Data Acquisition, Sampling, and Data Preparation Considerations for Quantitative Social Science Research Using Social Media Data

Zeina Mneimneh[1], Josh Pasek[1], Lisa Singh[2], Rachel Best[1], Leticia Bode[2], Elizabeth Bruch[1], Ceren Budak[1], Pam Davis-Kean[1], Katharine Donato[2], Nicole Ellison[1], Andrew Gelman[3], Erica Groshen[4], Libby Hemphill[1], William Hobbs[4], Brad Jensen[2], George Karypis[5], Jonathan Ladd[2], Amy O'Hara[2], Trivellore Raghunathan[1], Philip Resnik[6], Rebecca Ryan[2], Stuart Soroka[1], Michael Traugott[1], Brady T. West[1], and Stefan Wojcik[7]

March 15, 2021

[1] University of Michigan
[2] Georgetown University
[3] Columbia University
[4] Cornell University
[5] University of Minnesota
[6] University of Maryland
[7] Northeastern University

This series of white papers is funded by National Science Foundation awards #1934925 and #1934494 as a collaboration between Georgetown University and University of Michigan. To learn more about The Future of Quantitative Research in Social Science research project, visit www.smrconverge.org.

1. Project Overview

In 2019, a group of computer and social scientists began a project to converge these disciplines, with the aim of harnessing data from social media to improve our understanding of human behavior. People all over the world have started using social media, search engines, smart devices, and other technologies that record their moment-to-moment behaviors (often called, "digital traces"). Social media, in particular, provides a massive amount of information on the everyday activities, opinions, thoughts, emotions, and behaviors of individuals, groups, and organizations in near real-time. Today, most adults in the US use some form of social media (Perrin & Anderson, 2019) to share and discuss topics as wide-ranging as politics, employment, parenthood, leisure activities, travel, sports, and health, to name only a few, making these platforms potentially excellent sources of information on constructs relevant to all social science fields.

Expanding the availability and utility of this extremely rich but still underutilized data source in the social sciences, however, requires attention to the unique features of these data. Unlike many forms of standard social science data, social media data have no set structure and are not the product of a designed process initiated by the researcher to answer specific hypotheses or questions. Instead, the data are provided "as is," which often means they are raw, complex, and highly dense in nature. Moreover, these data involve unique bias concerns not typically at issue in traditional social science methods, including often not knowing who generated the data or what population they represent. The organic nature of the data, along with their magnitude and complexity, require methods for managing and structuring the data -- which are most commonly found within the toolbox of computer scientists. Together with social science methods, they can generate useful measures for social scientific inquiry and represent a fruitful convergence across disciplines.

While employing methods from computer science to wrangle digital trace data in order to answer social science questions has enormous potential, it also presents a number of challenges. First, neither field is well versed in the others' methods. So before social scientists can begin using ideas and algorithms from computer science, they need to learn how to work with large-scale unstructured organic data and understand the general principles, tools, and methods used by computer scientists. Likewise, computer scientists can reach inaccurate conclusions if they fail to understand key considerations and objectives within social science research that may not traditionally apply in computer science. Second, it is often unclear who or what, exactly, are behind the accounts that produce the data appearing in social media data sets, and how the entities creating trace data might relate to the larger groups of people that social science researchers would like to understand. For many questions, social media data contain information about the presence of a behavior, but do not have information about why or under what circumstances that behavior may have occurred (or may not have occurred) -- which is a key area of focus for many social scientists. They also do not have sufficient information about the demographics of the users participating in the conversation, including spatial/geographic location

information, further limiting its value for social scientists. Third, ethical questions around the use of digital trace data in research contexts requires collaboration across these disciplines. Understanding what we need to know about the data that are gathered, what other data are needed to supplement them to answer central research questions, and how to do so responsibly is critical for digital trace data to live up to their full potential.

Thus the convergence of methods and relevant theories between computer scientists and social scientists is a necessary condition for leveraging social media data to understand this increasingly important window into human societies. The advantages to social science in effectively harnessing these data are clear. But for computer scientists too, this convergence holds great opportunity. Designing algorithms with a new set of constraints and optimizing existing algorithms for this large-scale, real-time domain, all while addressing privacy, bias, misinformation, and algorithmic fairness concerns, will not only advance computer science research but can also help to make their methods, models, and tools more usable by social science researchers and more applicable to solving societal challenges.

To initiate this convergence, our group planned a set of topical meetings bringing together social scientists from multiple disciplines, including economics, psychology, political science, communication, demography, sociology, and survey methodology with data scientists and computer scientists with the goal of creating a common set of methodologies for how to study complex human behaviors using social media data in a scientifically rigorous manner. The topics of these meetings address each stage of the research process as we have defined it: study design; data acquisition, sampling, and data preparation; measurement and feature engineering; model construction; analysis and storytelling. At each meeting, we also discuss criteria for the responsible conduct of research with social media data.

This is the third in a series of white papers that will provide a summary of the discussions and future directions that are derived from these topical meetings, with this paper focusing on issues of data acquisition, sampling, and data preparation. These topics incorporate data collection methods, sampling strategies, population mismatch adjustments, and other data acquisition and data preparation decisions.

2. Key Differences Between Social Media and Designed Data

The properties of social media data that are generated as a product of users' behavior differ from the properties of data that are generated by a planned research method. For example, before collecting survey data, a social science researcher formulates specific hypotheses, designs a questionnaire that taps the constructs of interest, selects a sample that satisfies certain eligibility criteria, and collects respondents' answers to these survey questions. While the final sample characteristics are not fully under the researcher's control (since many sample members will not respond to the survey invitation), the "designed" constructs and characteristics of the respondents are known. Social media data, on the other hand, are not generated for the purpose of measuring a construct of interest; instead they are produced in response to stimuli unknown to

the researcher and made available by companies whose principal interest in doing so is not in providing data for basic research. Moreover, in many instances, the characteristics of the users generating the data, who they represent, and how to model their probability of generating the data are also unknown. Although some features of such observed organic data resemble observed qualitative data or data used in ethnographic studies (see Study Design white paper - Bode et al., 2020), social media data still lack essential information about the characteristics of the users (especially for researchers who are not part of the social media industry), and the full set of contextual factors that lead to their generated data.

Essential differences between the production of social media data and that of designed data are further complicated by variability in the design and purpose of different social media platforms. Social media platforms differ on the type of data generated, with some focused-on text, others images and videos, to name a few. Moreover, each platform has specialized actions that users can perform to interact with the shared content, such as likes, reposts, and up-votes. Finally, the persistence of the data on platforms also differs. While some are automatically erased after a certain period, others remain as long as the user does not delete them.

From a research design perspective, these significant data differences make designed data a better fit for a hypothesis-driven approach, while organic social media data align better with a data-driven research approach. In its purest form, data-driven research takes as a given the data that have been collected and uses a series of approaches to understand social phenomena using those data. In contrast, pure hypothesis-driven research rigidly follows a traditional view of the scientific method wherein scholars first derive a specific hypothesis from a theory and subsequently collect (or limit their observations to) only the data necessary to test the research question proposed. In reality, most research combines elements of the data-driven approach and the hypothesis-driven approach into a hybrid approach. That is, researchers iterate to some degree between, on the one hand, pre-specified analyses of data collected to test a specific hypothesis with predetermined measures and, on the other, analyses aimed at better understanding some phenomenon through the use of more exploratory analyses.

These differences between designed data and organic data pose essential questions about data acquisition, sampling, and data preparation. For example, since the data are generated directly by the user and some are available publicly, what methods are available for acquiring these data? What are the tools, infrastructure and knowledge that social scientists need to have on their team to collect these data? How should we think about user consent in the “public” space? Can these data that are generated organically be sampled in such a way as to represent a larger population? Can population inferences be made with certain types of adjustments? What kind of data preparation activities are needed for such diverse unstructured data? Do researchers need to identify and exclude certain users? What are common data cleaning approaches (e.g. removing flood words, stemming, etc.) that are relevant to social media data? And what types of data aggregation might be needed to deal with issues of missing information or privacy concerns?

This paper is a first attempt to delve into these questions that relate to data acquisition, sampling, and data preparation for social media data. The remainder of the paper is organized as follows. We begin with six sections that discuss different considerations broadly related to data acquisition, sampling, and data preparation. Section 3 focuses on data collection approaches and when to use different strategies. In Section 4, we discuss different sampling strategies. Depending on the study objective and the sampling strategy adopted, deciding on whether sample adjustment is needed and the methods available for adjustment are discussed in Section 5. Sections 6 and 7 focus on the activities undertaken after acquiring large scale social media data including storage, cleaning, and aggregation. The final essential consideration relating to ethical practices of collecting social media data are discussed in Section 8. Different suggestions for accelerating research and examples of existing shared resources are presented in Section 9. Finally, conclusions and recommendations are presented in Section 10.

It is worth noting that many of the considerations discussed around data collection (Section 3), data storage (Section 7), and informed consent (Section 9) have been the subject of discussions by the internet research community (e.g., Association of Internet Researchers (AOIR)) since the 1990s. AOIR working groups have produced a series of reports summarizing community standards regarding data collection, analysis, and reproduction in papers. The first of these was published in 2002; see Franzke and colleagues (2020) for the latest revision. Therefore, although these may be new issues for some disciplines, there is a set of practices already documented that might be useful for contemporary social media researchers to consider. As we go through the considerations in different sections, we refer to the best practices and other guidelines that are used for other types of organic, publicly available data when applicable.

3. Data Collection: Key Considerations

As in other areas of research, the specific approaches to collecting social media data can have important implications for the conclusions that can be reached using that data, the potential for different sources of error, and the capacity for other researchers to replicate and expand on original research conducted with those data. There are many different approaches for accessing social media data. Here we define the five most common approaches and accepted practices for collecting social media data to investigate a research question.

The first approach involves collaborating on a study with a social media company. Companies like Facebook have calls for proposals related to some of their interests (election integrity, misinformation, etc.), and give researchers free access to relevant group data and URL data. For example, Social Science One is a consortium of researchers that Facebook has given access to some data in order to better conduct meaningful social science research (Harvard's Institute for Quantitative Social Science, 2021).

