visibility, being perceived as female is associated with more visibility; however, this tendency flips among the most visible users were being perceived as male is strongly associated with more visibility".

– Contents from different semantic domains are treated differently, notably with respect to sharing, attention, and interaction patterns. Romero et al. [2011] found that distinct kinds of information tend to spread differently within a shared online environment, while Olteanu et al. [2015] show that during crises, distinct kinds of users tend to focus on different types of information; thus, depending on the users one focuses on, distinct patterns will be more salient in the data. Further, due to both automated mechanisms and human curation, social media also exhibits common forms of bias present in traditional news media [Lin et al. 2011; Olteanu et al. 2015; Saez-Trumper et al. 2013], including gatekeeping (preference for certain topics or stories), coverage (disparity in attention), and statement bias (differences in how the information is presented) [D'Alessio and Allen 2000].

– High-impact events, whether anticipated or not, are reflected on social media. Just like traditional media, high-impact sudden-onset events (e.g., disasters), as well as seasonal phenomena (e.g., Ramadan or Christmas) tend to be covered prominently by users in social media. Their prominence not only determines how likely users are to mention them, but also what are they likely to say after on social media [Fourney et al. 2015; Olteanu et al. 2017], just as the characteristics of a disaster situation leave a characteristic "print" on social media (e.g. time and duration), including variations in the kind of information that is posted, and by whom [Olteanu et al. 2015; Saleem et al. 2014]. Seasonal events such as winter or summer holidays, also influence the prevalence of topics and languages [Mocanu et al. 2013], as well as the mood users express [De Choudhury et al. 2013].

4.4. Non-individual accounts

> Def.: **Non-individual agents**. Interactions on social platforms that are not produced by individuals, but by accounts representing various types of organizations, or by automated agents.

Researchers have noted that "studies of human behavior on social media can be contaminated by the presence of accounts belonging to organizations" if such accounts are not accounted for [McCorriston et al. 2015]. This is important as having an active presence on social media is a common practice for organizations (such as NGOs, government, businesses, media). For instance, in a study

of the #BlackLivesMatter movement, about 5% of the Twitter accounts that included the #BlackLivesMatter hashtag in their tweets were organizations [Olteanu et al. 2016]. The content these accounts produce may be large in some cases, as they were found to be responsible for over 60% of crisis-related tweets on average [Olteanu et al. 2015].

Another important type of non-individual account are bots and spamming accounts which are increasingly prevalent [Abokhodair et al. 2015; Ferrara et al. 2014]. While such accounts are typically used to manipulate various measurements (typically for increased visibility), not all bots are harmful. Some of them are used to post important updates about weather or other topics, such as emergency alerts. Broadly, the challenge is how to effectively separate them from accounts operated by individual users [boyd and Crawford 2012; Crawford and Finn 2014; Ruths and Pfeffer 2014]. Note that simply identifying and removing them from the analysis will often not be enough as the behavior of such accounts (e.g., what content they post or who they "befriend") influences the behavior of human accounts [Ferrara et al. 2014; Wagner et al. 2012]. x

5. ISSUES INTRODUCED WHILE COLLECTING DATA

Def.: **Data collection biases**. Biases introduced due to the selection of data sources, or by the way in which data from these sources are acquired and prepared.

So far, we covered many ways in how choosing a certain data source affects the observations we make and thus the research results. This can be described as *source selection bias,* as datasets are strongly affected by their origin, due to platform-specific phenomena. The external/ecological validity of Type II research (§2.1) is affected by the data source choice; and the main reasons why this introduces biases have already been described: First, users of distinct platforms may have different demographics (population biases, §3.2). Second, users of distinct platforms may behave differently (behavioral biases, §3.3) due to functional biases (§4.1), as each platform enables different actions; or due to normative biases (§4.2), as each platform has its own written and unwritten norms.

Looking beyond source selection bias, or even if we consider only Type I research within the context of a specific social platform, several quality dimensions of the samples obtained from the overall platform data, like representativeness or completeness, have been questioned [González-Bailón et al. 2014b; Hovy et al. 2014; Tufekci 2014]. This section examines issues of data acquisition via crawling or APIs (§5.1), issues related to querying such APIs (§5.2), and data (post-)filtering (§5.3).

5.1. Data Acquisition

While biases and errors can occur at any step in a data analysis pipeline, the first opportunity to account for them is at data collection time.

– Many social platforms discourage data collection by third parties. Social media platforms may offer no programmatic access to their data, prompting researchers to use crawlers or "scrappers" of content, or may even actively discourage any type of data collection via legal disclaimers and technical counter-measures. For instance, LinkedIn's terms of service forbid "manual or automated software, devices, scripts robots, other means or processes to access, 'scrape,' 'crawl' or 'spider' the [s]ervices or any related data or information."[9] Trip Advisor forbids to "access, monitor or copy any content or information of this Website using any robot, spider, scraper or other automated means or any manual process for any purpose".[10] These are common terms of services for many online platforms.

To actively deter data collection, platforms make some data harder to gather or employ distinct interfaces when interacting with likely automated agents. Researchers collecting data from them against the wishes of their operators, typically use custom crawlers that perform some form of "stealth crawling" [Pham et al. 2016]. This creates gaps between the data a crawler can collect and what the platform shows to regular users [Gyongyi and Garcia-Molina 2005].

[9] https://www.linkedin.com/legal/user-agreement
[10] https://www.tripadvisor.com/pages/terms.html

– Programmatic access to data from a platform comes with limitations. Some platforms provide Application Programmer Interfaces (APIs) to access data, but they set limitations on the quantity of data that can be collected, and provide query languages of limited expressiveness [Bruns and Stieglitz 2014; González-Bailón et al. 2014b; Morstatter et al. 2013; Olteanu et al. 2014a] (we discuss the latter in §5.2). In general, legal and technical restrictions on API usage set boundaries to what data can be collected. One basis behind these limits is that social data is often proprietary, and considered a valuable asset by the platform's operator; having third parties copy substantial parts of it may reduce its competitive advantage. Rate limits and data quotas provide a way of preventing third parties, such as researchers, from collecting large datasets within a reasonable time frame.

– The platform may not capture all relevant data. The development efforts are naturally driven by the functionalities that are central to each platform. In some cases, developers may not be capturing and recording all user actions, e.g., to save development costs, minimize data storage costs, or simply because some user traces are easier to collect and measure than others.

For instance, social media researchers study the content people produce more often than they study the content people are exposed to. We often know what people write, but not what they read, and we may know who clicked on or "liked" something, but not who read it or watched it [Tufekci 2014]. While these may seem to capture different behavioral cues, they can be used to answer the same question, e.g., both what people write and read can be used to measure their interest in a topic. Yet, using one or another may result in different conclusions. For instance, exploring the differences between the explicit (created by users) vs. the implicit (interaction-based) social networks, Wilson et al. [2009] showed that the network constructed based on explicit links among users was significantly denser than the one based on user interactions.

– Platforms may not give access to all the data they capture. Some data collection APIs' restrictions stem from agreements between the platform and its users. For instance, social media datasets typically include only *public* content that has not been deleted, to which users have not explicitly forbidden access by setting their account as private, or to which users have given explicit access (e.g., through agreements or by accepting a social connection) [boyd and Crawford 2012; Maddock et al. 2015]. Free speech limitations, including censorship, may also rendered some content inaccessible [King et al. 2013]. This means that a fraction of relevant data is bound to be left out.

– Sampling strategies are often opaque. Depending on the social platform, the available APIs for accessing data may further limit what and how much of the public data we can collect; often without clear guarantees about the properties of the provided data. In practice, APIs may return only some of the elements matching a query. This can be, e.g., the top-k elements given some criteria (e.g., top-20 most watched videos), or a random sample of k elements from those matching a query. Yet, APIs may conceal specific details about how the returned data items are sampled from those satisfying a query [boyd and Crawford 2012; Bruns 2013; Maddock et al. 2015; Morstatter et al. 2013].

For instance, much research on Twitter relies on APIs that give access to at most 1% of the public tweets [González-Bailón et al. 2014a; Joseph et al. 2014; Morstatter et al. 2014, 2013]. These APIs were compared against the full data stream, finding disparities in the prevalence of content [Yates et al. 2016]. While some APIs seem to provide a uniform random sample of the relevant content [Morstatter et al. 2014], others lead to biases regarding e.g., the topical making of the relevant content [Morstatter et al. 2013] or the follower-followee network structure [González-Bailón et al. 2014a]. For an overview of the limits of various social platforms APIs, see Reuter and Scholl [2014].

5.2. Data Querying

Data access through an API involves communicating a set of criteria for selecting, ranking, and returning the data being requested. In general, these criteria constitute a *filter*, typically in the form of a *query*, and different APIs have different forms of expressing this query.

– APIs have limited expressiveness regarding information needs. Many APIs support various types of predicates to query data, which allow various degrees of expressiveness and of controls over the quality of the data collection [González-Bailón et al. 2014a]. They determine what possible

sampling/filtering strategies are allowed such as a list of geographical locations/regions, a list of keywords, a set of temporal intervals, or a list of platform users. Yet, the specific information needs of a particular research task might not be directly expressible within a particular endpoint of an API; and this may result in data loss and/or bias in the resulting dataset.

For instance, keywords-based sampling may over-represent content by traditional media [Olteanu et al. 2014a] or posted by social-media literate users, while geo-based samples may be biased towards users in the cities [Malik et al. 2015]. Further, not all relevant content may include the chosen keywords [Olteanu et al. 2014a] and not all content might be geo-tagged.[11]

In addition, the expression of complex information needs involving conjunctions or disjunctions of various elements may be limited, e.g., Twitter's Streaming API interprets a request containing geographical regions and keywords as a disjunction of both predicates.

– An information need may be operationalized to an API in different ways. The operationalization of relevant data given a query language is known as query formulation [Olteanu et al. 2014a; Sampson et al. 2015]. There may be more than one formulation to convey an information need to an API (e.g., "postings about the Olympic games"), and different choices may lead to distinct results.

In *data collections centered around locations*, even similar strategies to match relevant data e.g., locating social media messages using either their geo-tags or user self-declared geo-location, introduce different biases in terms of both user demographics and content [Pavalanathan and Eisenstein 2015]. In addition, the quality of geo-location tags was found to be dependent on time bounded phenomena [Dredze et al. 2016]. Studies relying on geo-tagged tweets often assume that the geo-tags "correspond closely with the general home locations of its contributors"; yet, by studying data from three distinct social platforms Johnson et al. [2016] found this locality assumption to hold in only about 75% of cases. In addition, in geo-tagged data is often hard to separate tourists from locals, or, more generally, it is hard to account for users' mobility [Malik et al. 2015].

In *data collections centered around samples of users*, the user-selection criteria may include features held at a lower rate by members of certain groups [Barocas and Selbst 2014], and the resulting proxy population might fail to correctly capture the population of interest [Ruths and Pfeffer 2014]. Further, filtering strategies centered around users may over or under-emphasize certain categories of users such as those that are highly-active on a target topic [Cohen and Ruths 2013].

Linking biases (§3.5) can be worsened by query formulations, which can even alter the communication networks that can be reconstructed from social media posts, and poorly specified queries to the data access APIs exacerbate the biases in network properties (e.g. clustering, degree of correlation) more than the APIs limitations [González-Bailón et al. 2014b].

– The choice of keywords in keyword-based queries shapes the resulting datasets. A recurrent discussion has been the problematic reliance on keyword-based sampling for building social media datasets [Bruns and Stieglitz 2014; Magdy and Elsayed 2014; Tufekci 2014]. Referring to the choice of keywords, González-Bailón et al. [2014b] emphasizes that it "is equivalent to specifying the boundaries of a data collection: working with the wrong list of keywords might cause relevant data to be missed." Different query assembling strategies lead to different performance trade-offs: relying on manually curated keywords about given events leads to more precise collections, but it also results in data loss; while, relying on general domain terms leads to more comprehensive collections at the cost of precision [Olteanu et al. 2014a].

Similar observations hold for hashtag-based collections: different hashtags used in the same context (e.g., during a political event) may be associated with distinct social, political or cultural frameworks, and, thus, the samples built on top of them may embed different dimensions of the data [Tufekci 2014]. Ultimately, hashtags are a form of social tagging (or folksonomies[12]), and even if we assume that all relevant data is tagged, their use is often inconsistent (varying formats,

[11] For instance, only 1%-2.9% on Twitter messages were geo-coded [Graham et al. 2014; Osborne and Dredze 2014], 1% on Facebook and 0.6% in Google Plus [Osborne and Dredze 2014], and only 60% of Twitter profiles contain a valid location [Hecht et al. 2011], while 70% of users were found to occasionally disclose their location [Li and Sun 2014].

[12] A form of ad-hoc categorization and labeling of the data within social systems [Specia and Motta 2007].

spellings or word ordering) [Potts et al. 2011]. While some attempts to standardize the use of hashtags in certain contexts exist (e.g., see OCHA [2014] for humanitarian crises or Grasso and Crisci [2016] for weather warnings) to better capture the main types of information that are of interest to relevant stakeholders, the data collections built on top of them may also miss relevant data, as they overlook possible communications among actors that may not respect these standards.

However, there are also some notable efforts to improve and automatize the data retrieval strategies to generate more optimal queries [Ruiz et al. 2014], by expanding and adapting user queries [Magdy and Elsayed 2014], by exploiting domain patterns for query generation and expansion [Olteanu et al. 2014a], by leveraging recent seen content to maintain a uniform sampling [Osborne et al. 2014], or by splitting the queries and run multiple in parallel [Sampson et al. 2015], in order to mitigate possible biases and improve the quality of the datasets at collection time (typically by improving their completeness or representativeness).

5.3. Data Filtering

Data filtering entails the removal of irrelevant portions of the data; sometimes this cannot be achieved during data acquisition, due to the limited expressiveness of an API or query language. The data filtering step at the end of a data collection pipeline is often called post-filtering, as it is done after the data has been acquired or obtained by querying (hence the prefix "post-"). In general, the choice to remove certain data items implies an assumption that these items are not relevant for a study. This is helpful when the assumption holds, and harmful when it does not.

– Outliers are sometimes relevant for data analysis. Outlier removal is a typical filtering step. A common example is to filter out inactive and/or unnaturally active accounts or users from a dataset. In the case of inactive accounts, Gong et al. [2015, 2016] found that a significant fraction of users, though interested in a given topic, choose to remain silent. Depending on the analysis task, there are implications to ignoring such users. For instance, studies that track users' opinions or interest over time at a population level by aggregating generated content may lead to misguided conclusions.

Similarly, research has shown non-human accounts have a steady presence on social media [McCorriston et al. 2015], as we discussed on §4.4. These accounts often have anomalous content production behavior, for instance, posting more content than regular accounts, or at more regular intervals. Despite not being "normal" accounts, they can influence the behavior of "normal" users [Wagner et al. 2012], and filtering them out may hide important signals.

– Text filtering operations may bound certain analyses. A typical filtering step is the removal of functional words and stopwords from textual content extracted from social media. Such words often come from standard pre-compiled lists of terms, including pronouns, articles, or prepositions; yet, they may embed useful signals about the valence of the text or the users' personality and psychology [Campbell and Pennebaker 2003; Pennebaker et al. 2003; Saif et al. 2014].

6. ISSUES INTRODUCED WHILE PROCESSING DATA

> Def.: **Data processing biases**. Biases introduced by data processing operations such as cleaning, enrichment, and aggregation.

The biases or assumptions of those designing and building the data processing frameworks can affect even the most basic levels of data organization, distorting datasets by altering the content, the structure, the organization, and the representation of the data [Barocas and Selbst 2014; Poirier 2015]. Data biases can be introduced, often in subtle manners, by data preparation operations like cleaning (§6.1), enrichment via manual or automatic procedures (§6.2), and data aggregation (§6.3).

6.1. Data Cleaning

The purpose of data cleaning is to ensure that the data faithfully represents the phenomenon being studied (e.g., to ensure construct validity). It corresponds to detecting and correcting errors and inconsistencies in the data, and typically it is considered successful when "cleaned" data can pass consistency and validation tests [Rahm and Do 2000].

Data cleaning may involve the removal of certain data elements, as well as the normalization by correction or substitution of incomplete or missing values. Such alterations can embed the scientist's beliefs about a phenomenon and the broader system into the dataset. While well-founded alterations improve a dataset's validity, data cleaning can also result in incorrect or misleading data patterns.

– Data representation choices and default values may introduce biases. Data cleaning involves mapping items, possibly from different data sources, to a common representation.[13] Such mappings may as well introduce subtle biases that affect the analysis results. For instance, if a social media platform allows both "textual" postings, and "image" postings, interpreting than an image posting without accompanying text has (i) null text, or (ii) text length of zero, can yield different results when computing average text length.

– The normalization of geographical references may introduce biases. We noted in (§5.2) that geographical references are a source of complexity when dealing with social data. Users of some social platforms have various choices for geographically annotating profiles and content. Cleaning may involve replacing missing values or making estimations to geo-locate objects within a location at a given geographical granularity (e.g., city or country level). This may introduce errors, for instance by mapping a description of a location to the coordinates of the center of the geographical bounding box containing a given location.[14]

6.2. Data Enrichment

Data enrichment involves adding annotations to data items, to be later used during the analysis phase. Annotations can range from simple categorical labels associated to each item, to more complex processing such as part-of-speech tagging or dependency parsing done on text. They can be obtained through either some form of (semi-)automatic classification, or through human-annotations (e.g., crowdsourcing, surveys), yet both are liable to errors [Cohen and Ruths 2013].

– Manual annotation often yields subjective and noisy labels, as many factors can affect the quality of human-annotations such as (i) unreliable annotators, (ii) poor annotation guidelines, (iii) poor category design, such as categories that are too broad, too narrow, or too vague, (iv) or insufficient information to make a reliable assessment [Cheng and Cosley 2013]. Though the goal of an assessment task is to provide human input, underspecification or appeal to subjective judgment can introduce unintended biases that are often difficult to detect. Take the example of user profiles annotation along some demographics factors. From inspecting a user profile, an annotator may be more likely to correctly identify a user gender than her age [Nguyen et al. 2014], and some categories may be easier to identify than others (e.g., "baby" may be a category in which annotators make less errors than "in his/her early 50s"). Such gaps across categories or data dimensions may introduce systematic biases in the data. Subjectivity and noise may be reduced by requiring more annotators for every item, but annotation budgets are finite, and thus there is a limit to how accurate manual annotation can be. Even when collecting straightforward binary judgments, a skewed class distribution may result in a poor representation of the less dominant class [Cohen and Ruths 2013].

– Automatic annotation through statistical or machine learning methods introduces errors. A wide range of automatic processes may be used to enrich data. Text can be processed through a Natural Language Processing pipeline that can be complex. Elements can be annotated with specialized classifiers or other types of annotators. What these processes have in common is that they apply some type of statistical or machine learning techniques, which are almost never 100% accurate.

For instance, automatic classification, a common operation of this kind, can introduce biases in the data. This is particularly problematic when the end goal is not the estimation of specific labels, but measuring the prevalence of these labels in the data (e.g., Gao and Sebastiani [2016] discuss why the distinction between the two tasks is important, and different evaluation metrics should apply).

[13] One example of this is Semantically Interlinked Online Communities (SIOC), an RDF standard for representing data from the social web: http://sioc-project.org/, accessed Nov. 2016.

[14] See, for example: https://medium.com/@ayman/the-social-concerns-of-geo-located-rectangles-9b361f34811d or http://fusion.net/story/287592/internet-mapping-glitch-kansas-farm/

However, many social data analyses rely on machine learned classifiers to classify first, and count later (e.g., Abbar et al. [2015]; Mislove et al. [2011]; Zagheni et al. [2014]).

In general, automatic classifiers used for data enrichment may not be robust across distinct datasets or not even across distinct classes of data within each dataset (e.g., it is easier to predict the political leaning of active users [Cohen and Ruths 2013]). For instance, Landeiro and Culotta [2016] show that when a confounding variable influences both the data features and the target class variable, the accuracy of a machine learned classifier can degrade rapidly if the the confound varies. Users of relatively common word embeddings such as those provided by `word2vec` or `GloVe` should be cautious about the gender biases introduced because of the data such embeddings are trained on [Bolukbasi et al. 2016; Caliskan-Islam et al. 2016].

6.3. Data Aggregation

Data aggregation is often performed in order to structure, organize, represent or transform data. Such transformations can either hide or give prominence to distinct, even divergent, patterns [Olteanu et al. 2014b; Poirier 2015]. For instance, consider pre-processing heuristics that aggregate the data to make it more manageable at the cost of losing information. The way in which these aggregations are done, or what information they compromise (e.g., aggregating content along users versus aggregating it along topics) may lead to different conclusions at a later-on analysis on the overall incidence of distinct topics across users (e.g., aggregating by user may give equal weight to each user's interests, while aggregating by topic may give more weight to content from highly active users). Furthermore, if the data is organized along a certain attribute (e.g., the presence of a keyword or hashtag), and there are multiple independent factors that result in the attribute taking a certain value, analyzing the data entries with this value is equivalent to conditioning on it, and may result in spurious patterns of association among these factors [Blyth 1972; Tufekci 2014].

7. METHODOLOGICAL PITFALLS WHEN ANALYZING THE DATA

We have discussed many kinds of biases in social data (before or after processing) that might slant a study results. This section surveys how the choice of methods to characterize (§7.1–§7.2), to make inferences and predictions (§7.3), as well as to distill relationships (§7.4) regarding user populations and behaviors, may also affect both the internal and external validity of social data research.

For instance, due to variations in data collection, but also due differences in the analysis methodology and measurement, Liang and Fu [2015] could not replicate or generalize 6 out of 10 known propositions from social media studies. However, selecting *the* methodology is hard as distinct approaches are often characterized by different strengths and weaknesses, and the choice typically reflects the researcher's perspective.

Using the right method, in the right place, and at the right time. Prior work raised concerns that the research agenda is opportunistically driven by access to data, tools or ease of analysis—e.g., readily available and easy to use programming libraries, or familiarity with certain research methods [Bruns 2013; Ruths and Pfeffer 2014; Tufekci 2014; Weller et al. 2015], or, as Baeza-Yates [2013] puts it, "we see a lot of data mining for the sake of it." Related concerns are:

– *Using data as a source of hypotheses* rather than a tool to test them or tailoring the research agenda based on data availability, which can result in bias in the type of questions being asked.
– *Multiple hypothesis testing* or testing multiple hypotheses until a significant, positive result is found–e.g., greedily test multiple features for classification tasks until finding one that delivers important improvements, instead of selecting it based on a priori hypotheses, termed as *feature hunting* [Ruths and Pfeffer 2014];[15] or, more generally,
– "*Type III Errors*" [Mitroff and Silvers 2010] or using the right method to answer the wrong (or poorly specified) question; like when a construct validity is uncertain as the variables may not correctly measure the theoretical constructs of interest, or when lacking theoretical grounding (and, e.g., comparing variables based on a posteriori knowledge of some results).

[15] This is also an example of *harking*, which is the practice of hypothesizing after the results are known [Kerr 1998].

– "*Type IV Errors*" [Adams and Hester 2012], or drawing the wrong conclusions about a method results, such as making the wrong assumptions about the data structure or about the independence of variables (e.g., collinearity and confounder bias), e.g., using only messages including a hashtag of interest to study the variables that determine its' usage [Tufekci 2014].

Hence, independently on data or methods being old, imported from other fields, adjusted, or new, an argument about their suitability and trade-offs is needed [Ruths and Pfeffer 2014; Tufekci 2014].

7.1. Qualitative Analyses

While the availability of large social datasets makes them suitable for quantitatively depicting behavior and populations, qualitative analyses ("small-N") are also used in social data research [boyd et al. 2010; Marwick 2014; Marwick et al. 2011; Tufekci 2014], either alone or in conjunction with quantitative methods.

Qualitative analyses tend to be in-depth, open-ended, and exploratory in nature answering questions about the *how*, *what* or *why* of a social phenomenon.[16] They help to coin hypotheses about phenomena to be quantified [Charmaz 2014], can be used for in-depth explorations of quantitatively inferred results or data samples to validate or discern the nuances of their social meanings [Cranshaw et al. 2012; Olteanu et al. 2017; Tufekci 2014], or to develop codebooks to quantitatively code larger corpora [Vieweg et al. 2010]. Further, qualitative methods like in-depth interviews with users can aid trace analyses and explore, for instance, how social media usage affects social ties [Burke and Kraut 2014] or what elements trigger changes in usage over time [Lampe et al. 2008].

Yet, though rich in details and illuminating when mixed with quantitative methods [Creswell and Clark 2011; Marwick 2014], qualitative methods have known limitations when used in isolation. They tend to compromise the generalizability (or external validity) for details [Trochim 2006], particularly due to their limited scope (e.g., limited sample size [Lampe et al. 2008], time [Burke and Kraut 2014] or context [Vieweg et al. 2010]). They also tend to be sensitive to the interpretation biases of researchers, are challenging to scale, and it is difficult to draw quantitative evidence from qualitative observations (e.g., we may learn that people sometimes reposts content they dislike, but we do not know how prevalent this is).

7.2. Descriptive Statistics

Descriptive analyses are the basis of many studies, quantitatively depicting social data with numerical or graphical summaries of quantities such as the average number of messages across user demographics [Olteanu et al. 2016], the geographical distribution of messages [Leetaru et al. 2013], or temporal associations among topics of interest [Fourney et al. 2015]. These analyses may measure the distribution, variability, or correlations among variables of interest, such as Java et al. [2007], one of the first studies to characterize the growth, the topological and geographical properties of Twitter with descriptive statistics. However, given that the goal of such work is to summarize complex datasets in a manageable way, they may also conceal important details, potentially misleading the reader into the wrong conclusions.

– Social data research often relies on counting entities; such as users, links, or messages, and computing summaries for those counts [Lazer et al. 2014; Salganik 2017]. Yet, simple counts can mislead if it is unclear what is counted and why. Salganik [2017] points to Back et al. [2010] that found a steady rise in feelings of anger on Sept. 11, 2001 after the attacks in NYC based on pager messages. The finding was later debunked by Pury [2011], showing the rise to be due to a repeated message coming from a single pager. Based on how and when a distinction is made between content created by users and what they re-shared from others (e.g., tweets vs. retweets), such confusions may also occur in other studies—e.g., when studying volume-based trends or the use of language, falling to make this distinction may lead to bias and may affect the study validity.

Count-based analyses are also sensitive to confounders and issues with construct validity. For instance, popular strategies to characterize the emotional state of users rely on counting affectively

[16] Qualitative research is a complex methodological area. See, e.g., the textbook by Silverman [2013].

positive and negative terms. Yet, Beasley and Mason [2015] indicate that these terms frequency is an imprecise measure for how users truly feel; while [Kıcıman et al. 2014] show how when conditioning for possible confounding factors, neighborhood maps created based on the frequencies of co-visits highlight differently shaped neighborhoods. In addition, as many measurements follow a power low, summaries of centrality or dispersion like averages or ranges are often distorted, leading to paradoxes such as the friendship paradox [Jackson 2016] or the majority illusion [Lerman et al. 2016]. This adds to the problem introduced when the objects being counted are obtained through an automatic classification approach, as we discussed in (§6.2).

– Correlational analyses are sensitive to bias and confounders. Many studies assume that co-occurring patterns reflect true relationships, a common task is the extraction of associations among dataset variables (e.g., sources and types of information [Olteanu et al. 2015]) or with elements of the off-line world (e.g., food mentions on social media and obesity rates [Abbar et al. 2015]).

Such assumptions can be problematic as social data may not accurately capture relevant offline or online populations [Hargittai 2007], or user behavior may be distorted by both online or offline phenomena [Olteanu et al. 2015; Ruths and Pfeffer 2014] (as seen in §3). Many datasets are built around dependent variables, with the inclusion of, e.g., users or content depending on the inclusion of the variable under analysis [Tufekci 2014]; and this may result in apparent patterns of association that hold only in the presence of that variable. The key challenge is how to distill between attributes that merely correlate and those that are causally related. For instance, Liang and Fu [2015] show that correlations found by prior work between the URLs inclusion in tweets and how retweeted they were might in fact be induced by URLs co-occurring often with hashtags and, thus, spurious. Causal analyses aim to address such issues, as we shall see in (§7.4).