The second strategy employs existing databases and archives that contain social media data. One example is Harvard’s Dataverse (Harvard Dataverse, 2021). They have data related to different research projects available for download. Typically, if the data are large social media

data sets, the post ids can be shared publicly, but behavioral and contextual data must be obtained from the social media company. If the data sets are small, some social media companies allow them to be shared for research purposes if they are anonymized. Unfortunately, few databases exist that are publicly shared (see Section 10 for examples of shared data resources). If the data are already available through a public archive, then no data collection activity is needed and the researcher can access the data directly.

A third method is to work with a data vendor that collects the data and sells it to researchers, both in its raw form and in aggregated formats. Example data vendors include Sysomos (2021) and Brandwatch (2021). Cost can be an issue when considering this route. However, if researchers do not have students or staff, or collaborators who have the knowledge and skills to write programs for collecting the data, then working with a data vendor or academic institute with an infrastructure that can process large-scale data are promising options.

A fourth approach to collect social media data is to use web scraping tools. Web scrapers or data scrapers are automated tools and software libraries that can be used to efficiently collect structured data from webpages. These automated scraping tools and libraries exist in many programming languages, including Java, Python, and R, and in some applications like Excel. Researchers using scrapers extract data from tables, lists, or other well-structured parts of a webpage. Data scraping is typically used in two scenarios. First, the researcher may have limited funds for data collection. Second, the social media platform may not have an interface for data collection (Application Program Interfaces) available to researchers, sometimes as a deterrent for capturing their public data. Legally, if the data are public, and an interface for data collection does not exist, the data can be scraped. Although this is continually being litigated, in 2019 the 9th Circuit Court of Appeals determined that scraping did not violate the Computer Fraud and Abuse Act (*HiQ Labs Incorporated v. LinkedIn Corporation*).

While the free nature of this approach makes it attractive, another large issue is that scraping data is sometimes harder than using company provided data collection tools and requires data extractors that are customized for each website. When scraping data, if the layout of the web page changes, the data scraper must be adjusted accordingly. For example, if a website switches from a list format to a table format for presenting data, the scraper will need to be rewritten to look for data within table cells. This means that data collection can get interrupted every time an update is made to the way information is presented, and the researchers need to continually monitor the web page and update the code. Also, if researchers try to scrape too much data too quickly from a website, the researcher may be blocked from using the site or rate limiting may be enforced. In such cases, IP anonymization services (that rotate between a pool of IPs) can be used to circumvent the blocking so long as the terms of service are still adhered to. When information is scraped from websites, the extraction can be aided if the content in the source websites is tagged using public schemas (e.g., schema.org). Such schemas tag the content of the website with meta-information to describe their structure, thereby allowing it to easily be placed into certain data storage options like a relational database.

Fortunately, for most of the major social media platforms, computer scientists and data scientists have created scraping libraries for scraping different platforms, making it easier for researchers who have less experience with web programming. They also make updates to these libraries as websites change. For example, Requests (Reitz, 2020) and BeautifulSoup (Richardson, 2020) are general purpose web scraping libraries that make requesting web pages and parsing website page text straightforward. Similarly, for Twitter data, Tweepy is a popular web scraping library that can be used to collect posts containing a particular keyword or hashtag (Roesslein, 2020).

Finally, the fifth and most popular way to collect social media data is through Application Program Interfaces (APIs). An API is a software interface that specifies how a computer program or application can interact with another application. APIs establish secure and standard ways for software applications to work with each other. The researcher is given a "key" by a social media platform so he/she can access specific data. Then the user can use this key to request data from the API. The most popular framework for authorization and authentication is oauth2 (IETF OAuth Working Group, 2021). Given the complexity of coding and the need for customizing web scraping algorithms, APIs, when available, are the most common approach for acquiring social media data. APIs have been developed by Twitter, Google, and other larger social media sites to allow researchers and others to access data.

Because APIs are typically designed by software developers at the social media platforms, there tends to be reasonable documentation, sample code, and tips shared about using them. APIs are also a reliable source of platform information since companies develop them as an additional source of revenue and are therefore incentivized to maintain them. APIs, however, vary in 1) the types of data they offer, 2) the amount of data a company will make available, and 3) and the cost of the data collected. In terms of types of data, it is typical for post-level queries to collect information such as post text, author of the post, date of post, likes, number of posts shared (e.g. retweet count), keyword, hashtag, channel, or group; user-level queries might gather biographies, other demographic information shared, lists of followers and individuals followed, and account statistics (such as date created, location, and total number of posts); group-level queries might elicit information about group purpose, members, and posts. However, it is rare for all of these options to be available. Typically, platforms require those using APIs to specify a time range and the type of data they want to collect. Also, it is common for old data to be less accessible through free APIs as some sites either archive or delete those data (though it is sometimes available through third-party vendors or through special requests).

Some sites have *streaming* APIs in which data are accessible as soon as they are posted, but old data are not accessible. While some platforms may limit data streams to a percentage of data, typically, streaming data is not a sample. Rather, it is the complete set of data posted. When using streaming APIs, interruptions in data collection can also yield missing data. Examples include network interruptions, company server failures, or poorly written code that cannot handle certain types of data failures. Moreover, a researcher needs to determine the features of the data

that are needed for the study. If there is a match between the data a researcher needs and the data obtainable from the API, then using an API is ideal. Even if there is an API that provides the specific data of interest, it is not unusual for companies to share some of their public user data freely and require payment for more extensive data sets.

Unfortunately, the properties of the sample acquired through the different types of APIs are not always clearly explained or ideal for researchers. For example, Twitter has historically had an option for academic decahose licenses for a fee which is considerably less than getting all the data (firehose option[8]). The decahose is a ten percent sample of posts. However, the exact nature of the randomness of the 10% sample from the decahose is not clear. The 1% API sample raises similar concerns (Baumgartner, 2019). Thus, studies that aim at inferring their results to the user population face challenges in making such inference. This is discussed further in the next section. Finally, the use of API data may be limited by additional terms of service, which can reduce the value of the data for research purposes (e.g. by prohibiting inference of specific user demographics).

The costs of API use and the challenge of maintaining API-based data sets can also introduce inequities into both research and reproducibility. Those using APIs are more likely to be parts of large organizations that have the resources to manage the data obtained through the API. And variations in access, coupled with limits in sharing the data (and, for streaming APIs, inconsistencies in the results of API data collection across those with access) can make research harder to replicate (Freelon, 2018a). Web scraping, on the other hand, is available to all levels of students and researchers, limited centrally by constraints on data storage, programming knowledge, and skills of any given research team. Thus, one can view web scraping as a more democratic data collection approach. That is one reason why developing portals for constructing variables from social media data that are available for all researchers is an important direction for future research.

In summary, all of these methods produce data that differ in the availability of different pieces of information, the accessibility of that data to other researchers, and whether the data collection can be clearly replicated at a later point in time. All are also complicated by constantly fluctuating terms of service, ownership, pricing plans, among other considerations. However, when conducting research using social media data, data collection considerations must be thought through with care.

4. Alternative Sampling Strategies and Study Designs

When researchers collect social media data, they typically do not consider (or they do not have access to) all of the cases that could be relevant to a particular research question (unless they are collaborating with or are part of the social media platform organization). Instead, it is common to sample from the universe of potential entities and only analyze cases that fulfill

[8] Twitter has just announced new, free access for researchers to the Twitter firehose. Because it is new, we have limited information about the process and the different challenges and considerations associated with using it.

certain inclusion criteria. How the sample is selected and what information is available on the sampled cases and its universe or population (from which the sample is selected) impact the type of conclusions and the generalizability of the findings. We begin by considering these issues when using social media data for studies that are descriptive in nature (Section 4.1). We then consider sampling strategies for studies that are inferential in nature (Section 4.2).

4.1. Social Media Studies with Descriptive Research Goals

Many common research designs using social media data follow a descriptive approach, where the research is focused on the data collected, and does not seek to make an inference about the larger population or generalize to one. Some of these designs examine how interventions influence online behavior among the individuals who are directly sampled and observed (Section 4.1.1). Others focus on comparing two groups of individuals that are already defined in the way the data are collected (Section 4.1.2). And yet others are focused on understanding things that happen to users on a platform (Section 4.1.3). While these design types are by no means exhaustive (for example, ethnographic studies that involve deep descriptions of specific populations are also not designed to be generalizable (see e.g., Small, 2009)), they give a sense of studies for which the research does not need to begin with a formal map for linking the sample gathered with a larger population.

4.1.1. Intervention studies

Some studies present an intervention with the goal of changing a behavior or a set of behaviors that are observed with digital trace data. Studies using these approaches are designed to understand cause-and-effect relationships (i.e. internal validity) and to test the impact of stimuli in real world environments (i.e. maximizing ecological validity). Because of these objectives, such studies often are not particularly concerned with the mechanisms by which individuals were selected into the sample. It is common for these studies to rely on self-selection such that users decide to participate without being individually recruited. Intervention studies often employ a design in which: 1) participants are consented to be part of the intervention specifically if contact with users is attempted, 2) informed consent to collect digital trace data is gathered, 3) baseline information are collected on all consented participants, 4) participants are randomized into different intervention groups, 5) digital trace data are collected and analyzed, and 6) post intervention self-reported data might also be collected. The randomization dimension of this design justifies the self-selection, as the randomized groups are assumed to be the same on all potential confounders and triggers except the intervention. Moreover, as in many randomized control studies of real world experiences, those who select into the study are among those who would be most affected by the stimulus outside the study. Thus, differences in digital trace behavior between the randomized groups can be reasonably attributed to the intervention. An example of an intervention study would be to recruit users of a specific platform and invite them to be part of an intervention that aims at increasing their physical activity. Users could be divided randomly into two groups: a control (that receives posts related to basic information

about the benefits of physical activity) and an intervention (where an automated algorithm generated encouragement and motivational exercise messages based on the user's posts).

4.1.2. Case-control studies

Case-control studies compare social media users who fulfill certain criteria (cases) and those who do not (controls) in order to investigate correlates of differences in digital trace behavior. Because observed digital behaviors are the outcome of interest, understanding and measuring the mechanism of selection into the sample is not always necessary for case-control studies. Such studies could also rely on self-selection and use the following design: 1) clear inclusion and exclusion criteria into the case/control groups are specified and measured, 2) participants are consented to be part of the study, particularly when direct contact with users is attempted, 3) informed consent to collect digital trace data is gathered, 4) similar baseline information on cases and controls are measured, 5) cases and controls are matched on baseline measures, and 6) digital trace data are collected and analyzed. Many of these studies proceed to use computational linguistic models to identify language differences between the cases and controls. The main objective could be to develop models to predict the probability of belonging to a case from their digital trace data.