7.3. Inferences and Predictions

Beyond social data use for descriptive purposes, many studies also aim to draw conclusions beyond the dataset under analysis. Such studies use smaller (more manageable) samples to make inferences about unseen or larger populations, or use historical known measurements to predict their current (*"nowcasting"*) or future (*forecasting*) values using social data [Asur et al. 2010; Salganik 2017].

However, many such tasks have proved harder than early results suggest, with many reporting pitfalls around attempts to infer users' political orientation [Cohen and Ruths 2013], mood [Beasley and Mason 2015], location [Jurgens et al. 2015b; Pavalanathan and Eisenstein 2015], or gender [Nguyen et al. 2014], as well as to predict exit polls [Gayo Avello et al. 2011] or flu incidence [Lazer et al. 2014]. Limiting factors include uncontrolled confounders [Landeiro and Culotta 2016; Pavalanathan and Eisenstein 2015], bias in training or testing datasets [Barocas and Selbst 2014; Cohen and Ruths 2013], dealing with ambiguous cases [Nguyen et al. 2014], non-stationary user population and participation over time [Diaz et al. 2016; Guerra et al. 2014; Jurgens et al. 2015b], construct validity [Beasley and Mason 2015], or even data representation [Barocas 2014].

– There are performance variations across and within datasets. Even when a model achieves a good overall accuracy, errors may be unevenly concentrated on certain classes of messages or users [Cohen and Ruths 2013; Pavalanathan and Eisenstein 2015; Tramer et al. 2015]. Hardt [2014] uses fake name detection as an working example to exemplify how data patterns found for a majority of users may not hold for a minority group, resulting in higher error rates for the minority group. Indeed, several empirical studies show that the performance of existing inferences models is sensitive to various user-related confounds such as age or gender [Dos Reis and Culotta 2015; Landeiro and Culotta 2016; Pavalanathan and Eisenstein 2015].

User-related confounds are not the only culprits. For instance, Denny and Spirling [2016] shows that topic modeling techniques such as Latent Dirichlet Allocation (LDA)—frequently used in investigations of textual content created or shared by users—are sensitive to common pre-processing steps. Further, sometimes it may be ambiguous to what category a certain data point belongs, such as the age of a user; in general there are limits in the current approaches used to predict demographics of users based on the messages they post in social media [Nguyen et al. 2014].

– The composition of test and training data samples impacts the results [Cohen and Ruths 2013; Jurgens et al. 2015b; Nguyen et al. 2014; Pavalanathan and Eisenstein 2015]. For instance, using data samples biased towards users whose gender [Rao et al. 2010] or political identity [Cohen and Ruths 2013] are easy to discern, leads to overoptimistic performance estimations that do not reflect those obtained on balanced or representative samples [Cohen and Ruths 2013; Nguyen et al. 2014].
– Distinct target variables, class labels, or data representations may lead to different results. When dealing with "fuzzy" constructs for which there is no gold standard, studies often end up using varying definitions and proxies for the target variable (e.g., political leaning) and class labels (e.g., democrats or republicans), which can lead to results that are hard to compare or generalize [Cohen and Ruths 2013; Wong et al. 2013]. Even for less ambiguous constructs (e.g., user location) there can be multiple competing proxies, whose choice can impact the results as well. Jurgens et al. [2015b]; Pavalanathan and Eisenstein [2015] observed that the accuracy of text-based geo-location of Twitter users varies across samples, and depends on whether user-supplied location or the GPS coordinates of their tweets are used as a proxy for user location.

In general, the *data representation* or *features* selected to represent an object, such as a user or a message, impacts the results of inference tasks on those sort of objects. For instance, even if a user sample is representative, certain features may occur at lower rates in the messages of certain users [Gong et al. 2015]. For a comprehensive discussion on these issues (beyond social data research), see Barocas and Selbst [2014].
– The choice of the objective function can misguide the inference task. Risks are also linked to the selection of the objective functions used to express various inference or prediction tasks [Wagstaff 2012]; such as using a wrong objective function that does not match the inference methodology [Gao and Sebastiani 2016], or one that leads to undesirable behavior during the learning process or that is expensive to reliably evaluate [Amodei et al. 2016]. Similarly, at times a concrete objective function will only approximate the true objective. For example, in a web search scenario, the true objective criterion may be user satisfaction, but it is approximated by behavioral signals such as clicks or reformulations [White 2016]. Moreover, these surrogate objectives themselves might also be based on imprecise measurement or biased modeling [Mehrotra et al. 2016], and have the potential to create self-fulfilling feedback loops when decisions are made based on the inference results and the outcomes are fed back into the models as training data [Barocas 2014].

7.4. Observational Studies

Beyond describing what is happening or making a prediction, many studies also want to determine *why* something is happening, that is, causation. For this, a study would typically seek to compute the effect of a *treatment* (e.g., receiving a recommendation or using a social convention) on users. To do so, the gold standard are randomized controlled experiments [Aral and Walker 2011; Bakshy et al. 2012; Muchnik et al. 2013], but when experimentation is impractical or violates ethical standards, many studies resort to observational data collected from social platforms. Yet, even with active experimentation determining causality is hard. Unless the observed data approximates a randomized experiment—this is, even in the presence of a *natural experiment*—observational studies are even more challenging due to the difficulty of accounting for the effects of uncontrolled confounds.

Short of identifying natural experiments, however, under strong assumptions, there are methods that help assess causation in observational studies and/or mitigate the effects of confounding or selection bias;[17] including matched analysis [De Choudhury et al. 2016; Sharma and Cosley 2016], instrumental variables analysis [Sharma et al. 2015], regression discontinuities [Malik and Pfeffer 2016], differences-in-differences [Carmi et al. 2012; Zagheni et al. 2014], and others. Yet, with all of these methods there are critical caveats and strong assumptions that must be accounted for; otherwise, they are susceptible to various validity issues [Oktay et al. 2010].

[17] A comprehensive review of causal inference with observational data is beyond the scope of this survey. For more background on the topic, see [Nichols et al. 2007].

– Social data may not capture the entirety of users' lives. A key assumption made by causal analyses of observational data is that unobserved covariates can be ignored as they are independent of users getting the treatment or not—dubed the *ignorability* assumption. However, it is possible that some unobserved covariates such as environmental factors or individual characteristics and actions may in fact affect users' propensity to get the treatment. Without significant domain expertise this assumption is often hard to assert. For instance, network studies of peer influence and social contagions suffer from a stubborn challenge of disambiguating such effects from homophily among peers and within communities [Christakis and Fowler 2007; Lyons 2011; Shalizi and Thomas 2011]. While Christakis and Fowler [2007] found obesity to spread through peer influence in social networks, others suggest that unobserved confounds correlated with the social network structure may be the culprit rather than peer influence [Cohen-Cole and Fletcher 2008; Lyons 2011].

– Peer effects due to platform affordances and conventions may weaken causal analyses. Other key assumption is that the effect of a treatment on an individual is independent of the treatment status of others. Alas, this assumption is often violated in the presence of common social features (e.g., hashtags, messaging, community support) that provide value through network effects [Ugander et al. 2013]. For instance, observed covariates may include the terms used by users. Yet, as Olteanu et al. [2017] acknowledge, a conversation on a topic may include content (re)sharing or hashtags and, thus, one user use of a term may have some effect on other users in an online community.

– The identification of (non-)treated users may pose internal and construct validity threats. Social media studies often rely on self-reports to identify users that had a treatment (e.g., who mention online to have lost their job) by searching for certain terms [Dos Reis and Culotta 2015; Olteanu et al. 2017; Proserpio et al. 2016]. However, it may happen that not everyone that uses these terms had the treatment (and not all users that had the treatment use the searched terms) [Dos Reis and Culotta 2015; Proserpio et al. 2016]. Further, to identify a *control* group—used as baseline in causal analyses—social media studies employ various sampling strategies including random sampling, network based sampling (e.g., friends or followers), or topical or domain based sampling (e.g., select users taking different drugs than the one under test) to identify similar users with those treated, but that have not received the treatment [Dos Reis and Culotta 2015; Olteanu et al. 2017; Pavalanathan and Eisenstein 2016]. Yet, different such strategies may lead to different degrees of similarity among the treated users and the control group, and, thus, to different results [Oktay et al. 2010].

– Selection bias and how treatment effects are estimated affects results generalizability. First, many methods compute only the local average treatment effects for a selected (sub)population [Nichols et al. 2007], which may limit the generalizability of results to users with different characteristics than those included in a study. This is important for social data studies that typically suffer from self-reporting biases (as mentioned above and in §3.3), and are thus limited to the association patterns captured by each working datasets, also noted by [De Choudhury et al. 2016; Olteanu et al. 2017]. Second, there can be heterogeneity in the effects of a treatment across users, and, as a result, the average treatment effect (even when calculated under sound assumptions or for randomized experiments) may not generalize to all treated users [Taylor et al. 2014].

8. CHALLENGES WITH THE EVALUATION AND INTERPRETATION OF FINDINGS

A last opportunity to account for biases and gauge the reliability of findings is at evaluation time. If that is not possible, disclaimers on a study limitations are needed [Tufekci 2014]. This section covers issues with metrics selection (§8.1), and with the assessment and interpretation of findings (§8.2). Finally, concerns about the reproducibility of studies and the lack of negative results and disclaimers are also discussed (§8.3).

8.1. Metrics Selection

Metrics are used to quantify a phenomenon (e.g., popularity or interest), or the performance of some method or tool. However, they might be inconclusive, or suffer from reliability and validity issues.

– The choice of metrics mould a research study take-aways. Metrics often attempt to quantify the relationship between system decisions and a desired objective. For example, the effectiveness of a web search engine might be quantified by the click-through rate on the search results page. However, metrics attempting to the measure the same thing (e.g., user satisfaction) may be inconsistent with one another depending on the context. Jurgens et al. [2015b]; Olteanu et al. [2014b] review how computing the same metric (e.g., precision) in a user-centered versus in an inference-centered fashion (geo-inferences for social media posts or item recommendations) may show different performance trends as, e.g., the later may be biased towards active users (that post on social media or recommend more often) and may obfuscate the performance distribution across users. A user-centered approach may offer more reliable estimates of the expected error for arbitrary users, relevant for applications operating rather on users (e.g., flu trends) [Jurgens et al. 2015b]. In general, result metrics are aggregates and, thus, sensitive to the way in which the aggregation is done (§6.3).
– The assessment of performance should account for the domain impact; as the adequacy of a certain performance level depends on the cost of errors in each domain. Wagstaff [2012] raises concerns about the pervasiveness of abstract metrics, such as precision and recall, explicitly ignoring or removing problem specific details. These abstract metrics enable comparisons across domains, but offer limited insights about the impact within each problem domain; for instance, 75% precision may be appropriate for some applications (e.g., identify cat pictures for an image search engine), but not for others (e.g., identify criminal activities for law enforcement). Even if a metric indicates an overall low error rate for a classification task, it is hard to know what that implies [Hardt 2014]. It may as well mean a high performance was obtained for one class, but a low one for another [Hardt 2014; Konstan et al. 2012].

Further, there are questions about the stability and validity of abstract metrics [Sokolova and Lapalme 2009]. In social media research, the number of posts has been used as a proxy metric for the interest in a topic [Chen et al. 2010]; yet, while this number may reflect production patterns, it may not reflect how much content on the topic users read. In addition to these metrics, there is a need for metrics that measure impact in the problem domain, such as "dollars saved, lives preserved, time conserved, effort reduced, quality of living increased" [Rudin and Wagstaff 2014; Wagstaff 2012]. Finally, in some cases, metrics may themselves be designed using a statistical model, subject to the same biases presented in §7.3 [Diaz 2016].

8.2. Results Assessment and Interpretation

Much research rests upon the assumption that online social traces reflect in some quantifiable way real-world phenomena [Asur et al. 2010; Kıcıman and Richardson 2015; Rost et al. 2013]. Yet, this assumption has been challenged due to concerns with construct validity and stability over time [Freelon 2014; Lazer 2015; Tufekci 2014]. Further, Rost et al. [2013] argue that data explicitly generated by users should in fact be interpreted as communicative rather than representative.
– The meaning of social traces may change with context; yet, this is hard to discern. Rarely will a social network reflect homogeneous relations between individuals. Social links between users can stem from friendship, trust or shared interests, and thus can embed different social cues [Tang et al. 2012]. Likewise, sharing content can be a sign of endorsement or interest, but users may also share content to ridicule, disapprove, or bully. The same mechanism or process may capture different signals across contexts [Tufekci 2014], but such distinctions are often unintelligible and hard to make in an automated fashion when looking at data in aggregate [Rost et al. 2013; Tufekci 2014].

This unintelligibility is subject to functional biases (§4.1), as it depends on the mechanisms available on each social platform (e.g., having a like button, but not a dislike one), as well as on variations in platform algorithms and mechanisms in response to users actions [Lazer et al. 2014]. It is difficult to account for what is or not in the data when researchers lack proper context [boyd and Crawford 2012]—e.g., for social media use in crises, it may be hard for a geographically distant researcher to fully gauge the cultural context and the event peculiarities [Crawford and Finn 2014]. Distinct methodological alternatives may also lead to varying interpretations of what it is in the data [Bruns

2013]. As Tufekci [2014] advises, small qualitative data pull-outs may be useful to unpack varying meanings of the same process (see §7.1).

– Analyses should go beyond studies confined to a single dataset or method. The confinement of many studies to one dataset or analysis method has prompted calls for more comprehensive studies [Fraustino et al. 2012; Ruths and Pfeffer 2014]. Results of different methods to collect, measure or process the data should be routinely juxtaposed [Ruths and Pfeffer 2014; Tufekci 2014]. When biases cannot be ruled out as the biasing factors are too complex or hard to untangle, running longitudinal, comparative, multi-datasets, cross-domain/platform analyses is advised [Bruns 2013; Gayo-Avello et al. 2013; Ruths and Pfeffer 2014; Shani and Gunawardana 2011; Weller et al. 2015].

If access to multiple datasets is limited, the analysis can be run on datasets altered to introduce or remove noise or biases [Ruths and Pfeffer 2014]. Alternatively, general patterns can be probed across different classes of data [Bobadilla et al. 2013; Cohen and Ruths 2013], as important data variations may exist not only across datasets, but also within datasets, across data demographics e.g., distinct classes of users [Cohen and Ruths 2013] or of items [Olteanu et al. 2014b]. To ensure that observed patterns are not evanescent trends due to platform changes (e.g., as users interaction patterns may change due to functional changes) [De Myttenaere et al. 2014; Ruths and Pfeffer 2014; Weller et al. 2015], a study can be replicated across time [Lazer et al. 2014; Liang and Fu 2015].

8.3. Disclaimers and Reproducibility

Finally, there is also a need to develop baselines and guidelines [Tufekci 2014; Weller and Kinder-Kurlanda 2015], to find common grounds regarding methodological approaches [Counts et al. 2014], and to better document home-grown tools and methodologies, as well as dataset provenance [Bruns 2013; Weller and Kinder-Kurlanda 2015]. While text mining or information retrieval research typically follows standard evaluation procedures and metrics, for many social media analysis tasks—such as crisis informatics—more effort is needed for developing standardized experimental methodologies when they do not exist [Bruns 2013; Diaz 2014].

– Disclaimers and negative results are overlooked. While failed studies or negative results are useful for learning about what hypotheses were rejected, or what datasets or methods are not suitable for a given problem, publications of negative results are scant [Gayo-Avello 2012; Ruths and Pfeffer 2014]. There is an unfortunate bias against publication of negative results [Fanelli 2012] describing failures to reproduce an existing result, or that discuss approaches that did not deliver the expected results such as the set of features that did not improve the classification performance, algorithms that failed to deliver an acceptable performance, or what evidence did not support ones hypothesis.

In addition, disclaimers about the limitations of an analysis are fundamental to good practice. If errors or biases cannot be ruled out, researchers must discuss the gaps and limitations in their working datasets, their methods and their assumptions [Crawford and Finn 2014; Ruths and Pfeffer 2014; Tufekci 2014]. The risk of ambiguous generalizability claims should be considered, and the assumptions under which the results would hold to other context (e.g., other domains, platforms or populations) should be clarified.

– There is a need to ease the task of sharing tools and data, which are cornerstone for the reproducibility and replicability of studies.[18]

Data sharing may consist of providing datasets, or the details (including source code) for gathering exactly or approximately the same datasets when data sharing is prohibited by terms of service or privacy constrains. It can reduce redundant, labor-intensive, and time-consuming data collection, making social data research more inclusive and narrowing existing data access gaps [Jurgens et al. 2015a; Weller and Kinder-Kurlanda 2015]. Yet, Hutton and Henderson [2015b] study of 505 papers mentioning a social network between 2011-2013 revealed that only about 6% of them share any data, while Zimmer and Proferes [2014] found that of 382 Twitter studies only about 5% use existing datasets collected by other researchers. A reason for this is the confusion around data sharing

[18] See [Drummond 2009] for the difference between reproducibility and replicability.

"*cans*" and "*cannots*," which we discuss in the next section, such as: What can be shared? How to make it accessible?

Tools sharing may include providing details (including source code) for understanding or executing an algorithm or for analyzing data. Beyond aiding reproducibility and future comparisons, the availability of tools may also enable the participation of those lacking the resources to create their own (e.g., non-CS researchers) [boyd and Crawford 2012; Bruns and Liang 2012]. Alas, releasing and maintaining code and tools is a laborious, non-trivial task and many researchers lack the incentives to do so. However, there are increasing efforts to release and open-source data and tools such as Jurgens et al. [2015b]; Kıcıman et al. [2014]; McCreadie et al. [2012].

9. ETHICAL CONSIDERATIONS

Ethics are important in all kinds of research. Trust is eroded by breaches of scientific ethics such as fabricated results, conflicts of interest, and plagiarism. The public support to scientific endeavors can be severely undermined by research that harms people, animals, or the environment, or that goes against values that are deeply held by society. In particular, research on human subjects is regulated by law in many jurisdictions, and given that data elements in social datasets represent people or groups of people [Diaz 2016]; research on social data can be considered human subjects research.

The fact that social data is often publicly accessible does not mean research done on is ethical [boyd and Crawford 2012; Zimmer 2010]. Indeed, ethical issues in social media research have been highlighted by recent cases, including the notorious Facebook contagion experiment, where researchers manipulated users' social feeds to include more or less of certain kinds of content based on the expressed emotions [Kramer et al. 2014]. The experiment was criticized as an intervention that affected the emotional state of unsuspecting users, who had not given consent to participate in the study [Hutton and Henderson 2015a]. This incident was followed by an unprecedented move by the SIGCOMM 2015 Program Committee[19] which decided to accept a paper on measuring censorship [Burnett and Feamster 2015] on the condition of placing a prominent note at the top of the paper highlighting their ethical concerns [Narayanan and Zevenbergen 2015], drawing further attention to the issue.

Scientists [Barocas and Selbst 2014; Chou 2015; Dwork and Mulligan 2013; Kenneally 2015] and journalists [Hill 2014; Kirchner 2015; Miller 2015] have urged scientists and practitioners to carefully scrutinize their use of social data against a variety of possible ethical pitfalls, such as breaching users privacy [Goroff 2015], or enabling racial, socioeconomic or gender-based profiling [Barocas and Selbst 2014; Chou 2015]. However, only a small number of publications (between 2007 and 2012) relying on data from Twitter was found to include any discussion or even an acknowledgement of ethical challenges [Zimmer and Proferes 2014]. On the other hand, while surveying various works for this survey, we noticed recent papers that mentioned or addressed ethical issues in their research, such as Fourney et al. [2015]; Kumar et al. [2015]; Minkus et al. [2015].

We outline key concepts and principles of human subjects research ethics in the next section. We then organize the discussion on specific ethical problems in social data research w.r.t. the principles of autonomy (§9.2), beneficence (§9.3) and justice (§9.4). Given that our treatment of the subject is purposefully schematic, the interested reader can find more information in Bowser and Tsai [2015]; boyd and Crawford [2012]; Grimmelmann [2015]; Metcalf and Crawford [2016], among others.

9.1. Concepts and Principles

In 1947, in the aftermath of World War II, the Nuremberg Code[20] provided the foundation for the development of human subjects research ethics, manifested in the Declaration of Helsinki [World Medical Association 1964] and the Belmont report [Ryan et al. 1978]. The latter is based on three ethical principles,

— autonomy: experiments should show respect for individuals (§9.2);

[19] SIGCOMM is a top-tier conference on computer networking.

[20] Available at http://www.cirp.org/library/ethics/nuremberg/, accessed December 2016.

— beneficence and non-maleficence: experiments should minimize risk for research participants and maximize benefits for society (§9.3);
— justice: the risks and benefits of experiments should be fairly distributed (§9.4).

These principles are brought to practice into three primary areas of application: informed consent, assessment of risks and benefits, and selection of subjects—covered in following subsections.

Social media research is different from clinical trials. Many processes to ensure ethical compliance in human subject research were developed in the medical profession for the purposes of clinical trials, which involve testing the effect of a treatment on actual patients. These treatments may have harmful, sometimes severe and irreversible unexpected effects. In contrast, the harm that common types of current social media research can produce is often of a different nature, such as suffering a breach of privacy, or being exposed to disturbing images.

In this context, Bowser and Tsai [2015] advocate for a process that starts from the distinction between four types of research that are social media specific, characterized by (i) whether experimental subjects are aware that their data is being used for research; and (ii) whether experimental subjects are subject to any type of intervention or manipulation. The four types are:

(1) awareness and manipulation: e.g., a lab-based image annotation study;
(2) awareness without manipulation: e.g., an opt-in study of nutrition and exercise self-monitoring;
(3) no awareness with manipulation: e.g., the A/B testing of a feature on a social media platform;
(4) no awareness and no manipulation: most observational studies involving data acquired from a social media platform.

Each type involves a somewhat different process, see Bowser and Tsai [2015] for details.

Ethical choices in research require deliberation. Ethical choices are difficult, among other reasons, because they may involve different and sometimes conflicting values. Particularly, the value of producing research outcomes that are valuable for society may collide with values such as privacy for the people who contribute the data used for research. For instance, we may infer a statistical model whose accuracy can be increased with more details about people—at the expense of a loss of privacy from the data contributors. Data analysis may in fact be necessary to provide important services, and solutions that balance between privacy and accuracy should be considered [Goroff 2015]. For instance, the United Nations Office for the Coordination of Humanitarian Affairs seeks a balance between the imperative to save lives and the responsibility to do no harm: "Absolute protection would make humanitarian response impractical by not allowing the collection of any information, while the public listing of personal details would likewise endanger lives" [UN OCHA 2014].

People in the computing profession have varying degrees of preparation when it comes to addressing ethical problems, and may approach them from the wrong angle, e.g., from a too narrowly utilitarian, or a too strictly legalistic perspective. Instead, these issues are best addressed through informed deliberation and conversation. This is why Institutional Review Boards (IRBs) are important. Researchers at many universities are often required to submit their research proposals to IRBs for approval, but even in cases where this is not mandated by an institution, it constitutes a good practice. IRBs set common standards within an institution, provide researchers a framework to think critically about consequences, and show to others that careful decisions have been made for a study.

9.2. Respect to Individual Autonomy

The concept of individual autonomy, i.e., the capacity of individuals to make autonomous decisions, is usually expressed in human subject research as the need for **informed consent.** Participants should give researchers explicit permission, often in written, to use their data for research; furthermore, they should be free to withdraw this permission at any point. Informed consent requires researchers to *disclose* all relevant information for a decision on whether to participate or not in a research study, to a potential participant that is *capable* of evaluating this information, so that he or she can *voluntarily* decide without any coercion or pressure whether to participate or not. Research with online social data poses particular challenges to the practice of informed consent.

– Obtaining consent from millions of users is impractical. Studies that leverage user data from millions of social media users often do it without any kind of consent from them [Hutton and Henderson 2015a; Zimmer 2010]. User data may have been provided freely online for anyone to access it, but it is inherently sensitive as users might not have anticipated a particular use of their data, especially when created in a context-sensitive space and time [boyd and Crawford 2012]. This becomes even more delicate when analyzing user demographic attributes [Chou 2015]. While asking consent might be often seen as impractical [boyd and Crawford 2012], we should note that there are a few efforts to design methodologies for acquiring consent while minimizing the burden on the participants [Hutton and Henderson 2015a].

Even if we were to accept the notion that by placing their information in online public spaces, user consent for research is implied, "people privacy preferences depend on their circumstances" [Crawford and Finn 2014] (and these preferences may or may not be reflected in privacy settings, which users rarely change [Wang et al. 2011]). Take the case of social media use in crisis situations by vulnerable populations that may publicly share personal information to assist others or ask for help. Such disclosures are closely coupled with the context, and, thus, data use and share should be extensively scrutinized and the privacy of these users should be protected [Crawford and Finn 2014].

– The terms of use of a social platform may not constitute informed consent for research. By signing up for a social media platform, users accept their terms of use, which often contain clauses allowing research for testing new features or for other purposes. The acceptance of terms of use may not fulfill the criteria of informed consent, as the often vague language alluding to "research use" does not involve a disclosure of the relevant elements of a specific research program. For instance, the aftermath of the Facebook emotional contagion experiment [Kramer et al. 2014] suggests that users were in fact not sufficiently informed about the nature of this research, including its risks and benefits. Even if experiments were described clearly in a specific informed consent form for this type of experiment, the intimate nature of social platforms may require ongoing, dynamic consent, as is found in disciplines such as ethnography [American Anthropological Association 2004].

9.3. Beneficence and Non-Maleficence

Researchers in all disciplines should ensure that their actions are beneficial and do not cause harm. Research on social data is associated to specific types of harm, of which privacy breaches exposing the identity or the personal information of participants are perhaps the most obvious type [Crawford and Finn 2014; Zimmer 2010].

– Data about individuals can harm them if exposed. As our "offline" and "online" lives converge [Vieweg et al. 2015], and public and private spaces online appear as less separable [Crawford and Finn 2014], privacy breaches become more dangerous. Privacy breaches can have harmful consequences [Barocas and Selbst 2014] such as stalking, identity theft, discrimination or blackmailing [Gross and Acquisti 2005]. Social data can potentially reveal more about individuals than what they think. While internet users may decide to withhold certain information about themselves online (e.g., age, gender, sexual orientation, religious views), such information can be to a large extent inferred from other digital traces [Kosinski et al. 2013].

In general, privacy breaches are possible as publicly shared datasets can be combined to gain insights about private individuals without their knowledge [Crawford and Finn 2014; Goroff 2015; Gross and Acquisti 2005; Horvitz and Mulligan 2015]. To minimize risks, the sharing and archival of data embedding personal information [Zimmer and Proferes 2014], as well as the use of content explicitly deleted by users should be cautiously handled and anonymization should be considered [Almuhimedi et al. 2013; Crawford and Schultz 2014; Maddock et al. 2015]. This anonymization does not only mean the removal of personally identifiable information, but careful processing of data to avoid re-identification through combinations of apparently non-sensitive attributes [Ohm 2010].

– Research outcomes can be used to do harm in unforeseen ways. Inferences made for a purpose can be used for another. In the early 1930s, Germany's secret police maintained "pink lists" of sus-

pected homosexuals, which were later used by the Nazis to arrest them.[21] A more recent example was the arrest of protesters in Baltimore, USA, based on information provided by social media.[22] In addition to the fact that inferences drawn from social data may be incorrect in many ways, as this survey emphasizes, inferences that are too precise may be harmful as well, for instance, by enhancing the capacity to finely discriminate among people into ever-smaller groups [Barocas 2014].

9.4. Justice

An ideal of justice in research is that risks and benefits should be justly apportioned, and that research does not contribute to create injustice. Independently of the particular conception of justice being used, these questions require at the onset to know the individuals or groups that will be burdened by research, and the ones that will benefit from the results.