Such studies and models have been found to be promising for constructs such as flu (Wojcik et al., 2021), suicide, or depression (Chancellor et al., 2019; De Choudhury et al., 2013; Resnik et al., 2020). A topic model, for example, may discover latent categories that already correspond to known constructs (e.g. a topic relating to social isolation) or that identify other characteristics a researcher might not have already thought to look for. The model might also identify ways that known constructs might be expressed differently in different groups. For example, topic modeling related to depression in college students identifies a known category, vegetative, i.e. energy-level as expressions of being late for class, taking naps, etc. (Resnik et al., 2015). Although sometimes classifications generated in one arena will not translate to others, sampling is typically not a major concern for these types of case control studies, where the goal is to develop a linguistic or machine learning model that identifies latent features that can be used to separate the cases and the controls.

4.1.3. Simulated account studies

This group of studies focuses on understanding certain properties of the social media platform by intentionally inducing or generating social media posts. For example, Pfeffer and colleagues (2018) investigated potential sampling biases in Twitter APIs by producing tweets that matched the timestamps used by the API to generate systematic 1% and 10% samples. A similar type of study could also simulate users by creating fake accounts that allow researchers to collect platform data, examine messaging patterns, or investigate certain properties of the platform. An example would be simulating certain Youtube user profiles and investigating which videos get pushed/advertised to these users depending on profile characteristics. For this type of research there is no real user sampling occurring.

4.1.4. Considerations in descriptive research study designs

Although the sampling mechanism for the above study types is not a required feature or is at times irrelevant, the generalizability of the findings outside the specific environment of the intervention, the case/control group, and the non-simulated users is sometimes still important. In these situations strict generalizability might not be attainable, and the notion of transferability becomes more relevant. We define transferability as applying the model used in an initial set of studies/samples to a completely new data set as a starting point. The motivation is not to treat the new data set/sample as the same as the initial ones but to start with a nonrandom prior (model), update it and refine it as more information is gathered about the new sample. For example, if a researcher is interested in modeling suicidal ideation using Facebook posts, but has already built a model using Twitter posts, the features used for the Twitter model can serve as initial priors for the Facebook model. In this context, a valuable result is typically one that is generative across contexts, but there is no assumption that what is discovered in one arena will replicate in another.

4.2. Social Media Studies With Inferential Research Goals

Population inferential research assumes the data come from and represent a larger population to which the generated statistics can generalize. This definition of population inference is also different from the notion of inference in the machine learning environment. There, the goal is to build a model using observed data in order to predict an outcome (e.g. the example of case control studies where depression status or suicidal ideation is being classified and predicted). The outcome being predicted is not assumed to reflect some larger population. Instead, the model is used to make good classification or categorization decisions on examples that have not been seen before, having properties similar to the data the model was trained on. However, when scholars have a goal of rendering conclusions about more than simply the data that were collected, they need to consider how the processes of generating that data may have led to differences between the data collected and the population of interest, a complication that renders population inferential goals more challenging than descriptive ones in most cases.

Historically, this type of inference has been associated with probability sample designs - where every unit in the population has a known nonzero chance of being selected. When the selection probability of each sampled unit is known, it is possible to statistically adjust for differences in the likelihood that various individuals were selected (thereby eliminating biases) and to estimate attributes of the larger population with known variance. This notion of probability sampling is also a more democratic way of giving everyone the chance of being included in the science that is used to generate public policies or make decisions that affect the welfare of communities. Thus, probability sampling is built on a system that is seen as more equitable, open, and transparent. This type of sampling, however, requires having access to a list (or a frame) of population units or creating a list/frame where the probability of inclusion is known.

Unfortunately, the conditions necessary for probability sampling are only rarely available for social media studies. For example, accounts on social media do not always uniquely identify

people, resulting in a complicated mapping between the sampled unit and the target population. Moreover, not all people on social media post at similar rates or in similar ways -- leading to unequal probability of selection and potentially biased conclusion if such differences are related to the research outcome and are not accounted for. Most importantly, the probabilities of selection are unknown to the researchers and may be configured by certain data acquisition tools in a way that will over or underrepresent certain posts or users. For example, Pfeffer and colleagues (2018) demonstrate how the 1% and 10% Twitter APIs are not true probability samples of the Twitter population as they are based on a millisecond timestamp window of the post. Thus, not only will spammers and in-sync bots have a higher unknown chance of being included in the API samples, but others who tweet at a regular interval and miss the millisecond window of selection may never be selected. Such conditions pose challenges for inferential research, but to varying degrees depending on the purpose and design of the study.

Still, many social scientists are interested in using social media data to make inferences to a variety of different populations. They may want to say something about the content of posts, attributes of accounts, features of users, or perhaps even a larger population. The extent to which a given sample might represent one of these populations or could be used to generate an accurate understanding of those populations depends on the sampling procedures employed, the ability and quality of mapping the sample to the population, and the magnitude of the sample-population differences. In this section, we consider one common inferential population -- that of the larger community of social media users or posts – and articulate some of the methods that can be used toward this end.

4.2.1 Probability digital panel studies

One method for collecting social media data where the researcher has control over the selection mechanism involves recruiting respondents to panel studies using traditional probability sampling techniques and then collecting their social media data. These are panels that are similar to existing US general internet population panels such as Ipsos/GfK, AmeriSpeak, SSRS, and Understanding America. They initially rely on selecting a probability sample of users through an address-based sample (ABS) design (Harter et al., 2016) or random digit dialing (RDD) telephone sample (Waksberg, 1978; Lepkowski et al., 2008) and then inviting respondents to be part of an online panel where the user's social media data are collected and analyzed. Such designs need to account for the penetration rate of the platform, survey response rates, consent to share and link social media data, and sample retention. Given all these considerations, they tend to require extremely large initial samples. Take for example Twitter. Assuming a population penetration rate of 22% (Perrin & Anderson, 2019), a survey response rate of 9%, and consent rate to provide access to digital trace data of 30% (Mneimneh et al., 2020), the panel would need 1.7 million participants to reach a sample of 10,000 users who are willing to provide their Twitter public data. This sample would also need to be refreshed to keep the panel current. Whether the resulting participant pool is still representative of the population is

also an open question, but one that potentially can be statistically adjusted by utilizing methods such as Multilevel Regression and Post-stratification (MRP; Gelman & Little, 1997).

Such initiatives will require a multi-institution effort to absorb the cost and maintain the demand of such panels. Such a multi-data source initiative is still restrictive since the initial samples are much smaller than what one might collect straight from APIs and thus is not leveraging the full scale and potential of social media data. However, it could provide the needed descriptive information about the social media platform users, help researchers understand the probabilities and triggers of both posting and "silent" users, and validate some of the social media measures, advancing the understanding of social media data properties and propelling the next step of leveraging the full potential of the social media data.

4.2.2 Non-probability user-sample studies

Researchers have mainly attempted to sample social media users through a number of non-probability approaches. One strategy includes generating a random sample of names, using the platform search to identify users with that name, then scraping the data or using an API to return the user's information and posts that map to associated platform user ids (Moore et al., 2013). This approach, however, has many limitations, including the difficulty of understanding the eligible universe of specific names in a platform, measuring the probability of inclusion of a specific name, and the erroneous inclusion or exclusion of certain users because of complications like fake names. Another approach is generating random IDs to create a list of potential existing user IDs, selecting a random sample of IDs from the generated list, and using the API to return public information available on the users who have these ID (Berzofsky et al., 2018; Singh et al., 2015). In addition to the inefficiency of this approach and the difficulty to ascertain that the selected ID fulfill the inclusion criteria for the study, many social media platforms including Twitter stopped using a sequential approach for assigning IDs, making this approach unsuitable for generating a random sample.

Thus, the largest majority of studies that use samples of social media users rely on a third approach which starts with identifying inclusion criteria for the user, such as those that post a certain keyword or hashtag (and/or use a specific language, or provide a geocode for location for a specific geography) and then using the platform API to select a sample of posts that fulfill the inclusion criteria. Once the posts are returned, the users that generated the posts are identified and constitute the sample. As discussed earlier, this approach does not qualify as a random probability sample. Thus, when such studies attempt to generate point estimates, such as overall support of a political candidate, from such samples, inferring the results to the larger social media user population is problematic (cf. Jungherr, 2015; Beauchamp, 2017).

In all of these approaches, indices that measure the potential sampling bias arising from the nonrandom selection are needed (Andridge et al., 2019; Little et. al, 2020). However, the biggest challenge in using such measures is that they require: 1) the availability of auxiliary variables that are correlated with the measure/outcome of research on all users in the sample, and 2) at least sufficient statistics (means, variances and covariances) of the same auxiliary variables

on the *non-selected* users (i.e. the rest of the user population that are not selected). First, this requires that such auxiliary measures that are correlated with the outcome of interest are identified. Second, for such measures to be available for the researchers on all the sampled users, the researchers either need to collaborate with the social media platform itself (and in this situation the researchers might as well select a probability sample), or they need to be able to predict such measures (such as gender, age, education, political affiliations, assuming these are correlated with the outcome) on the selected users from existing trace data such as images, text posted, and friends' information, or they need to link the sampled users to another data source where some of these measures are available (such as voter record registration; Wang et al., 2015; Grinberg et al., 2019).

Unfortunately, collaborating with platforms has been proven to be difficult for most researchers. Predicting simple auxiliary measures such as demographics from trace data, while relatively successful for certain measures such as gender (Chen et al., 2015), still needs significant improvement for other measures such as age and education (Chen et al., 2015; Zhang et al., 2016), and these measures will still be missing on many selected users who do not leave enough digital trace data for accurate prediction. Also, linking users across data sources suffers from mismatches, additional sample restrictions and limited auxiliary variables. Finally, for the second necessary condition- the availability of sufficient statistics (means, variances, and covariances) for the auxiliary measures on the non-sampled users- such statistics will need to be released by the industry regularly or generated from other data sources (such as large scale surveys that identify the specific social media population) so that indices of sampling biases can be calculated and inference to the general user population from the non-probability sample can be made.

4.2.3 Non-probability keywords/hashtags sample studies

The final common design we discuss also relies on a non-probability sample approach, where the inclusion criteria are certain key phrases, keywords, or hashtags, and the unit of analysis is a *post* rather than a user. An example would be a study that aims at measuring support for Black Lives Matter by selecting tweets based on relevant hashtags such as #blacklivesmatter or #BLM. Similar to Section 4.2.2, if point estimates such as sentiments towards an event are to be generated from such samples and inferred to the overall (non-selected) platform, added auxiliary information about the selected and non-selected entities are needed. In addition, in this design, the selection itself is based on the measure or outcome of interest. This means that inclusion of an entity in a sample is determined by not only the API black box algorithm but also on a self-selection mechanism. That is, only the individuals who decided to use this key phrase at the time of selection will have the chance to be included. This self-selected set might have different overall characteristics than the rest of the population and could also exhibit different associations, making the need to calculate sampling bias indices like those discussed in Section 4.2.2 even more important.

Moreover, different keywords and hashtags are related to their specific cultural and socio-political framework, and their frequency of use changes over time even if their triggers remain active. For example, Tufekci (2014) discusses how hashtags are generally used to attract attention to a new topic or event, and once the topic is known, the hashtag might be seen as wasteful to the character limit of the platforms though the topic/event will continue to be discussed. Because of this challenge, analyses of hashtags that do not consider how they might be a product of human behavior can be biased for some purposes. In these contexts, she recommends that hashtag analyses start from understanding the principle of the users' behavior, and then the selection needs to switch to the user who generated the hashtags rather than continue to be based on the hashtag itself. In addition, when relying on hashtags, supplementing social media data with other data sources, e.g. surveys, could help with the changing contextual nature of hashtags and self-selection that is tied to a specific term. Thus again, hybrid designs where multi-data sources, e.g. cheaper social media data and more expensive survey data, are leveraged during different phases of the study could provide a more efficient mixed method approach. For example, one could envision phase one of a study where both survey data and social media data are linked to understand how these two data sources could be mapped and calibrated. Then phase two could proceed with only the cheaper data source, social media data, where adjustments of estimates could be made based on what is learned from phase one. Such alternating phases might need to be repeated several times to account for changes in language and the composition of the self-selected individuals as the study progresses or as different studies build on each other's findings. The notion of supplementing one method with another to better interpret the results and understand the limitations of each method is not new (Campbell & Fiske, 1959). As new data sources emerge, the need for multi-method multi-data source approaches will increase and will likely become the new norm. The literature has many examples of researchers calling for the need to supplement social media data with other types of data sources (Singh et al., 2020a; Lazer et al., 2020; Grinberg et al., 2019; Paparrizos et al., 2016).

Thus, depending on the objective of the research (descriptive vs. inferential) certain types of sampling approaches are more appropriate than others. Studies that aim to generate statistics to be inferred to a larger population beyond the sample at hand need to either rely on a probability sample design (as described above) or assess the gap between the sample and the target population to determine what type of adjustment, if any, is needed. Section 5 describes in more detail when adjustment is usually needed.

5. Adjusting Samples

Depending on the type of study and the research objective, several forms of adjustments to map the sampled data to its originating population or universe have been mentioned above. This section delves deeper into the general topic of adjustment (Section 5.1) and the specific use of inferred demographics for adjustment (Section 5.2).

5.1 Different Approaches to Sample Adjustment

Understanding the links between the data gathered and the target population is important because this relationship helps to clarify the substantive meaning of the data gathered and analyses conducted with respect to a particular research goal. It also highlights the sorts of procedures that should be used to most efficiently reach that goal. That is, if a scholar has in mind a particular attribute they would like to estimate (i.e., an estimand), they need to consider what if any features of the data collection might differentiate between the sample-based estimate and the target population-based estimate for that parameter. A sample adjustment is the approach used to account for these differences.

In general, considerations about whether, how, and when to make sample adjustments depend on 1) the type of statistical inference that researchers are hoping to make, 2) the size of the gap between the units for which observations were made and the target population, and 3) the information that is known about the nature of that gap. In some cases, the method for gathering data can ensure a small gap between the data collection and target population; when a data gathering approach yields a random probability sample of the target population, the only differences between the two are random differences due to sampling and either random or non-random differences due to nonresponse error. In other cases, the differences between sampled units and a target population are well understood as a function of research design; this sort of design-adjustable difference exists when data from social media posts are used to generate understandings of the social media users who create those posts or vice versa. That is, researchers can determine how many posts are associated with a particular user and can figure out how conclusions about posts might apply to users (or how conclusions about users might influence the distribution of posts). And in yet other cases, the link between the sample and the population is unknown, such as when a study attempts to use social media to say something about a broader population (e.g. the people who use a site or the population as a whole).

Different fields of research have taken different tacks toward thinking about the gaps between sample and target population. Scholars of biological and psychological processes, for instance, have often presumed that the phenomena observed in any given population will emerge similarly in others, leading then to conclude that probability sampling may be unnecessary (cf. Sears, 1986; Druckman & Kam, 2011). Fields attempting to describe entire societies, in contrast, have focused heavily on obtaining samples that reflect population parameters by design, to ensure that variations in the estimand of interest (e.g. an attitude) reflect the population distribution of that estimand (Cornesse et al., 2020). And in other domains of research, scholars have focused their attention on ensuring that the estimates they generate are not idiosyncratic by assessing whether similar results can be achieved across differing populations (cf. Krupnikov & Levine, 2014). These field-based differences are often driven by distinctions in the theoretical likelihood that the results obtained under some circumstances would be likely to replicate elsewhere. They are also based on an understanding that studies designed to maximize external validity (i.e. the confidence in being able to generalize to the entire population) will tend to

suffer in the precision of the conclusions they can make about the effect they are studying. Further, as a practical matter, samples that use broader populations tend to have more variance in the estimates that are obtained from them. This is because the presence of covariates among sampled respondents serves as an additional source of random error. Sample sizes need to be larger in broad populations to address these challenges, which can pose problems when researchers are investigating rare phenomena.

There are a number of different sample adjustment techniques, each based on different models of the nature of the distinction between the sample and the population. These approaches include: 1) simply generating estimates from the data gathered, 2) weighting the data to address known or presumed differences, 3) accounting for differences in the process of building models of relationships, and 4) replicating analyses across samples with differing relationships to triangulate the range of potential answers that could emerge. The appropriateness and efficacy of these different approaches likely depends on the topic of research and the type of conclusion that researchers wish to make. The relative value of different methods is a subject of disagreement among scholars who study these issues. To illustrate the various approaches and considerations that underlie them, we briefly discuss a few considerations and possible strategies for dealing with them in the paragraphs below.

The case for adjusting or modeling for differences between the sample data collected and the target population is the strongest when there are sizable, well-understood differences between the two that are theorized to bias the estimand of interest. For instance, when scholars using a social media site have a random sample of posts and wish to estimate the prevalence of some posting behavior among the set of users who generated those posts, it is possible to use the number of posts per user to address the distinctions between estimates gathered from the posts and what would be observed had users instead been sampled. When the differences between data and target population are small, the importance of adjustment diminishes because unadjusted estimands are as likely to describe the population as adjusted ones. Alternatively, when large differences between data and target populations are poorly understood, the process of adjusting data can serve to introduce as much error as it can mitigate. And finally, if there is little reason to think that the differences between the sample and the target population will yield a biased estimate of a target parameter, then adjustments may not be helpful. This can occur either if the sample is a microcosm of the target population or if the estimand of interest is not sensitive to variations in the composition of the sample.

Historically, one of the most common approaches for adjusting for the differences between samples and target populations has been the use of weights, where weights are added to the sample to ensure that it better maps to the target population. For this kind of weighting, units representing undersampled populations are typically treated as more important in data analysis than units representing populations that are more prevalent in the sample. Weights used to adjust for these sorts of differences come in two types — those that are meant to invert the probability of selection/sampling relative to the target population (based on the sampling procedure and

assuming that the probability of selection is known) and those that are designed to adjust for "known" deviations from the target population on variables that are regarded as important covariates of both selection/sampling and the estimand (often demographic variables). Although the theoretical basis for adjusting for variability in the former is somewhat stronger than for the latter, both work off of a similar model, wherein the data collected from underrepresented populations is treated as a proxy for the data that could not be collected due to the non-presence of others in those populations. On a pragmatic level, this can cause issues because it introduces additional variance in the effort to reduce an unknown bias. Theoretically, it is also tenuous in that it relies on an untested assumption that present members of a group serve as a strong reflection of the group members who are not present. The validity of this assumption is often difficult to assess, but may be particularly problematic in the case of social media, where benchmarks for assessment are often poor and norms of online presence and posting behavior can often vary within and across groups.

Attempts to use weights to address discrepancies between the sampled population and the target population can run aground under a variety of circumstances. The process of producing weights can be error prone (e.g., if weighting is conducted using variables that are inferred about the sample rather than observed without accounting for prediction error; see Section 5.2), the assumption that the probability of missing data is conditional on measured covariates and nothing else is regularly violated, the inflated sampling variance introduced through the use of weights can swamp the bias that they are designed to correct, and the generation of statistical estimates based on weights produces substantively incorrect values (e.g., the standard errors for many inferential statistics are wrongly estimated). Furthermore, the weights are often treated as if they are values representing units in the data set rather than attributes designed to facilitate a translation between sample and population (Gelman, 2007).

Alternative approaches to weighting have been developed with the goal of allowing researchers to generate inferences about larger populations. These include attempts to assess the variability of estimates when similar estimators are drawn from different populations or platforms and the use of analytic techniques that aim to understand the implications of conditional variations in estimates attributable to variations in the composition of the sample drawn. Small area estimation, multilevel regression and poststratification (MRP), and various missing data imputation techniques have been employed to this end (Loux et al., 2019; Marchetti et al., 2015). Although there are a number of theoretical advantages to generating formal estimates of the sensitivity of conclusions across populations, the value of these approaches depends on an understanding of the parameters on which the sample and the population differ and the availability of auxiliary information for the sample and the larger population (Cornesse et al., 2020).