– The digital divide may influence research design. The *digital divide* is the gap that exists among and within countries with respect to access to information and communication technologies. This gap has many manifestations, including the *data divide*: a lack of availability of high-quality data about developing countries and underprivileged communities [Cinnamon and Schuurman 2013]. Together, the digital divide and the data divide can be important source of bias on the questions that are asked and the populations that are chosen for research [boyd and Crawford 2012; Counts et al. 2014]. They can focus the research agenda on so-called "first-world problems," such as finding a restaurant, for which data is widely available.

– Research outcomes can be made more or less available to people. Providing information to people about how their data are used is an important element concerning their autonomy [Horvitz and Mulligan 2015]. This transparency can also lead to a more just allocation of research benefits. Ideally, people should be given access to research results and artifacts that resulted from the study of their personal data [Crawford and Finn 2014; Gross and Acquisti 2005]. Research outcomes can be beneficial if they are shared widely, ideally via open access to research results and data, to the extent that it does not compromise the privacy of users. Alas, more often than not, the way in which user data is processed and analyzed to support decision making remains "*black-boxed*" [Poirier 2015].

Furthermore, a failure to make data available may further deepen the data divide [Bruns 2013] as well as the gap between those that have the computational skills needed to analyze large volumes of data and those who lack them [boyd and Crawford 2012; Weller and Kinder-Kurlanda 2015]. This can be further exacerbated by the raise of "embedded" researchers that have privileged access to social platforms and are able to access data that is unavailable to broader groups [Crawford and Finn 2014; Ruths and Pfeffer 2014].

– Algorithms and research outcomes can lead to discrimination. The use of social data can result in discrimination against protected groups [Barocas and Selbst 2014]. The reliance on automated decision making processes based on statistical methods, can inherit, propagate, or even amplify the biases and prejudice present in the training data with respect to various factors such as race, age, gender or socioeconomic groups [Barocas and Selbst 2014; Crawford and Schultz 2014; Kirchner 2015; Miller 2015]. This problem is often referred to as *algorithmic discrimination* [Kirchner 2015]. While such a result is often unintentional, it can have a variety of consequences: companies could use such information to practice price steering and discrimination [Hannak et al. 2014]; or users employment, credit or housing prospectives may be affected due to being stereotyped and profiled based on their race [Barocas and Selbst 2014]. This is concerning as the existing laws often cannot properly handle such issues [Barocas and Selbst 2014; Crawford and Schultz 2014].

---

[21] "Persecution of Homosexuals in the Third Reich," US Holocaust Memorial Museum, https://www.ushmm.org/wlc/en/article.php?ModuleId=10005261. Accessed December 2016.

[22] "Facebook, Twitter, and Instagram surveillance tool was used to arrest Baltimore protesters." The Verge, October 2016. http://www.theverge.com/2016/10/11/13243890/facebook-twitter-instagram-police-surveillance-geofeedia-api. Accessed December 2016.

10. DISCUSSION: TRENDS AND FUTURE DIRECTIONS

The last few years, particularly the last two, have seen a growing interest among researchers and practitioners on probing known limitations of social datasets and social data methods; ethical challenges have also been brought to the forefront. In the light of this trend, we believe that the need to identify, quantify, and address data biases, as well as methodological and ethical challenges around the use of social data, will remain a persistent and important issue for years to come.

However, eliminating all biases on social data is *unlikely*, perhaps even *undesirable*. Biases that bound the general applicability of solutions may help boost the performance of dedicated solutions [Guerra et al. 2014; Olteanu et al. 2014a; Yan et al. 2011] or inform the design of specific systems [Lerman and Hogg 2014; Olteanu and Pierre 2012]. In other cases, the solutions to various limitations might pull in opposite directions—e.g., in a user classification problem, one may be faced with the decision to err against a minority group to ensure accurate results for the majority, or vice versa. Even in cases where trade-offs are fully understood, which are not common, it may be unclear what would be the best way to balance different aspects. Ultimately, as we stressed earlier in this survey (§1–§2), whether a research method or a dataset is adequate (or not) depends on the research question being asked, the context in which the research takes place, and, fundamentally, on the goals of the researcher(s).

We foresee two broad trends for future research and discussions around the limits of social data. We believe that the skepticism around easy answers and out-of-the-box solutions will continue to grow (§10.1), and we see increasing efforts towards addressing these questions and developing data standards and methodological best practices (§10.2). We conclude this survey with pointers to further readings on these topics (§10.3).

10.1. A Trending Skepticism Towards Easy Answers

Following the well-known "hype cycle,"[23] the phase of "inflated expectations" on social data research has perhaps already passed. A growing number of research fora that critically examine research have emerged on disciplines that either focus on social data, or often use it. This includes the Fair and Transparency in Machine Learning Workshop (FATML)[24], the Fair and Transparency on Web Workshop (FAT/WEB)[25], and the Ethics in Natural Language Processing Workshop[26], in addition to focalized one-off discussions such as the special issue on "Social and Technical Trade-Offs" in the Journal of Big Data [Barocas et al. 2016]. These venues often adopt an ethical framework based on fairness and transparency, to motivate discussions around the consequences of built-in biases in working datasets and methodologies and about unethical uses of social data. Policies around these concerns may eventually emerge, as activity on the policy dimensions of these biases is increasing [Crawford et al. 2016; Goodman and Flaxman 2016; US White House 2016].

These efforts are embedded in a context of a broad reflection of common needs of computing research across the board, such as "the need for increasing awareness for what it is actually analyzed" (e.g., data and phenomena) [Ruths and Pfeffer 2014], or the need to understand various dimensions of the automated behavior of platform specific mechanisms (e.g., design and algorithms) [Sandvig et al. 2014]. In this context, the use of social data for commercial and research purposes remains a core area of concern [boyd and Crawford 2012; Salganik 2017; Sandvig et al. 2014].

10.2. A Shift From Raising to Addressing Concerns About Social Data

There are a current efforts going beyond the identification of potential issues around the use of social data, and into frameworks to audit and mitigate those issues.

There is a growing interest in auditing existing social software systems. In some cases bias can be hard to discover without a thorough, in-depth examination of a dataset or system. Sandvig

[23] Gartner Hype Cycle, http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp. Accessed Dec. 2016.
[24] http://www.fatml.org/
[25] https://fatweb.github.io/
[26] http://ethicsinnlp.com/

et al. [2014] argue that scrutiny is required even when a social software system appears to satisfy users' needs, as there can be "subtle patterns of problematic behavior" that may be hard to discern. Kulshrestha et al. [2017] introduces a framework for auditing search systems on social media platforms by differentiating between various sources of bias, such as bias due to the content created by users or due to ranking algorithms. This audit sometimes requires access to proprietary systems, which needs explicit permission, which is likely to be denied if the goal is to expose flaws. Reverse engineering systems or using them in unanticipated way to expose their bias may be illegal under the US Computer Fraud and Abuse Act (CFAA), which has been challenged in court by a group of researchers.[27]

The space of solutions addressing social data limits is growing. While we have briefly noted efforts to improve data collection (§5.2), the benefits of combining qualitative and quantitative methods (§7.1), or using open-source data and tools (§8.3), our survey has been mostly concerned with highlighting potential issues around the use of social data. There is a growing number of works that try to offer solutions to these issues, in the form of guidelines, standards, or new methodological approaches. One important direction, for instance, is towards employing techniques from the causal inference literature that can lead to more robust research results such as Landeiro and Culotta [2016]; Olteanu et al. [2017]; Proserpio et al. [2016]. Another direction is to employ standardized evaluation protocols when testing new tools or methodologies [Diaz 2014; Jurgens et al. 2015b].

10.3. Further Reading

For more discussion on the issues we cover in this survey, we recommend the books by Salganik [2017] and O'Neil [2016], the tutorials by Castillo et al. [2016] and Weber et al. [2016], the talks by Wallach [2014] and Diaz [2016], and the following papers: boyd and Crawford [2012]; Nguyen et al. [2016]; Ruths and Pfeffer [2014]; Tufekci [2014]; and Amodei et al. [2016], among others.

REFERENCES

Sofiane Abbar, Yelena Mejova, and Ingmar Weber. 2015. You Tweet What You Eat: Studying Food Consumption Through Twitter. In *Proc. of CHI*.

Norah Abokhodair, Daisy Yoo, and David W McDonald. 2015. Dissecting a social botnet: Growth, content and influence in Twitter. In *Proc. of CSCW*.

Kevin MacG Adams and Patrick T Hester. 2012. Errors in systems approaches. *International Journal of System of Systems Engineering* 3, 3-4 (2012).

Hazim Almuhimedi, Shomir Wilson, Bin Liu, Norman Sadeh, and Alessandro Acquisti. 2013. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proc. of CSCW*.

American Anthropological Association. 2004. Statement on Ethnography and Institutional Review Boards. (2004). http://www.americananthro.org/ParticipateAndAdvocate/Content.aspx?ItemNumber=1652

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *arXiv preprint arXiv:1606.06565* (2016).

Chris Anderson. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired* 16, 07 (2008).

Monica Anderson. 2015. Men catch up with women on overall social media use. *Pew Research Center* (2015).

Sinan Aral and Dylan Walker. 2011. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management science* 57, 9 (2011).

Anne Archambault and Jonathan Grudin. 2012. A longitudinal study of Facebook, Linkedin, & Twitter use. In *Proc. of CHI*.

Sitaram Asur, Bernardo Huberman, and others. 2010. Predicting the future with social media. In *Proc. of WI-IAT*.

Anne Aula, Rehan M Khan, and Zhiwei Guan. 2010. How does search behavior change as search becomes more difficult?. In *Proc. of CHI*.

[27] Sandvig v. Lynch: Challenge to CFAA Prohibition on Uncovering Racial Discrimination Online. American Civil Liberties Union (ACLU), June 29th 2016. https://www.aclu.org/cases/sandvig-v-lynch-challenge-cfaa-prohibition-uncovering-racial-discrimination-online, Accessed Dec. 2016.

Mitja D Back, Albrecht CP Küfner, and Boris Egloff. 2010. The emotional timeline of September 11, 2001. *Psychological Science* (2010).
Lars Backstrom, Eytan Bakshy, Jon M Kleinberg, Thomas M Lento, and Itamar Rosenn. 2011. Center of Attention: How Facebook Users Allocate Attention across Friends. In *Proc. of ICWSM*.
Ricardo Baeza-Yates. 2014. Wisdom of Crowds or Wisdom of a Few? *Web Engineering* (2014).
Ricardo Baeza-Yates and Yoelle Maarek. 2012. Usage data in web search: benefits and limitations. In *Scientific and Statistical Database Management*. Springer.
Ricardo A Baeza-Yates. 2013. Big Data or Right Data?. In *AMW*.
Mossaab Bagdouri and Douglas W Oard. 2015. On Predicting Deletions of Microblog Posts. In *Proc. of CIKM*.
Eytan Bakshy, Dean Eckles, Rong Yan, and Itamar Rosenn. 2012. Social Influence in Social Advertising: Evidence from Field Experiments. In *Proc. of EC*.
Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015).
Solon Barocas. 2014. Data Mining and the Discourse on Discrimination. In *Proc. of KDD Workshop on Data Ethics*.
Solon Barocas, danah boyd, Sorelle Friedler, and Hanna Wallach. 2016. Special Issue on Social and Technical Trade-Offs. *Big Data* 4, 1 (2016).
Solon Barocas and Andrew D Selbst. 2014. Big Data's Disparate Impact. *Soc. Sci. Research Network Working Paper Series* (2014).
Asaf Beasley and Winter Mason. 2015. Emotional states vs. emotional words in social media. In *Proc. of WebSci*.
Paul Bennett, Alexander Fishkov, and Emre Kıcıman. 2015. Persona-ization: Searching on behalf of others. In *Proc. of SIGIR Workshop on Social Personalization and Search*.
Michael S Bernstein, Andrés Monroy-Hernández, Drew Harry, Paul André, Katrina Panovich, and Gregory G Vargas. 2011. 4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community. In *Proc. of ICWSM*.
Parantapa Bhattacharya, Saptarshi Ghosh, Juhi Kulshrestha, Mainack Mondal, Muhammad Bilal Zafar, Niloy Ganguly, and Krishna P Gummadi. 2014. Deep twitter diving: Exploring topical groups in microblogs at scale. In *Proc. of CSCW*.
Colin R Blyth. 1972. On Simpson's paradox and the sure-thing principle. *J. Amer. Statist. Assoc.* 67, 338 (1972).
Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez. 2013. Recommender systems survey. *Knowledge-Based Systems* 46 (2013).
Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Quantifying and Reducing Stereotypes in Word Embeddings. *CoRR* abs/1606.06121 (2016).
Anne Bowser and Janice Y Tsai. 2015. Supporting Ethical Web Research: A New Research Ethics Review. In *Proc. of WWW*.
danah boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society* 15, 5 (2012).
danah boyd and Nicole B Ellison. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication* 13, 1 (2007).
danah boyd, Scott Golder, and Gilad Lotan. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proc. of HICSS*.
Axel Bruns. 2013. Faster than the speed of print: Reconciling 'big data' social media analysis and academic scholarship. *First Monday* 18, 10 (2013).
Axel Bruns and Yuxian Eugene Liang. 2012. Tools and methods for capturing Twitter data during natural disasters. *First Monday* 17, 4 (2012).
Axel Bruns and Stefan Stieglitz. 2014. Twitter data: what do they represent? *Information Technology* 56, 5 (2014).
Moira Burke, Lada Adamic, and Karyn Marciniak. 2013. Families on Facebook. In *Proc. of ICWSM*.
Moira Burke and Robert E Kraut. 2014. Growing closer on facebook: changes in tie strength through social network site use. In *Proc. of CHI*.
Sam Burnett and Nick Feamster. 2015. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. In *Proc. of SIGCOMM*.
Aylin Caliskan-Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Language necessarily contains human biases, and so will machines trained on language corpora. (2016). in preparation.
R Sherlock Campbell and James W Pennebaker. 2003. The secret life of pronouns flexibility in writing style