Whether through weighting or other techniques, the strongest case for adjusting data occurs when a given sample differs from a target population in relatively small, well understood, and systematic ways. Yet often, adjustments are based not on measured data, but on inferred data

— specifically presumed demographic attributes of social media users. This is likely the case for two interrelated reasons. First, researchers attempting to adjust social media estimands are often building off a long literature on statistical adjustment from survey methodology, which has largely relied on demographic variables as the basis for such adjustments. Second, such data are often available at the population level such that the discrepancies between the sample and population can be readily discerned. The availability of accurate benchmarks at the population level renders all forms of adjustment considerably easier (cf. Buskirk & Dutwin, 2017; Diaz et al., 2016). But their use will only aid estimation under two conditions: if the benchmarks that exist are measured similarly in the sample and in the population and if variable values are correlated with the estimand in ways that would help mitigate errors. In the survey context, the former is usually true by design, though considerable questions remain about the use of demographics for the latter. In the social media context, differences in measurement and utility can both undermine adjustment (Barberá & Rivero, 2015).

How should researchers make a decision about whether and how to adjust social media data that have been collected? The answer is almost certainly field-specific, should depend on the research question that is being asked, and is also sensitive to the availability of auxiliary information that can be used as part of that adjustment. For studies of interventions on social media, the only adjustments likely to be necessary are those that ensure that the unit of analysis for a study is in line with the target of the intervention -- that is via translations between levels of analysis such as posts, users, and groups. For case control studies, again the goal is not to generate any inferences beyond the population of analysis, so again such adjustments may not be particularly relevant.

Where the aim is to make inferences about the population of social media users, there are important questions about how observed posts and accounts relate to individuals that presumably need to be addressed. But a number of studies have raised questions about whether enough is known about the relationships between these to make solid inferences. In a series of studies attempting to model election results using social media data, for instance, scholars have found inconsistent relationships between expressed preferences on social media and electoral behaviors. Some of this is clearly attributable to individuals who are not present in social media. But even when estimates are compared by subsetting survey data to look like the kinds of people who use social media (Pasek & Dailey, 2019) or who post about the topic of interest on social media (Pasek et al., 2019), there are still irreconcilable differences. For example, Pasek et al. (2019) found that the strongest correlations between social media-derived trends and those derived from surveys existed for a vanishingly small subset of the population.

5.2 Inferring Demographic Attributes from Social Media Data

Studies that aim to infer estimates beyond the sample and particularly to a larger population using social media data face a challenge that much of the most common data used for translating between sample and population in other social sciences is less reliable in the social media context. In survey research, for instance, the most common strategy for understanding the

generalizability of a sample involves assessing both the extent to which the sample shares the same distributions of demographic variables with the population and the extent to which the statistical inferences a researcher wishes to make are related to those variables. For social media data that is not linked to some type of self-report data, however, this is a more challenging process because the availability and format of reported demographics on social media users vary greatly by platform and in many instances this information is not present. For example, many social media users do not explicitly report their age or gender. Thus, scholars may predict such demographic attributes for the user based on other information available on the users (e.g. names, pictures; see Mislove et al., 2011). The quality of those inferences, however, can be highly variable depending on the model, the ground truth methods, and the specific demographic characteristic being predicted.

There are various algorithms that attempt to ascertain characteristics of social media users as a function of profile information, posting behavior, or other indicators (see Hind and Joinson, 2018 for a systematic review). Despite a number of approaches to generate these sorts of data and sometimes effective algorithms to do so, there are reasons to worry about biases in both the set of data that can be classified, the ground truth used, and in the quality of the classifications that can be rendered. For instance, when algorithms misassess the age of an individual on social media, there is reason to worry that the misassessment may itself be related to other variables of interest (Wienberg, 2017). Because the amount of data available for making these inferences varies across individuals and because different groups of individuals have different norms for sharing information that can be used to make those inferences, these sorts of biases are commonplace (Freelon, 2018b).

Our ability to impute covariates that can be used for both understanding the conditional nature of our inferences and adjusting estimates based on social media data depends on the quality of the training data available as well as the coverage of the attributes that are used for training. For some covariates (e.g., education) it can be very challenging to make inferences due to the lack of good training data (see Ladd et al., 2020). Even for others for which there is more research (e.g. political ideology), accuracy is only around 85% on a two-category prediction task and coverage for the U.S. population may be as low as 30%, depending on the activity threshold the researcher chooses to decide when to impute (e.g., if the researcher is using text or connections to politicians or media outlets as the basis for inference). The relevant set of covariates to impute will also vary across research questions. For example, for inferences related to politics, factors like party identification, ideology, education are important. But for inferences related to consumer behavior, they may be less relevant. Hence, theory and not just empirics should be used to guide the choice of covariates.

Many common computer science tools for estimating covariates also fail to propagate uncertainty in user-level covariate inferences. That is, they often provide a deterministic “most likely” estimate for classifications or provide aggregate rather than individual-level assessments of accuracy. While there are a number of reasons for this, the two most common are their focus

on aggregate assessment of methods they are developing and concerns around privacy of the individuals used to build the models. This can be problematic for social scientific purposes as correspondence between inferred traits and criterion variables will be systematically weakened for groups where trait inferences are less accurate (Blackwell et al., 2017; Meng, 2018). Strategies such as introducing error to equalize accuracy across individuals or using multiple imputations can sometimes assist in these situations. For example, variables in the ultimate data set being prepared for analysis that get imputed/predicted as a function of other variables, e.g., the age of a social media user, could be imputed/predicted multiple times, in order to propagate the uncertainty in that prediction process. The ultimate data set would then include all M imputed values for a given variable in separate variables, so that multiple imputation analyses of the final data set can be performed. The point of this is to ensure that uncertainty in variables that are predicted is carefully accounted for in the subsequent analyses, using standard combining rules that have been developed for multiple imputation analyses (Little & Rubin, 2019).

The use of inferred demographic data can also raise ethical issues, especially because researchers might be capable of inferring private traits that are useful for inference but could also be employed for nefarious purposes (e.g. race/ethnicity, sexual orientation, religion, COVID-19 status, etc.). To date, IRBs have done little to scrutinize these issues, meaning that ethical lapses can sometimes bypass ethical review. Thus, researchers must consider the potential implications of their inferences and focus on developing a better set of standard practices moving forward. Similar issues about the implications of violating site terms of service, which often prohibit some types of inference, also need to be considered.

6. Data Linkage

Some of the previous sections mention the importance of linking social media data to other data sources, whether to assess the validity of the social media metrics, update assumptions, weight observations, or interpret conclusions. The difficulty of linking different forms of data with social media data and the value of doing so depends on the level of the linkage (macro vs. micro), the quality of the links that can be made, the sub-selection of the linked sample, and the precision with which those links describe the social media data generation process. At a highest level, demographic and geographic representation in social media data can be compared with high quality official population statistics. The largest and most timely sources for these comparisons in the U.S. are the Bureau of Labor Statistics' Current Population Survey (see U. S. Bureau of Labor Statistics, 2021) and the Census Bureau's American Community Survey (see U. S. Census Bureau, 2021). This involves linking at the population aggregate level to assess differences between the sample and another measure of the population. Yet, the relevance of such linkage to the research question has to be carefully assessed as certain social media populations do not represent the general population. Similar links can be drawn between social media data and data from trend surveys to assess the extent to which different sorts of samples capture similar trends in metrics over time (e.g. Diaz et al., 2016; O'Connor et al., 2010; Pasek et al., 2018; Pasek & Dailey, 2019).

A second type of data linking involves connecting aggregate-level information between social media data and another source. Links between social media and aggregate data can often be used to generate estimates of parameters that could not be derived without such links. For instance, by combining social media data with location-based data sets, scholars can test the impacts of things that happen on the local level on the data that are generated. These sorts of studies could help flag systematic issues in voting (Mebane et al., 2017), identify people who were displaced by a natural disaster (Wang & Taylor, 2014), or examine the link between Donald Trump's speech and the Capitol Riot on January 6, 2021 (Bajak et al., 2021). Social media data on members of particular social groups can also be compared to other sources of information about those groups. Although the value of such data can be enormous, it is important to recognize the potential for considerable biases when correlating individual attributes of group members with aggregate statistics for those same individuals. Group members who happen to exhibit a particular behavior on a particular social media site might be quite different from those members who are not on the site or whose site behavior is not observed. This means that social media aggregates may not reflect the group to which they are being compared.

In many cases, it is even possible to link social media data with other data from the same individuals. This could include links between various sorts of digital trace data, links to official records, or links to data derived from surveys. Depending on the features of the research, the proportion of respondents who have overlapping data across modes may be low and a moderate proportion may rebuff researchers' permission requests (Stier et al., 2020). Nonetheless, even when complete data cannot be obtained, these sorts of links can be fruitful. Moreover, to adhere to the ethical research practices, generating these sorts of links typically requires consent from the individuals who generate the data. Obtaining this consent is sometimes difficult, and can influence the conclusions that can be drawn if consenters are different from non-consenters on the measures of interest.

The literature on linking social media data to other sources of data at an individual level is starting to emerge (Henderson et al., 2019, Al Baghal et al., 2020, Mneimneh et al., 2020). Rates of consent to link typically range from 25% to 45% specifically among panel respondents. Consent rates among non-panel respondents can be lower. A number of individual-level factors, social/environmental factors, and study design factors could affect consent to link rates (Mneimneh et al., 2020). Not surprisingly, one of the most important predictors of consent to link is the individual's concern about privacy. Design level factors could include the mode of collecting the consent (telephone, face-to-face, or web), the placement of the consent request, as well as the consent phrasing. Future research that explores ways of maximizing consent rate, minimize linkage bias, while adhering to research ethical practices is needed to achieve individual level linked data sets that could provide insight into the quality of social media data and help improve prediction algorithms.

It is important to note however, that the process of linking between data streams can also be a source of error. Where links are made using group attributes, concerns about imputed data

quality can be important (see section 5.2). For individual-level links, additional challenges emerge in being sure that the links generated are accurate ones and in minimizing biases due to the set of people willing to consent to linking data. Links made on the basis of names can encounter challenges based on the relative density of similar names in some areas (cf. Hughes, et al., forthcoming), and can be highly dependent on both the accuracy of official records and the flexibility of matching software for generating links (Berent et al., 2016). And the accuracy of the linked data may itself be lacking (cf. Jürgens et al., 2020). Hence, linked data studies must be clear on exactly how links were generated and the resulting match rates, and these studies should ideally test the sensitivity of conclusions to minor variations in matching procedures.

7. Data Storage

Another key background decision for researchers involves how the data they collect should be stored. This too can have critical implications for analysis, adjustment, and replicability. When the samples are small and the collected data are public, using comma delimited files or spreadsheets may be reasonable approaches. As the data become larger and the number of attributes or data fields increases, organizing all data in a single rectangular matrix becomes inefficient and relational databases or other structured databases can be a good option.

Relational databases organize data into a set of tables containing rows (data records) and columns (data fields). The structure allows for easy identification of relationships that exist between data fields, i.e. mapping data fields in a course table to data fields in a student table, as well as efficient processing of large volumes of data. Well known examples of relational databases include Postgres, MySQL, and Oracle (Silberschatz et al., 1997). If the data being stored are documents, particularly longer ones, then using a database designed for text can be a good option. Examples include MongoDB and ElasticSearch (database and search engine). Finally, if data sets are very large, and possibly streaming, then using cloud-based distributed databases like BigQuery, Dynamo, HBase, or Databricks are good options. These databases are designed to distribute data optimally across a large number of servers and efficiently query them. Another popular option is to maintain text files in JSON or XML format and then use distributed and parallel compute infrastructures to process large volumes of the text files. Both JSON and XML formats are similar to comma separated files, but allow for more flexibility, including nesting of fields.

Acquiring and analyzing large scale social media data sets requires resources for large-scale data storage. Most social science researchers that use designed data are not trained on how to design and manage different types of databases, issues to consider when designing databases, or how to structure large data sets. This can be an impediment to research. One approach social scientists have taken is to build consortiums that work together on data collection and data storage issues. There are a number of NSF-funded initiatives and grants centered around making computing infrastructure more readily available for large-scale initiatives. One place to connect to those resources is visiting the NSF Big Data Innovation Hubs (2021). There are also universities and institutes working on access to different social media data sets for researchers at

their universities. For example, both Georgetown University and the University of Michigan provide access to Twitter data to their faculty. If social media data are going to be a significant part of a research portfolio, then having sufficient storage is important. Ultimately, companies will always have more storage capacity than universities, so finding ways to not replicate data sources is important for universities.

In addition to resources, data security is an essential feature of data storage. Survey researchers and experimental psychologists are used to keeping anonymized data sets under fairly light security, but data sets that include people's personal identifiable information (PII), such as contact information, are typically stored on highly secure computers, either with full encryption or air gapped. The typical standard is different in computer science. Computer scientists working with public social media data do not generally store data with different levels of security or encryption depending on the identifiability of users on the platforms. Instead, they keep all data at one specific security level, e.g. password protected. Which standards should apply to public social media data is an outstanding question. If the data are used in conjunction with survey data or other data sets containing PII (including other social media), then we recommend using the more stringent security level to ensure the protection coincides with the expectation of the research subjects. We also suggest storing both the original collected data (raw data), and the extracted data. It is important to archive the raw data to allow for future improved extraction techniques to be used on the data at a later time.

Finally, users sometimes choose to delete content or delete their accounts after the data have been collected for a research study. Because researchers do not typically monitor accounts on social media after collecting the data, they may not realize that some of the data have been removed. However, most user agreements with social media companies obligate us as researchers to also delete the data. Whenever a researcher uses API data, he/she has to go through a series of agreements that describes the requirements for collecting and storing social media from the specific platform. Unfortunately, researchers do not think about this issue because it is different from consented research. This issue is not just an ethical one, but also one that impacts replication. The same data removal standards should be applied to scraped data. We need to be consistent with user agreements about how to use the data, but then move beyond them as social scientists focus on specific questions and specialized groups. Since this is a constantly changing area, systematic and agreed upon rules still do not exist. It is important to begin establishing consistent ethical standards for use of different forms of social media data.

8. Data Preparation Considerations

One of the core areas of discussion throughout this white paper has been data management. While different fields define data management in different ways, we choose to define it broadly. We consider data management to encompass data collection, data storage, and data preparation. This section focuses on the last of the data management components, data preparation. How data are prepared depends upon the research question, the study design, the sampling strategy, and the data itself. What we focus on in this section are considerations that we

believe are among the most important or most complicated given the challenges associated with using social media data. The three main components that we discuss are data exclusion, data cleaning, and data aggregation. It is natural that these preparation steps depend on the type of analysis taking place. That is, does the analysis focus on whether a behavior or attitude exists, the prevalence of that behavior or attitude, the changing prevalence of that behavior or attitude, or how that behavior or attitude relates to other variables? Therefore, when applicable, we also discuss how the specific data processing component relates to the type of analysis.

8.1 Data Exclusion

An important part of processing any set of social media data for analysis is considering whether or not certain content should be excluded from the ultimate data set (depending on the context and the definition of the target population). Towards that end, we begin with a set of questions researchers should consider after data collection and before excluding data:

1) Might some of the data that are not currently relevant become relevant in the future? It is not always clear upfront which parts of the data will be important for machine learning models; therefore, if possible, maintaining data that may be relevant down the line makes sense.
2) How would the analysis change if some of the data are excluded? If it will not change, why exclude? If it will change, does it affect the interpretation of the conclusions?
3) Are there data that the researcher cannot analyze? For example, data that are too difficult to study with some methods may include text in foreign languages or video data. The difficulty may be a result of idiosyncrasies in the data that need to be controlled for or because the researcher just does not have the tools.
4) How much data are being excluded? If the vast majority of the data on a social media site is not deemed relevant, does this affect the assessment of whether the topic can be studied well with this specific social media data? Will there be more selection concerns? Does something like topic modeling need to be done to identify who should or should not be in the sample?
5) If the researcher thinks that some data are/were generated at random, would a control group account for this randomness without having to remove the data from the analyses? In other words, is the randomness in the dependent variable or independent variable?
6) What is the impact of bots or other non-personal accounts on the research study? If the research involves discussions, speakers, or content consumption, keeping these accounts in the study may be reasonable. But if the researcher does not want to study these accounts (e.g. the bots are in a disconnected network, or do not have any attitudes that can be measured), should the researcher remove identified accounts from the analyses or conduct robustness checks to assess whether removing them makes any difference at all? We note that removing bots or other non-personal accounts might be more important for certain topics where the researcher expects attempts to manipulate social media content (e.g. politics), particularly if the research uses counts of people. Also, distinctions

between various types of non-personal accounts (bots, cyborgs, corporate accounts, etc.) may be necessary for some research questions, though this may be difficult to determine without access to network information or additional user information that are sometimes unavailable to researchers. How would this distinction impact the study, if at all?

7) At a macro-level, when considering exclusion, how does the exclusion impact replicability, interpretability, and measurement error?
8) At a micro-level, what parts of a post are relevant to the research question? Researchers may want to consider excluding parts of posts that are less relevant to the research question, e.g. URLs or emojis, or users who are not active enough on the platform or have multiple accounts.

In general, whenever data exclusions are made, the decision to exclude them need to be explained and documented. Otherwise, the rationale and history of the different exclusions will be lost over time. As an example, a typical set of data exclusions would include advertisements, languages the researcher cannot process, geographic restrictions, spam (for post-level analysis), bots through using a tool like *Botometer* or *Russian trolls*, or by assigning a probability of an account being a bot, and data that have been removed by platform users.

Another approach for thinking about data exclusion could be to construct variables that would serve as indicators of relevance or authenticity. One way to accomplish this is to select a random sample (possibly stratified in some fashion) of units (posts, tweets, etc.), and code them (in terms of predictors, indicators of interest, etc.) for the training of a classifier. The size of this random sample would be constrained by the available budget for data preparation and the number of data points necessary from each stratum in order to build a reliable classifier. For example, one could have a study that involves a network analysis at the user level. Such analysis might first start by estimating the probability that any given user account is a bot (and then this probability would be assigned to all units, such as posts, associated with that user). Or, one could code the posts for relevance, code other auxiliary information, and then develop a classifier for predicted probability of relevance. Once a sample data set has been coded, standard cross-validation techniques could be applied to develop an effective classifier. This classifier could then be applied to a larger sample when creating the ultimate data set for analysis. The resulting probabilities (of relevance/authenticity/etc.) need not be used to make ultimate decisions about dropping units of analysis (e.g., posts), but might instead serve as covariates, stratifiers, or importance indicators for the substantive data analysis.

8.2 Data Cleaning

Most data sets require some level of cleaning that depends on the type of study or model used. For example, if researchers are interested in the most frequent words used in posts about Black Lives Matter, removing common language stop words would constitute an important cleaning step. Although there are a number of different data cleaning best practices, we lack a common definition for data cleaning for social media data that is applicable irrespective of the context or data type (e.g., image or text). Therefore, it is critical for papers to include a

description of the data cleaning steps and, when possible, provide the code used to clean the data. In general, all code used to pre-process and clean data should be provided alongside any data set that eventually gets used for analysis. This is vital for transparency and reproducibility.

Though a large number of data cleaning steps have been employed for different types of data, we highlight the ones that are most applicable to social media data. First, it is important to construct clear metadata about the data set, including understandable variable names and consistent value labels, item-missing data rates for individual variables, methods used to impute the missing values (if any), and a list of cleaning steps. There are groups such as the Inter-university Consortium for Political and Social Research (ICPSR) that have developed standards for certain types of data. Adapting existing standards for social media data is an important step.

Many view data cleaning as more of an art than a science. Each project may need data in a different format, especially when examining text or images. Computer scientists tend to separate data and their cleaning from their use or the applications that use the data. But it is unclear whether such an approach is possible or desirable for social media data. For researchers who have available data storage, many different "approximations" of the original text or image data can be created and stored for use by different studies, along with the original unprocessed data. As an example, one approximation for text may be to convert it to lowercase and to remove punctuation. Another approximation may be to create a bag of words model of the text (Manning, 2008). Another advantage of this approach is that some data cleaning is very expensive. For example, on a large corpus, it is cheaper to not include lemmatization or stemming during pre-processing. Therefore, implementing this type of data cleaning on a large corpus before knowing that it is needed for analyses may not be cost effective; instead data collectors might prefer to have these steps taken on demand.