and physical health. *Psychological science* 14, 1 (2003).
Eyal Carmi, Gal Oestreicher-Singer, and Arun Sundararajan. 2012. Is Oprah contagious? Identifying demand spillovers in online networks. *Identifying Demand Spillovers in Online Networks, .NET Institute Working Paper* (2012).
Carlos Castillo, Fernando Diaz, Emre Kıcıman, and Alexandra Olteanu. 2016. A Critical Review of Online Social Data: Limitations, Ethical Challenges, and Current Solutions. ICWSM Tutorials. (2016). http://www.aolteanu.com/SocialDataLimitsTutorial/
Ilknur Celik, Fabian Abel, and Geert-Jan Houben. 2011. Learning semantic relationships between entities in twitter. In *Proc. of WISE*.
Daphne Chang, Erin L Krupka, Eytan Adar, and Alessandro Acquisti. 2016. Engineering Information Disclosure: Norm Shaping Designs. In *Proc. of CHI*.
Kathy Charmaz. 2014. *Constructing grounded theory*. Sage.
Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein, and Ed Chi. 2010. Short and Tweet: Experiments on Recommending Content from Information Streams. In *Proc. of CHI*.
Justin Cheng and Dan Cosley. 2013. How annotation styles influence content and preferences. In *Proc. of Hypertext*.
Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2014. How Community Feedback Shapes User Behavior. In *Proc. of ICWSM*.
Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. Antisocial Behavior in Online Discussion Communities. In *Proc. of ICWSM*.
Sophie Chou. 2015. Race and the Machine: Re-examining Race and Ethnicity in Data Mining. http://www.sophiechou.com/papers/chou_racepaper.pdf. (2015).
Nicholas A Christakis and James H Fowler. 2007. The spread of obesity in a large social network over 32 years. *New England journal of medicine* 357, 4 (2007).
Jonathan Cinnamon and Nadine Schuurman. 2013. Confronting the data-divide in a time of spatial turns and volunteered geographic information. *GeoJournal* 78, 4 (2013).
Raviv Cohen and Derek Ruths. 2013. Classifying Political Orientation on Twitter: Its Not Easy!. In *Proc. of ICWSM*.
Ethan Cohen-Cole and Jason M Fletcher. 2008. Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis. *Bmj* 337 (2008).
Scott Counts, Munmun De Choudhury, Jana Diesner, Eric Gilbert, Marta Gonzalez, Brian Keegan, Mor Naaman, and Hanna Wallach. 2014. Computational social science: CSCW in the social media era. In *Proc. of CSCW Companion*.
Justin Cranshaw, Raz Schwartz, Jason I Hong, and Norman Sadeh. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proc. of ICWSM*.
Kate Crawford. 2013. The hidden biases in big data. *HBR Blog Network* 1 (2013).
Kate Crawford and Megan Finn. 2014. The limits of crisis data: Analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal* (2014).
Kate Crawford and Jason Schultz. 2014. Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College. Law School. Boston College Law Review* 55, 1 (2014).
Kate Crawford, Meredith Whittaker, Madeleine Clare Elish, Solon Barocas, Aaron Plasek, and Kadija Ferryman. 2016. *The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*. Technical Report. AI Now. artificialintelligencenow.com
J.W. Creswell and V.L.P. Clark. 2011. *Designing and Conducting Mixed Methods Research*. SAGE Publications. https://books.google.com/books?id=YcdlPWPJRBcC
Dave D'Alessio and Mike Allen. 2000. Media bias in presidential elections: A meta-analysis. *Journal of communication* 50, 4 (2000).
Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proc. of WWW*.
Sauvik Das and Adam Kramer. 2013. Self-Censorship on Facebook. In *Proc. of ICWSM*.
Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Major life changes and behavioral markers in social media: case of childbirth. In *Proc. of CSCW*.
Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proc. of CHI*.
Munmun De Choudhury, Yu Ru Lin, Hari Sundaram, Kasim Candan, Lexing Xie, and Aisling Kelliher. 2010. How does the data sampling strategy impact the discovery of information diffusion in social media?. In *Proc. of ICWSM*.

Munmun De Choudhury, Meredith Ringel Morris, and Ryen W. White. 2014. Seeking and Sharing Health Information Online: Comparing Search Engines and Social Media. In *Proc. of CHI*.

Arnaud De Myttenaere, Bénédicte Le Grand, Boris Golden, and Fabrice Rossi. 2014. Reducing Offline Evaluation Bias in Recommendation Systems. *arXiv preprint arXiv:1407.0822* (2014).

Matthew James Denny and Arthur Spirling. 2016. Assessing the Consequences of Text Preprocessing Decisions. *Available at SSRN 2849145* (2016).

Nicholas Diakopoulos. 2016. Accountability in algorithmic decision making. *Commun. ACM* 59, 2 (2016).

Fernando Diaz. 2014. Experimentation Standards for Crisis Informatics. *SIGIR Forum* 48, 2 (2014).

Fernando Diaz. 2016. Worst Practices for Designing Production Information Access Systems. In *ACM SIGIR Forum*.

Fernando Diaz, Michael Gamon, Jake Hofman, Emre Kıcıman, and David Rothschild. 2016. Online and Social Media Data As an Imperfect Continuous Panel Survey. *PlosONE* 11, 1 (2016).

Joan DiMicco, David R Millen, Werner Geyer, Casey Dugan, Beth Brownholtz, and Michael Muller. 2008. Motivations for social networking at work. In *Proc. of CSCW*.

Yuxiao Dong, Omar Lizardo, and Nitesh V Chawla. 2016. Do the Young Live in a "Smaller World" Than the Old? Age-Specific Degrees of Separation in a Large-Scale Mobile Communication Network. *arXiv preprint arXiv:1606.07556* (2016).

Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health from Twitter. In *Proc. of AAAI*.

Mark Dredze, Miles Osborne, and Prabhanjan Kambadur. 2016. Geolocation for Twitter: Timing Matters. In *Proc of. NAACL*.

Chris Drummond. 2009. Replicability is not reproducibility: nor is it good science. In *Proc. of Workshop on Evaluation Methods for Machine Learning*.

Maeve Duggan. 2015. The Demographics of Social Media Users. *Pew Research Center* (2015).

Maeve Duggan, Nicole B Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. 2015. Demographics of key social networking platforms. *Pew Research Center* (2015).

Susan Dumais, Robin Jeffries, Daniel M Russell, Diane Tang, and Jaime Teevan. 2014. Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI*. Springer.

Cynthia Dwork and Deirdre K Mulligan. 2013. It's not privacy, and it's not fair. *Stanford Law Review Online* 66 (2013).

Kate Ehrlich and N Sadat Shami. 2010. Microblogging Inside and Outside the Workplace. In *Proc. of ICWSM*.

Hamid Ekbia, Michael Mattioli, Inna Kouper, G Arave, Ali Ghazinejad, Timothy Bowman, Venkata Ratandeep Suri, Andrew Tsou, Scott Weingart, and Cassidy R Sugimoto. 2015. Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology* 66, 8 (2015).

Daniele Fanelli. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* 90, 3 (2012).

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2014. The rise of social bots. *arXiv preprint arXiv:1407.5225* (2014).

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preotiuc-Pietro. 2016. Analyzing Biases in Human Perception of User Age and Gender from Text. In *Proc. of ACL*.

Adam Fourney, Ryen W White, and Eric Horvitz. 2015. Exploring time-dependent concerns about pregnancy and childbirth from search logs. In *Proc. of CHI*.

Julia D. Fraustino, Brooke Liu, and Yan Jin. 2012. *Social Media Use during Disasters: A Review of the Knowledge Base and Gaps*. Technical Report. Science and Technology Directorate, U.S. Department of Homeland Security. http://www.start.umd.edu/sites/default/files/files/publications/START_SocialMediaUseduringDisasters_LitReview.pdf

Deen Freelon. 2014. On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media* 58, 1 (2014).

Wei Gao and Fabrizio Sebastiani. 2016. From classification to quantification in tweet sentiment analysis. *Social Network Analysis and Mining* 6, 1 (2016).

Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. 2013. Cultural dimensions in twitter: Time, individualism and power. In *Proc. of ICWSM*.

Venkata Rama Kiran Garimella, Ingmar Weber, and Sonya Dal Cin. 2014. From "I love you babe" to "leave me alone"-Romantic Relationship Breakups on Twitter. In *Social Informatics*. Springer.

Daniel Gayo-Avello. 2012. Ï Wanted to Predict Elections with Twitter and all I got was this Lousy Paper¨-A Balanced Survey on Election Prediction using Twitter Data. *arXiv preprint arXiv:1204.6441* (2012).

Daniel Gayo Avello, Panagiotis T Metaxas, and Eni Mustafaraj. 2011. Limits of electoral predictions using

Twitter. In *Proc. of ICWSM*.
Eni Mustafaraj Markus Strohmaier Harald Schoen Gayo-Avello, Panagiotis Takis Metaxas, Daniel Peter Gloor, Harald Schoen, Daniel Gayo-Avello, Panagiotis Takis Metaxas, Eni Mustafaraj, Markus Strohmaier, and Peter Gloor. 2013. The power of prediction with social media. *Internet Research* 23, 5 (2013).
Paolo Giardullo. 2015. Does 'bigger' mean 'better'? Pitfalls and shortcuts associated with big data for social research. *Quality & Quantity* (2015).
Eric Gilbert and Karrie Karahalios. 2010. Widespread Worry and the Stock Market. In *Proc. of ICWSM*.
Jim Giles. 2012. Making the links. *Nature* 488 (2012).
Tarleton Gillespie. 2015. Platforms intervene. *Social Media+ Society* 1, 1 (2015).
Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009).
Sharad Goel, Jake M Hofman, and M Irmak Sirer. 2012. Who Does What on the Web: A Large-Scale Study of Browsing Behavior. In *Proc. of ICWSM*.
Scott A Golder and Michael W Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333, 6051 (2011).
Wei Gong, Ee-Peng Lim, and Feida Zhu. 2015. Characterizing Silent Users in Social Media Communities. In *Proc. of ICWSM*.
Wei Gong, Ee-Peng Lim, Feida Zhu, and Pei Hua Cher. 2016. On Unravelling Opinions of Issue Specific-Silent Users in Social Media. In *Proc. of ICWSM*.
Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014a. Assessing the bias in communication networks sampled from Twitter. *Social Networks* 38 (2014).
Sandra González-Bailón, Ning Wang, Alejandro Rivero, Javier Borge-Holthoefer, and Yamir Moreno. 2014b. Assessing the bias in samples of large online networks. *Social Networks* 38 (2014).
Bryce Goodman and Seth Flaxman. 2016. EU regulations on algorithmic decision-making and a right to explanation. In *ICML Workshop on Human Interpretability in Machine Learning*.
Daniel L Goroff. 2015. Balancing privacy versus accuracy in research protocols. *Science* 347 (2015).
Mark Graham, Scott A Hale, and Devin Gaffney. 2014. Where in the world are you? Geolocation and language identification in Twitter. *The Professional Geographer* 66, 4 (2014).
Valentina Grasso and Alfonso Crisci. 2016. Codified Hashtags for Weather Warning on Twitter: an Italian Case Study. *PloS Currents Disasters* 1 (2016).
James Grimmelmann. 2015. *The Law and Ethics of Experiments on Social Media Users*. Technical Report 15. University of Maryland.
Nir Grinberg, Mor Naaman, Blake Shaw, and Gilad Lotan. 2013. Extracting Diurnal Patterns of Real World Activity from Social Media. In *Proc. of ICWSM*.
Ralph Gross and Alessandro Acquisti. 2005. Information revelation and privacy in online social networks. In *Proc. of Workshop on Privacy in the Electronic Society*.
Tom Gruber. 2008. Collective knowledge systems: Where the social web meets the semantic web. *Web semantics: science, services and agents on the World Wide Web* 6, 1 (2008).
Pedro Calais Guerra, Wagner Meira Jr, and Claire Cardie. 2014. Sentiment analysis on evolving social streams: How self-report imbalances can help. In *Proc. of WSDM*.
Zoltan Gyongyi and Hector Garcia-Molina. 2005. Web spam taxonomy. In *Proc. of AIRWeb*.
Aniko Hannak, Piotr Sapiezynski, Arash Molavi Kakhki, Balachander Krishnamurthy, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring personalization of web search. In *Proc. of WWW*.
Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring price discrimination and steering on e-commerce web sites. In *Proc. of IMC*.
Moritz Hardt. 2014. How big data is unfair: Understanding sources of unfairness in data driven decision making. *Medium. Available at: https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de* (2014).
Tim Harford. 2014. Big data: A big mistake? *Significance* 11, 5 (2014).
Eszter Hargittai. 2007. Whose space? Differences among users and non-users of social network sites. *Journal of Computer-Mediated Communication* 13, 1 (2007).
Eszter Hargittai, Lindsay Fullerton, Ericka Menchen-Trevino, and Kristin Yates Thomas. 2010. Trust online: Young adults' evaluation of web content. *International journal of communication* 4 (2010).
Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H Chi. 2011. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. In *Proc. of CHI*.
Kashmir Hill. 2014. Facebook Added 'Research' To User Agreement 4 Months After Emotion Manipulation Study. *Tech* (2014).
Lichan Hong, Gregorio Convertino, and Ed H Chi. 2011. Language Matters In Twitter: A Large Scale Study.