Here we list some different types of data cleaning that have been used for social media data. For text, these include: lowercasing text, stop word removal (i.e., common words that are not content rich like "the" or "and"), flood word removal (i.e., non-traditional stop words specific to a domain), abbreviation expansion, part of speech tagging of words in a sentence, stemming/lemmatization of words based on sentence structure (e.g. lemmatizing the word "ran" to "run"), emoji coding, punctuation removal or coding, and URL removal or coding. Some difficulties that exist with social media text data include posts in multiple languages, the encoding used, whitespace inconsistencies, misspellings, and poor grammar usage. For images, standardization steps include ensuring a minimum resolution of each image, object extraction, and repixelizing/gray scaling. Ultimately, it can be difficult to determine what "noise" is in the data a priori. The noise may be very task specific. For example, noise in the raw data may look different for clustering, topic modeling, and prediction tasks. This is another reason for keeping multiple approximations when possible.

To what extent does data cleaning affect the analyses? Ultimately, we do not know the answer to this question for most analyses of interest, and even conceptualizing this is difficult if there are many considerations and decision points involved. If we imagine a tree with all of the

decision points as branches, we might consider how many of those branches reach the same conclusion as the one reported. If the one reported is different from others, has the paper thoroughly explained the data cleaning choices made to rule out the branches that reached different conclusions? Given the sheer number of potential permutations involved, it would be helpful for social scientists to know when the performance gains from some data cleaning steps are unlikely to affect hypothesis tests. What robustness checks would help determine that? How do we distinguish between measurements that are possible with a cleaned data set compared to an idiosyncratic result attributable to arbitrary cleaning choices? These questions need to be explored in depth to understand the relationship between cleaning, measurement quality, and impact on researcher questions.

8.3 Aggregation

There are many motivations for aggregating data. First, it may be necessary to protect the privacy of individuals or groups in the study. Second, a researcher may have data that are too narrowly focused (e.g. per post or per day) and may need to reduce the granularity of the data to compare the groups of interest or to complete the analysis given computational resource constraints. Third, it may be the case that the detailed data are too sparse and aggregation is necessary in order to answer the research question. For example, a computational linguist may be interested in understanding language change. Daily data may not have enough examples and it may also be too small a unit of time to tackle the research question so weekly aggregates may be in order. Next, a researcher may have data at different resolutions that need to be combined for the analysis. In this case, the researcher will need to harmonize data by aggregating some of the data to the level of other data. Finally, there may be missing data at one level of resolution, but not at the aggregate level. For example, a data set may include city, or county or zip code, but not necessarily all of these geographical identifiers for a single individual. In this case, aggregating to state allows for a consistent analysis that uses all the research subjects. All of these motivations for aggregating data apply to social media data.

Social media data are richly layered. There are a number of classic hierarchical structures for aggregation that are natural fits for social media data, including user, time, platform, geographies, topics, and also those based on personal characteristics (e.g., age). For example, the user aggregation hierarchy would include levels for the post, the user, the community, and the population. When we consider time, that easily aggregates from second to minute to day to month to year.

When researchers consider aggregation, there are a number of questions they need to answer. First, does their research question require a single level of analysis or multiple levels? Then they can decide which levels are most relevant given the unit of analysis, the unit of observation, and the target population. Next, they need to consider if any new bias is being introduced by aggregating. For example, sometimes researchers aggregate because they do not have the detailed level of data on all units. Researchers also need to consider what is lost with aggregation and whether that matters. For example, did rare events or actions get lost? Does

micro-level missingness impact the research or the conclusions?

Social media data are typically very long-tailed. For many analyses that social scientists conduct, researchers might study posts when their hypotheses and research questions are focused on individuals. This could come from a lack of awareness about the biases that can be introduced by disproportionate posting behavior in social media data or from the fact that sometimes the relevant covariates to correct for these issues are unavailable. This choice of users vs. posts can be extremely consequential. Studying posts might represent the behaviors and conversations of only a small fraction percentage of users, even in samples with thousands or millions of data points (Wojcik & Hughes, 2019). Understanding the relationship between aggregation and the underlying data distribution is important for rendering accurate conclusions. One approach to deal with long-tailed data is to build a hierarchy of clusters as units of analysis: For example, individual posts are nested within time periods, posts are also nested within users, and users are nested within geographies. Social scientists often use multilevel models to both decompose and explain variability in a given outcome measured at the lowest level of the data (e.g., a sentiment expressed in a post) across the different clusters. In this sense, data aggregation considerations should be carefully linked to the ultimate analytic goals of the researcher. This approach of multi-level extraction and computation should be balanced against the available resources and power. Thinking of aggregation as a multilevel model and then assessing which parts of the hierarchy are computationally feasible is one approach for making this decision. Such decision processes need to be documented and shared to guide future directions of aggregations.

Finally, it would be informative to create a taxonomy of different types of possible aggregations and then let researchers map their research questions to those approaches. The decision to choose a particular level is not just connected to the research question, but also to the availability of the data and the amount of processing needed to build and analyze the different aggregation levels. In computer science, aggregation is a fundamental operation of most database systems (at least data can be rolled up to different temporal and spatial levels). How do we transfer this type of operation to aggregate social media data properly? Given all the missing data on social media, we would need to develop rules for default approaches for filling in the gaps in data as we aggregate, perhaps including new metrics of reliability and potential bias.

9. Ethical Considerations

Ethical considerations in the use of social media data should be considered at all study phases from the design of consent language, the point of data collection, the process of data dissemination, and the future use of the disseminated data. Numerous researchers have written about the ethical considerations of using social media data (Chancellor et al., 2019; Benton et al., 2017; O'Connor, 2013; Lane et al., 2014; Zimmer, 2010; Sloan et al., 2020; Singh et al., 2020b) given the differences in the nature of social media data relative to the traditional design data. We expect this area to keep evolving as new platforms emerge, service agreements change, and novel study designs are introduced. In this section, we highlight some of the main considerations

that were raised during the workshop and add to the body of literature on IRB requirements, consent, and data dissemination of social media data.

9.1 IRB

Historically, public data which comprised de-identified databases of human subject data that are available to the public were exempt from IRB review. This notion has led many researchers using social media data to argue that, in general, using any kind of public data does not require IRB approval. Yet, in reality, many forms of publicly available social media data include identifiable personal information and can be used to estimate attributes of individuals that the ethical review process was designed to prevent. This raises the question of whether and when this type of data should be subject to various levels of ethical scrutiny. Indeed, the availability of human subject social media data that are rich and voluminous have opened the door for researchers from a diverse set of disciplines to start using these data for what had been more carefully regulated human subject research. Yet, disciplines vary in their training and in their perceptions of the requirement of seeking IRB approvals. For example, many computer scientists do not seek IRB approval for social media research. In contrast, psychologists, sociologists and survey researchers are used to seeking IRB approval for all research that involves human subject data. We recommend that researchers first consider the public nature of the data and make sure that they fully understand what pieces of the data are public and whether any identifying information might be collected in the process of gathering the data. Second, researchers from any discipline who are using any public human subject data that are not already anonymized would be well served to seek IRB approval. In many instances, such approval would be exempt from full IRB review as long as there is no direct interaction with the subjects and the research conducted does not cause additional risk or harm to the subjects who are sharing the information publicly. This type of consistency in seeking IRB approval would enhance cross-university and cross-discipline collaborations in the use of social media data.

9.2 Informed Consent

Conditional on an IRB approval, the second ethical consideration when using human subject social media data is whether informed consent needs to be garnered from the users generating the data. This consideration needs to take into account the specific social media platform, as platforms differ on the amount and type of public data shared and the purpose of sharing the data. For example, Yelp restaurant reviews are designed to be public and used to help the public with restaurant selection; thus if such reviews are used for restaurant preferences, consent to use the reviews may not be needed as the data usage is governed by the platform. In contrast, SnapChat is designed for private communications. Therefore, any analysis of SnapChat would need user consent. In reality, most sites fall somewhere in between, with user privacy often protected in practice by the relatively limited dissemination of what is technically a public post. Thus, providing a framework on when informed consent is needed would advance the systematic proper use of social media data for human subject research.

We list here one possible general hierarchy that could contribute to establishing such a framework. We think that consent is needed when: 1) a password was provided in order to get access to the data, 2) there is a direct communication between the researcher and the subject, 3) specific subgroups (including vulnerable populations) are studied or highly sensitive information is gathered, and 4) when social media data are linked to another sources of data (be it another social media data, survey data, administrative data, etc.). Data linkage usually provides additional information about the user that might allow the researcher to learn or predict certain characteristics that the user did not intend to share publicly. It is like putting pieces of the puzzle together and now having a clear picture of the whole. In all these cases, there is a reasonable chance that people believe their data are either private or will not be combined with other data. This is why we believe it is important to obtain consent in these cases. It is also important to note that the consent language presented to the user should have all the components of the consent language typically required by IRBs in addition to new components that might be unique to social media data such as the duration of the data collection (i.e. for how long in the future will the research continue to gather such public data).

The research scenario that triggers the most debate and where requiring consent is still controversial is the use of large volumes of public social media data that require no communication with the user and that do not link to other sources such as when using APIs or scraping social media sites to collect publicly available information. Given the amount of data and the possibility that the researcher does not have a direct way to contact the respondent, it might not be feasible to obtain informed consent from every subject (O'Connor, 2013). In this situation, Benton and colleagues (2017) recommend issuing a statement of responsibility on the research group website and any publication that uses the publicly collected data. Benton and colleagues (2017) specify that the statement should include "a description of the type of data that are collected, how they are being protected, and the types of analyses that will be conducted using it", thereby providing the researcher with a measure of accountability. Site terms of service and user agreements may also play a role in determining whether these uses are acceptable without explicit consent.

9.3 Data Aggregation and Dissemination

In general, researchers should not be making any identifiable data public, even if these data were collected from a public platform. This includes sharing verbatim content from posts that could identify the user since the context around the post is typically removed and many users may not fully understand the public nature of their posts. In this situation, we recommend aggregating the data before sharing it publicly. The aggregated data should also be assessed for the potential to identify individuals especially when the aggregation is done on rare groups or on very specific topics.