In *Proc. of ICWSM*.
Eric Horvitz and Deirdre Mulligan. 2015. Data, privacy, and the greater good. *Science* 349 (2015).
Dirk Hovy, Barbara Plank, and Anders Søgaard. 2014. When POS datasets don't add up: Combatting sample bias. In *Proc. of LREC*.
James Howison, Andrea Wiggins, and Kevin Crowston. 2011. Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems* 12, 12 (2011).
David John Hughes, Moss Rowe, Mark Batey, and Andrew Lee. 2012. A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior* 28, 2 (2012).
Luke Hutton and Tristan Henderson. 2015a. " I didn't sign up for this!": Informed consent in social network research. In *Proc. of ICWSM*.
Luke Hutton and Tristan Henderson. 2015b. Towards reproducibility in online social network research. *IEEE Transactions on Emerging Topics in Computing* (2015).
Matthew O Jackson. 2016. The Friendship Paradox and Systematic Biases in Perceptions and Social Norms. *Available at SSRN* (2016).
Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. Why we twitter: understanding microblogging usage and communities. In *Proc. of WebKDD*.
Isaac Johnson and Brent Hecht. 2015. Structural Causes of Bias in Crowd-derived Geographic Information: Towards a Holistic Understanding. (2015).
Isaac L Johnson, Subhasree Sengupta, Johannes Schöning, and Brent Hecht. 2016. The Geography and Importance of Localness in Geotagged Social Media. In *Proc. of CHI*.
Adam N Joinson. 2008. Looking at, looking up or keeping up with people?: motives and use of facebook. In *Proc. of CHI*.
Kenneth Joseph, Peter M Landwehr, and Kathleen M Carley. 2014. Two 1% s don't make a whole: Comparing simultaneous samples from Twitter's streaming API. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer.
David Jurgens, Tyler Finethy, Caitrin Armstrong, and Derek Ruths. 2015a. Everyone's Invited: A New Paradigm for Evaluation on Non-Transferable Datasets. In *Proc. of ICWSM*.
David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015b. Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. In *Proc. of ICWSM*.
Erin Kenneally. 2015. How to throw the race to the bottom: revisiting signals for ethical and legal research using online data. *ACM SIGCAS Computers and Society* 45, 1 (2015).
Norbert L Kerr. 1998. HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review* 2, 3 (1998).
Emre Kıcıman. 2010. Language differences and metadata features on Twitter. In *Web N-gram Workshop*.
Emre Kıcıman. 2012. OMG, I have to tweet that! A study of factors that influence tweet rates. In *Proc. of ICWSM*.
Emre Kıcıman, Scott Counts, Michael Gamon, Munmun De Choudhury, and Bo Thiesson. 2014. Discussion Graphs: Putting Social Media Analysis in Context. In *Proc. of ICWSM*.
Emre Kıcıman and Matthew Richardson. 2015. Towards Decision Support and Goal Achievement: Identifying Action-Outcome Relationships From Social Media. In *Proc. of KDD*.
Gary King. 2011. Ensuring the Data-Rich Future of the Social Sciences. *Science* 331, 6018 (2011).
Gary King, Jennifer Pan, and Margaret E Roberts. 2013. How censorship in China allows government criticism but silences collective expression. *American Political Science Review* 107, 02 (2013).
Lauren Kirchner. 2015. When discrimination is baked into algorithms. *The Atlantic* (2015).
Joseph Konstan, John Riedl, and others. 2012. Recommended for you. *IEEE Spectrum* 49, 10 (2012).
Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proc. of the National Academy of Sciences* 110, 15 (2013).
Adam DI Kramer. 2010. An unobtrusive behavioral model of gross national happiness. In *Proc. of CHI*.
Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *National Academy of Sciences* 111, 24 (2014).
Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, IIEST Shibpur, India Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying Search Bias: Investigating Sources of Bias for Political Searches in Social Media. In *Proc. of CSCW*.
Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides. In *Proc. of Hypertext*.
Sang Jib Kwon, Eunil Park, and Ki Joon Kim. 2014. What drives successful social networking services? A comparative analysis of user acceptance of Facebook and Twitter. *The Social Science Journal* 51, 4 (2014).

Cliff Lampe, Nicole Ellison, and Charles Steinfield. 2006. A Face (book) in the crowd: Social searching vs. social browsing. In *Proc. of CSCW*.

Cliff Lampe, Nicole B Ellison, and Charles Steinfield. 2008. Changes in use and perception of Facebook. In *Proc. of CSCW*.

Virgile Landeiro and Aron Culotta. 2016. Robust Text Classification in the Presence of Confounding Bias. In *Proc. of AAAI*.

David Lazer. 2015. Issues of Construct Validity and Reliability in Massive, Passive Data Collections. (2015). http://citiespapers.ssrc.org/issues-of-construct-validity-and-reliability-in-massive-passive-data-collections/

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343 (2014).

David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, and others. 2009. Life in the network: the coming age of computational social science. *Science* 323, 5915 (2009).

Kalev Leetaru, Shaowen Wang, Guofeng Cao, Anand Padmanabhan, and Eric Shook. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* 18, 5 (2013).

Kristina Lerman and Rumi Ghosh. 2010. Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. In *Proc. of ICWSM*.

Kristina Lerman and Tad Hogg. 2014. Leveraging Position Bias to Improve Peer Recommendation. *PLoS ONE* 9, 6 (2014).

Kristina Lerman, Xiaoran Yan, and Xin-Zeng Wu. 2016. The "majority illusion" in social networks. *PloS one* 11, 2 (2016).

Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proc. of KDD*.

Chenliang Li and Aixin Sun. 2014. Fine-grained location extraction from tweets with temporal awareness. In *Proc. of SIGIR*.

Linna Li, Michael F Goodchild, and Bo Xu. 2013. Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr. *Cartography and Geographic Information Science* 40, 2 (2013).

Hai Liang and King-wa Fu. 2015. Testing Propositions Derived from Twitter Studies: Generalization and Replication in Computational Social Science. *PloS one* 10, 8 (2015).

Q Vera Liao, Wai-Tat Fu, and Markus Strohmaier. 2016. # Snowden: Understanding Biases Introduced by Behavioral Differences of Opinion Groups on Social Media. In *Proc. of CHI*.

Bang Hui Lim, Dongyuan Lu, Tao Chen, and Min-Yen Kan. 2015. # mytweet via Instagram: Exploring User Behaviour across Multiple Social Networks. In *Proc. of ASONAM*.

Yu-ru Lin, James P Bagrow, and David Lazer. 2011. More voices than ever? quantifying media bias in networks. In *Proc. of ICWSM*.

Janne Lindqvist, Justin Cranshaw, Jason Wiese, Jason Hong, and John Zimmerman. 2011. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Proc. of CHI*.

Yabing Liu, Chloe Kliman-Silver, and Alan Mislove. 2014. The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior. In *Proc. of ICWSM*.

Russell Lyons. 2011. The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy* 2, 1 (2011).

Jim Maddock, Kate Starbird, and Robert Mason. 2015. Using Historical Twitter Data for Research: Ethical Challenges of Tweet Deletions. In *Proc. of CSCW Workshop on Ethics*.

Walid Magdy and Tamer Elsayed. 2014. Adaptive Method for Following Dynamic Topics on Twitter. In *Proc. of ICWSM*.

Momin M Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population Bias in Geotagged Tweets. In *ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*.

Momin M Malik and Jürgen Pfeffer. 2016. Identifying Platform Effects in Social Media Data. In *Proc. of ICWSM*.

Adam Mann. 2016. Core Concepts: Computational social science. *Proc. of the National Academy of Sciences* 113, 3 (2016).

Alice E Marwick. 2014. Ethnographic and qualitative research on Twitter. In *Twitter and society*. Peter Lang.

Alice E Marwick and others. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011).

Winter Mason, Jennifer Wortman Vaughan, and Hanna Wallach. 2014. Computational social science and social

computing. *Machine Learning* 95, 3 (2014).
James McCorriston, David Jurgens, and Derek Ruths. 2015. Organizations Are Users Too: Characterizing and Detecting the Presence of Organizations on Twitter. In *Proc. of ICWSM*.
Richard McCreadie, Ian Soboroff, Jimmy Lin, Craig Macdonald, Iadh Ounis, and Dean McCullough. 2012. On building a reusable Twitter corpus. In *Proc. of SIGIR*.
Caitlin McLaughlin and Jessica Vitak. 2012. Norm evolution and violation on Facebook. *New Media & Society* 14, 2 (2012).
Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* (2001).
Rishabh Mehrotra, Amit Sharma, Ashton Anderson, Fernando Diaz, Hanna Wallach, and Emine Yilmaz. 2016. Auditing Search Engines for Demographic Bias in Performance. In *Proc. of Workshop on Data and Algorithmic Transparency*.
Jacob Metcalf and Kate Crawford. 2016. Where are Human Subjects in Big Data Research? The Emerging Ethics Divide. *The Emerging Ethics Divide* (2016).
Loizos Michael and Jahna Otterbacher. 2014. Write Like I Write: Herding in the Language of Online Reviews. In *Proc. of ICWSM*.
Claire Cain Miller. 2015. When Algorithms Discriminate. *Hidden Bias* (2015).
Hannah Jean Miller, Jacob Thebault-Spieker, Shuo Chang, Isaac Johnson, Loren Terveen, and Brent Hecht. 2016. "Blissfully Happy" or "Ready toFight": Varying Interpretations of Emoji. In *Proc. of ICWSM*.
Tehila Minkus, Kelvin Liu, and Keith W Ross. 2015. Children Seen But Not Heard: When Parents Compromise Children's Online Privacy. In *Proc. of WWW*.
Alan Mislove, Sune Lehmnan, Yong-Yeol Ahn Ahn, Jukka-Pekka Onnela, and J Niels Rosenquist. 2011. Understanding the Demographics of Twitter Users. In *Proc. of ICWSM*.
Ian I Mitroff and Abraham Silvers. 2010. *Dirty rotten strategies: How we trick ourselves and others into solving the wrong problems precisely*. Stanford University Press.
Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The Twitter of babel: Mapping world languages through microblogging platforms. *PloS one* 8, 4 (2013).
Fred Morstatter, Jürgen Pfeffer, and Huan Liu. 2014. When is it biased?: assessing the representativeness of Twitter's streaming API. In *Proc. of WWW Companion*.
Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. In *Proc. of ICWSM*.
Abbe Mowshowitz and Akira Kawaguchi. 2005. Measuring search engine bias. *Information processing & management* 41, 5 (2005).
Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social influence bias: A randomized experiment. *Science* 341, 6146 (2013).
Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me?: message content in social awareness streams. In *Proc. of CSCW*.
Arvind Narayanan and Bendert Zevenbergen. 2015. No Encore for Encore? Ethical Questions for Web-Based Censorship Measurement. *Ethical Questions for Web-Based Censorship Measurement* (2015).
Nasir Naveed, Thomas Gottron, Jérôme Kunegis, and Arifah Che Alhadi. 2011. Searching microblogs: coping with sparsity and document quality. In *Proc. of CIKM*.
Edward Newell, Stefan Dimitrov, Andrew Piper, and Derek Ruths. 2016a. To Buy or to Read: How a Platform Shapes Reviewing Behavior. In *Proc. of ICWSM*.
Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016b. User Migration in Online Social Networks: A Case Study on Reddit During a Period of Community Unrest. In *Proc. of ICWSM*.
Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *arXiv preprint arXiv:1508.07544* (2016).
D-P. Nguyen, R. B. Trieschnigg, A. S. Doğruöz, R. Gravel, M. Theune, T. Meder, and F. M. G. de Jong. 2014. Why Gender and Age Prediction from Tweets is Hard: Lessons from a Crowdsourcing Experiment. In *Proc. of ACL*.
Austin Nichols and others. 2007. Causal inference with observational data. *Stata Journal* 7, 4 (2007).
Dimitar Nikolov, Diego FM Oliveira, Alessandro Flammini, and Filippo Menczer. 2015. Measuring online social bubbles. *PeerJ Computer Science* 1 (2015).
Shirin Nilizadeh, Anne Groggel, Peter Lista, Srijita Das, Yong-Yeol Ahn, Apu Kapadia, and Fabio Rojas. 2016. Twitters Glass Ceiling: The Effect of Perceived Gender on Online Visibility. In *Proc. of ICWSM*.

Andre Oboler, Kristopher Welsh, and Lito Cruz. 2012. The danger of big data: Social media as computational social science. *First Monday* 17, 7 (2012).

UN OCHA. 2014. *Hashtag standards for emergencies*. Technical Report.

Anne Oeldorf-Hirsch, Brent Hecht, Meredith Ringel Morris, Jaime Teevan, and Darren Gergle. 2014. To search or to ask: the routing of information needs between traditional search engines and social networks. In *Proc. of CSCW*.

Paul Ohm. 2010. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review* 57 (2010).

Hüseyin Oktay, Brian J Taylor, and David D Jensen. 2010. Causal discovery in social media using quasi-experimental designs. In *Proc. of Workshop on Social Media Analytics*.

Alexandra Olteanu, Carlos Castillo, Nicholas Diakopoulos, and Karl Aberer. 2015. Comparing Events Coverage in Online News and Social Media: The Case of Climate Change. In *Proc. of ICWSM*.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2014a. CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises. In *Proc. of ICWSM*.

Alexandra Olteanu, Anne-Marie Kermarrec, and Karl Aberer. 2014b. Comparing the Predictive Capability of Social and Interest Affinity for Recommendations. In *Proc of. WISE*.

Alexandra Olteanu and Guillaume Pierre. 2012. Towards Robust and Scalable Peer-to-peer Social Networks. In *Proc. of the 5th Workshop on Social Network Systems*.

Alexandra Olteanu, Onur Varol, and Emre Kıcıman. 2017. Distilling the Outcomes of Personal Experiences: A Propensity-scored Analysis of Social Media. In *Proc. of CSCW*.

Alexandra Olteanu, Sarah Vieweg, and Carlos Castillo. 2015. What to expect when the unexpected happens: Social media communications across crises. In *Proc. of CSCW*.

Alexandra Olteanu, Ingmar Weber, and Daniel Gatica-Perez. 2016. Characterizing the Demographics Behind the #BlackLivesMatter Movement. In *Proc. of AAAI Spring Symposia*.

Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown. http://www.worldcat.org/isbn/0553418815

Miles Osborne and Mark Dredze. 2014. Facebook, Twitter and Google Plus for Breaking News: Is There a Winner?. In *Proc. of ICWSM*.

Miles Osborne, Ashwin Lall, and Benjamin Van Durme. 2014. Exponential Reservoir Sampling for Streaming Language Models. In *Proc. of ACL*.

Raphael Ottoni, Diego Las Casas, João Paulo Pesce, Wagner Meira Jr, Christo Wilson, Alan Mislove, and Virgilio Almeida. 2014. Of Pins and Tweets: Investigating How Users Behave Across Image-and Text-Based Social Networks. In *Proc. of ICWSM*.

Raphael Ottoni, Joao Paulo Pesce, Diego B Las Casas, Geraldo Franciscani Jr, Wagner Meira Jr, Ponnurangam Kumaraguru, and Virgilio Almeida. 2013. Ladies First: Analyzing Gender Roles and Behaviors in Pinterest. In *Proc. of ICWSM*.

Aditya Pal and Scott Counts. 2011. What's in a@ name? How Name Value Biases Judgment of Microblog Authors. In *Proc. of ICWSM*.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and consequences in geotagged twitter data. *arXiv preprint arXiv:1506.02275* (2015).

Umashanthi Pavalanathan and Jacob Eisenstein. 2016. Emoticons vs. emojis on Twitter: A causal inference approach. In *Proc. of AAAI Spring Symposia*.

Sai Teja Peddinti, Keith W Ross, and Justin Cappos. 2014. On the internet, nobody knows you're a dog: a twitter case study of anonymity in social networks. In *Proc. of COSN*.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003).

Kien Pham, Aécio Santos, and Juliana Freire. 2016. Understanding Website Behavior Based on User Agent. In *Proc. of SIGIR*.

Barbara Poblete, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. Do all birds tweet the same?: characterizing twitter around the world. In *Proc. of CIKM*.

Lindsay Poirier. 2015. Reforming Mis-care in Big Data Analysis. (2015).

Liza Potts, Joyce Seitzinger, Dave Jones, and Angela Harrison. 2011. Tweeting disaster: hashtag constructions and collisions. In *Proc. of SIGDOC*.

Chris Preist, Elaine Massung, and David Coyle. 2014. Competing or aiming to be average?: normification as a means of engaging digital volunteers. In *Proc. of CSCW*.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PloS one* 10, 9 (2015).

Davide Proserpio, Scott Counts, and Apurv Jain. 2016. The psychology of job loss: using social media data to characterize and predict unemployment. In *Proc. of WebSci*.

Cynthia LS Pury. 2011. Automation can lead to confounds in text analysis Back, Küfner, and Egloff (2010) and the Not-So-Angry Americans. *Psychological science* (2011).

Giovanni Quattrone, Licia Capra, and Pasquale De Meo. 2015. There's No Such Thing as the Perfect Map: Quantifying Bias in Spatial Crowd-sourcing Datasets. In *Proc. of CSCW*.

Kira Radinsky, Krysta Svore, Susan Dumais, Jaime Teevan, Alex Bocharov, and Eric Horvitz. 2012. Modeling and predicting behavioral dynamics on the web. In *Proc. of WWW*.

Filip Radlinski, Paul N Bennett, and Emine Yilmaz. 2011. Detecting duplicate web documents using click-through data. In *Proc. of WSDM*.

Erhard Rahm and Hong Hai Do. 2000. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.* 23, 4 (2000).

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proc. of SMUC*.

Laura Reed and danah boyd. 2016. *Who Controls the Public Sphere in an Era of Algorithms? Questions and Assumptions*. Technical Report. Data & Society Institute.

Paul Resnick, R Kelly Garrett, Travis Kriplean, Sean A Munson, and Natalie Jomini Stroud. 2013. Bursting your (filter) bubble: strategies for promoting diverse exposure. In *Proc. of CSCW Companion*.

Christian Reuter and Simon Scholl. 2014. Technical limitations for designing applications for social media. In *Mensch & Computer*.

Matthew Richardson. 2008. Learning about the world through long-term query logs. *ACM Transactions on the Web* (2008).

Daniel M Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proc. of WWW*.

Alex Rosenblat, Tamara Kneese, and danah boyd. 2014. Networked Employment Discrimination. *Open Society Foundations' Future of Work Commissioned Research Papers* (2014).

Mattias Rost, Louise Barkhuus, Henriette Cramer, and Barry Brown. 2013. Representation and communication: Challenges in interpreting large social media datasets. In *Proc. of CSCW*.

Cynthia Rudin and Kiri L Wagstaff. 2014. Machine learning for science and society. *Machine Learning* 95, 1 (2014).

Eduardo Ruiz, Vagelis Hristidis, and Panagiotis G Ipeirotis. 2014. Efficient filtering on hidden document streams. In *Proc. of ICWSM*.

Derek Ruths and Jürgen Pfeffer. 2014. Social media for large studies of behavior. *Science* 346 (2014).

KJ Ryan, JV Brady, RE Cooke, DI Height, AR Jonsen, P King, K Lebacqz, DW Louisell, D Seldin, E Stellar, and others. 1978. *The Belmont report: Ethical principles and guidelines for the protection of human subjects of research*. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research.

Diego Saez-Trumper, Carlos Castillo, and Mounia Lalmas. 2013. Social media news communities: gatekeeping, coverage, and statement bias. In *Proc. of CIKM*.

Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2014. On stopwords, filtering and data sparsity for sentiment analysis of Twitter. In *Proc. of LREC*.

Hassan Saif, Yulan He, and Harith Alani. 2012. Alleviating data sparsity for twitter sentiment analysis. In *Proc. of CEUR Workshop*.

Haji Mohammad Saleem, Yishi Xu, and Derek Ruths. 2014. Effects of disaster characteristics on Twitter event signature. *Procedia engineering* 78 (2014).

Matthew J. Salganik. 2017. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press.

Justin Sampson, Fred Morstatter, Ross Maciejewski, and Huan Liu. 2015. Surpassing the Limit: Keyword Clustering to Improve Twitter Sample Coverage. In *Proc. of Hypertext*.

Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry* (2014).

Tatjana Scheffler and Christopher CM Kyba. 2016. Measuring Social Jetlag in Twitter Data. In *Proc. of ICWSM*.

Sarita Yardi Schoenebeck. 2013. The Secret Life of Online Moms: Anonymity and Disinhibition on YouBeMom.com. In *Proc. of ICWSM*.

Doug Schuler. 1994. Social computing. *Communications of ACM* 37, 1 (1994).

H Andrew Schwartz, Greg Park, Maarten Sap, Evan Weingarten, Johannes Eichstaedt, and Margaret Kern.

2015. Extracting Human Temporal Orientation in Facebook Language. In *Proc. of NAACL HLT*.
Julia Schwarz and Meredith Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proc. of CHI*.
Cosma Rohilla Shalizi and Andrew C Thomas. 2011. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research* 40, 2 (2011).
Guy Shani and Asela Gunawardana. 2011. Evaluating recommendation systems. In *Recommender systems handbook*. Springer.
Amit Sharma and Dan Cosley. 2016. Distinguishing between personal preferences and social influence in online activity feeds. In *Proc. of CSCW*.
Amit Sharma, Jake M. Hofman, and Duncan J. Watts. 2015. Estimating the Causal Impact of Recommendation Systems from Observational Data. In *Proc. of EC*.
Martin Shelton, Katherine Lo, and Bonnie Nardi. 2015. Online Media Forums as Separate Social Lives: A Qualitative Study of Disclosure Within and Beyond Reddit. *Proc. of iConference* (2015).
David Silverman. 2013. *Doing qualitative research: A practical handbook*. SAGE Publications Limited.
Fabrizio Silvestri. 2010. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval* 4, 12 (2010).
Meredith M Skeels and Jonathan Grudin. 2009. When social networks cross boundaries: a case study of workplace use of Facebook and Linkedin. In *Proc. of GROUP*.
Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009).
Lucia Specia and Enrico Motta. 2007. Integrating folksonomies with the semantic web. In *The semantic web: research and applications*. Springer.
Michele Starnini, Andrea Baronchelli, and Romualdo Pastor-Satorras. 2016. Temporal correlations in social multiplex networks. *arXiv preprint arXiv:1606.06626* (2016).
Abhay Sukumaran, Stephanie Vezich, Melanie McHugh, and Clifford Nass. 2011. Normative influences on thoughtful online participation. In *Proc. of CHI*.
Jie Tang, Tiancheng Lou, and Jon Kleinberg. 2012. Inferring Social Ties Across Heterogenous Networks. In *Proc. of WSDM*.
Sean J Taylor, Lev Muchnik, and Sinan Aral. 2014. Identity and opinion: A randomized experiment. *Available at SSRN 2538130* (2014).
Jaime Teevan, Daniel Ramage, and Merredith Ringel Morris. 2011. #TwitterSearch: a comparison of microblog search and web search. In *Proc. of WSDM*.
Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, and Chris Parnin. 2016. *Gender bias in open source: Pull request acceptance of women versus men*. Technical Report. PeerJ.
A Torralba and AA Efros. 2011. Unbiased look at dataset bias. In *Proc. of CVPR*.
Florian Tramer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2015. Discovering Unwarranted Associations in Data-Driven Applications with the FairTest Testing Toolkit. *arXiv preprint arXiv:1510.02377* (2015).
William M. Trochim. October 20, 2006. The Research Methods Knowledge Base, 2nd Edition. (October 20, 2006). http://www.socialresearchmethods.net/kb/
Zeynep Tufekci. 2014. Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. In *Proc. of ICWSM*.
Johan Ugander, Brian Karrer, Lars Backstrom, and Jon Kleinberg. 2013. Graph Cluster Randomization: Network Exposure to Multiple Universes. In *Proc. of KDD*.
UN OCHA. 2014. *Humanitarianism in the age of cyber-warfare*. UN OCHA Policy Development and Studies Branch.
US White House. 2016. Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights. *Executive Office of the President, White House* (2016).
José Van Dijck. 2013. "You have one identity": performing the self on Facebook and LinkedIn. *Media, Culture & Society* 35, 2 (2013).
Sarah Vieweg, Oliver L Haimson, Michael Massimi, Kenton O'Hara, and Elizabeth F Churchill. 2015. Between the Lines: Reevaluating the Online/Offline Binary. In *Proc. of CHI Extended Abstracts*.
Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proc. of CHI*.
Michail Vlachos, Christopher Meek, Zografoula Vagena, and Dimitrios Gunopulos. 2004. Identifying similarities, periodicities and bursts for online search queries. In *Proc. of SIGMOD*.
Yana Volkovich, Salvatore Scellato, David Laniado, Cecilia Mascolo, and Andreas Kaltenbrunner. 2012. The

Length of Bridge Ties: Structural and Geographic Properties of Online Social Interactions. In *Proc. of ICWSM*.

Claudia Wagner, Silvia Mitter, Christian Körner, and Markus Strohmaier. 2012. When social bots attack: Modeling susceptibility of users in online social networks. *Making Sense of Microposts* (2012).

Kiri Wagstaff. 2012. Machine learning that matters. *arXiv preprint arXiv:1206.4656* (2012).

H Wallach. 2014. Big data, machine learning, and the social sciences: Fairness, accountability, and transparency. In *NIPS Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Gang Wang, Sarita Y Schoenebeck, Haitao Zheng, and Ben Y Zhao. 2016. "Will Check-in for Badges": Understanding Bias and Misbehavior on Location-Based Social Networks. In *Tenth International AAAI Conference on Web and Social Media*.

Yang Wang, Gregory Norcie, Saranga Komanduri, Alessandro Acquisti, Pedro Giovanni Leon, and Lorrie Faith Cranor. 2011. I regretted the minute I pressed share: A qualitative study of regrets on Facebook. In *Proc. of SOUPS*.

Ingmar Weber, Claudia Wagner, Markus Strohmaier, and Luca Maria Aiello. 2016. Computational Social Science for the World Wide Web (CSSW3). In *WWW Tutorials*.

Katrin Weller, GE Gorman, and GE Gorman. 2015. Accepting the Challenges of Social Media Research. *Online Information Review* 39, 3 (2015).

Katrin Weller and Katharina E Kinder-Kurlanda. 2015. Uncovering the Challenges in Collection, Sharing and Documentation: The Hidden Data of Social Media Research?. In *Proc. of ICWSM*.

Ryen White. 2013. Beliefs and biases in web search. In *Proc. of SIGIR*.

Ryen W. White. 2016. *Interactions with Search Systems*. Cambridge University Press, Cambridge.

Fons Wijnhoven and Oscar Bloemen. 2014. External validity of sentiment mining reports: Can current methods identify demographic biases, event biases, and manipulation of reviews? *Decision support systems* 59 (2014).

Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. 2009. User Interactions in Social Networks and Their Implications. In *Proc. of EuroSys*.

Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang. 2013. Quantifying Political Leaning from Tweets and Retweets. In *Proc. of ICWSM*.

World Medical Association. 1964. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects (re-published 2002). *Journal of postgraduate medicine* 48, 3 (1964).

Shaomei Wu, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Who says what to whom on Twitter. In *Proc. of WWW*.

Yusuke Yamamoto and Katsumi Tanaka. 2011. Enhancing credibility judgment of web search results. In *Proc. of CHI*.

Xin Yan, Raymond Y.K. Lau, Dawei Song, Xue Li, and Jian Ma. 2011. Toward a Semantic Granularity Model for Domain-specific Information Retrieval. *ACM Trans. Inf. Syst.* 29, 3 (2011).

Jiang Yang, Meredith Ringel Morris, Jaime Teevan, Lada A Adamic, and Mark S Ackerman. 2011. Culture Matters: A Survey Study of Social Q&A Behavior. In *Proc. of ICWSM*.

Taha Yasseri, Robert Sumi, and János Kertész. 2012. Circadian patterns of wikipedia editorial activity: A demographic analysis. *PloS one* 7, 1 (2012).

Andrew Yates, Alek Kolcz, Nazli Goharian, and Ophir Frieder. 2016. Effects of Sampling on Twitter Trend Detection. In *Proc. of LREC*.

Elad Yom-Tov. 2016. *Crowdsourced Health: How What You Do on the Internet Will Improve Medicine*. Mit Press.

Muhammad Bilal Zafar, Parantapa Bhattacharya, Niloy Ganguly, Krishna P Gummadi, and Saptarshi Ghosh. 2015. Sampling content from online social networks: Comparing random vs. expert sampling of the Twitter stream. *ACM Transactions on the Web* 9, 3 (2015).

Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, and others. 2014. Inferring international and internal migration patterns from twitter data. In *Proc. of WWW Companion*.

Zengbin Zhang, Lin Zhou, Xiaohan Zhao, Gang Wang, Yu Su, Miriam Metzger, Haitao Zheng, and Ben Y Zhao. 2013. On the validity of geosocial mobility traces. In *Proc. of HotNets*.

Michael Zimmer. 2010. "But the data is already public": on the ethics of research in Facebook. *Ethics and information technology* 12, 4 (2010).

Michael Zimmer and Nicholas John Proferes. 2014. A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management* 66, 3 (2014).