Furthermore, since many researchers are using public data to predict and construct new variables, these constructed variables need to be fully scrutinized for their potential to reconstruct the identity of the user before the variables are shared. If such constructed or aggregate data

provide researchers with access to any potentially identifiable data, researchers are greatly encouraged to protect that data through the use of models that obscure identification or by limiting use to enclaves that allow sharing these sorts of data for research purposes. The federal statistical system (and Census Bureau in particular) has decades of experience providing researchers access to confidential, respondent-level data using data enclaves that allow analysis but prevent researchers from accessing individual-level records. Many institutions also have special enclaves for sharing data. Researchers are encouraged to see if data for studies of interest already exist in enclaves as it saves researchers time and helps to justify building these government and academic data stores.

10. Accelerating Research in This Area

Throughout this white paper, we have articulated a series of considerations that researchers should take into account as they utilize data from social media sources. Across these considerations, it is clear that much remains to be done to better understand the data generating process that leads to the data sets used in this domain, to describe the procedures employed in data collection and processing, and to recognize the implications of those choices for any conclusions that may be drawn. In this section, we articulate a series of concrete steps that can be undertaken by individual researchers, funders, and the community of scholars to improve the application of social media data to social scientific questions as data science, computer science, and the social sciences converge in this area. We also discuss some tools that already exist as a starting point toward these objectives.

10.1 Understanding the Data Generating Process

One of the central factors driving questions about the representativeness of the data and the types of conclusions that can be reached is how the data we observe on social media sites are generated. This is a function of understanding who uses particular platforms and how those users do so. To date, there have been a handful of studies that attempt to interrogate the demographic profiles of users as compared to the larger public (e.g., Pew Research, 2018) or that link social media data with survey data from the same set of users (e.g. Guess et al., 2019; Haenschen, 2020). Our understanding of the inferences that could be made from social media data sets, however, would be enormously aided by additional work in this area. We still need to establish basic statistics about what kinds of individuals use particular platforms (both demographic and psychographic), how different kinds of users differ in their behaviors on those platforms and offline, and how those processes vary over time. Much of this can be determined using large-scale panels that allow scholars to link social media data with other data sets at the individual level and to track those individuals over time. Funding and collaborative mechanisms that bring together scholars to produce these sorts of data sets would help establish benchmarks for the field.

Integrations with social media companies can also be important in providing a lens on how key parts of the data generating and collection process may shape the data researchers can

access. Social media companies do not typically provide clear information about public vs. private posting behavior, the differences between hidden and active users, or how exactly various APIs determine what information to provide in response to queries. These sorts of information would be helpful for diagnosing why different methods reach different conclusions when they do. Agreements with industry might also help to fill in additional covariates about individual users if mechanisms can be put in place to ensure data security and address privacy concerns. Currently, these sorts of access are irregularly available, with some well-connected researchers able to access data that others cannot. Challenges introduced by corporate restrictions on how that data can be used furthers questions of replicability, equity, and ethics. Encouraging greater transparency from social media companies and facilitating collaborations between industry and scholars would aid in these objectives.

10.2 Streamlining Collection and Processing Procedures

Although many tools exist for collecting, adjusting, and preprocessing social media data, too little has been done to consolidate and organize the tools that are available and to streamline their collective use. Individual researchers have been left to piece together approaches that differ depending on training and awareness. Further, guidelines for some procedures have not been well established. It is therefore critical to establish standards for identifying the tools that exist to gather, adjust, process, and link different types of data, as well as to provide repositories where those tools can be identified and considered by researchers. The ability to identify example use cases for these tools will also facilitate innovation and allow researchers to assess the impact of sampling and processing methods on research outcomes, to determine the relative robustness of research conclusions, and to identify where differing decisions across studies may be responsible for different results across studies.

10.3 Identifying the Importance of Choices for Outcomes

As scholars consider the robustness of their findings and work toward reproducible research, they need to articulate the choices that they made in data collection, describe the decision processes that yielded those choices, and consider how sensitive their findings are to alternative approaches. Among the first objectives in this area is providing a clear articulation of the set of processes that researchers could engage in, so that they can more clearly describe what occurred at each data acquisition phase. In addition, scholars need to develop tools for testing the sensitivity of results to the methodological choices they make. They also need to articulate standards for assessing the impact of methodological choices on analytical outcomes. Although this seems like a relatively straightforward task, the number of permutations of methods that are possible can quickly make it impossible to completely explicate the parameter space. For this reason, it would be helpful to establish statistical methods to assess the relative sensitivity of particular choices across arbitrary combinations of parameters to identify which choices are likely to have a substantial impact and which are likely to be de minimis. These types of tools exist for assessing the sensitivity of machine learning models and feature selection (see Pedregosa et al., 2011 as an example), but need to be expanded to consider sampling and data

preparation decisions. New tools should also be developed to turn these sorts of estimates into formal assessments of the robustness of results to alternative choices and specifications.

10.4 Shared Data Resources

We have mentioned a few times the importance of shared data resources for social media data given their scale and interest for social science research. Throughout the meeting, different shared resources were identified. We have begun to create a list of shared resources that is available on our website (http://smrconverge.org/resources/). We hope the community will continue to add to those. We highlight a few here that were mentioned multiple times throughout the meeting.

In terms of public, available data sets, archive.org shares the Twitter "sprinkle" data set which is the 1% sample of their posts. Facebook shares some of their group level and URL data through CrowdTangle (2021). They are also giving academics access to data related to elections, Black Lives Matter, and other issues of the day. All Reddit data is publicly available for download using the Pushshift.io API (2021). And Google has their Google Trends interface and API that release time series search intensities for any search keywords (Google, 2021). These are just a handful of examples, but are good starting points for those less familiar with using social media data.

Over the last decade, there have been a number of papers related to creating data archives and data enclaves (cf. Hemphill et al., 2021). There are two main types of data enclaves which store confidential data: those that allow for secure network access to the confidential data, and those that reside on physical machines that are not networked, where researchers are required to access the data from a monitored room. A number of enclaves have been developed around healthcare, e,g, the Coleridge Initiative (2021), and the UMD/NORC Mental Health Data Enclave (2021).

Finally, there are also a number of code-bases and libraries for collecting and preparing social media data. Pushshift.io has data and an easy to use Python library for accessing Reddit data, StackExchange data, and a limited amount of Gab data. The code masks away the complexity of data collection and parsing. Tweepy (Roesslein, 2020) and PythonTwitter (The Python-Twitter Developers, 2019) are Python programs for collecting Twitter data. PythonYouTube is a package for collecting YouTube data (Kun, 2020).

While this list of shared resources is short, it highlights the types of resources researchers and companies are developing. It is also a reminder of the need for resources for a broader set of social media platforms.

11. Conclusion

The convergence of theories and methods used by computer scientists and social scientists is a necessary condition for leveraging social media data to understand this increasingly important window into human societies. How researchers approach this task as they

gather and prepare their data, however, has critical implications for the hurdles they will face in data management, as well as for the types of conclusions they can reach. To that end, this white paper focuses on issues related to data acquisition, sampling, adjustment, storage, and preparation, including exclusions, data cleaning and data aggregation.

As we articulate throughout the white paper, a series of considerations guide our ability to answer different questions with the data that have been gathered, as well as the strength of the conclusions they generate and the transferability of those conclusions across time, populations, and domains. One consideration involves the extent to which researchers are conducting hypothesis-driven versus data-driven research. When researchers start with a specific hypothesis, collect data narrowly targeted to that hypothesis, and analyze that data, the results provide a robust test of the theory they have proposed. Either the data provides evidence for the hypothesis, or it does not. In contrast, when researchers examine the patterns that emerge from data collection, their conclusions can be more difficult to interpret. The likelihood that some relations they observe are artifactual is higher because the number of correlations tested is itself far larger. Some of the correlations that are observed by chance will be due to idiosyncrasies in the data or will be endogenous to factors that were not considered.

Another key consideration involves decisions related to accessing data. Although a more comprehensive and scientifically rigorous approach could be used when the researcher can acquire the data directly from the complete universe of units by collaborating with social media platforms, the majority of the research that has been produced using social media data rely on acquiring the data through web scraping and APIs. Researchers comparing these approaches should consider the necessary level of knowledge and skills, the changing nature and structure of social media data, service agreements set by the platforms, and the computing power needed. Their choices have implications for the results they obtain because these different strategies do not have equal likelihoods of obtaining the same data, and each can yield biases samples in different ways.

Researchers' objectives as to whether they are hoping to make generalizable conclusions or instead focus on describing the data obtained have implications for a variety of processes, ranging from the choice of methods for gathering data to the extent to which the data they gather needs to be subjected to various modeling strategies in order to reach desired conclusions. Some data collection strategies are only capable of answering descriptive questions whereas others are better suited for inferential research. In some cases, strategies akin to survey weighting and small area estimation can help to bridge these differences. In others, linking across data sources can provide the basis for inferences. While several statistical adjustment methods exist to account for potential sampling biases introduced through the selection process, the application of these methods to social media data is still questionable as many of them rely on collecting a number of covariates or auxiliary information that are related to the outcome of interest and the availability of population level summary statistics.

Challenges associated with data management can similarly complicate research. Social media data sets can often become unwieldy for researchers who are accustomed to working with rectangular data sets that are inefficient for sparse data or who are not familiar with the cloud storage and computing systems that are often necessary for analysis at scale. Data preparation steps also can have important implications for conclusions, yet are not currently well standardized across the different fields conducting social media research. Similar variation exists in how ethical standards are applied across fields.

In reviewing the state of the field, we believe that a number of concrete steps need to be taken to foster future research in this area. In addition to identifying a series of concrete efforts that should be pursued in this domain, articulating the considerations for researchers in this area highlights a need for better documentation and transparency about the series of decisions that researchers make and the likely impact of those decisions on conclusions. This not only takes the form both of clearer descriptions of the product and rationale for each of the considerations we identify, but also calls for sensitivity analyses to assess formally how variations in many of those choices might influence results.

The importance of documentation, dissemination of codes, and robustness checks also applies for data cleaning and aggregation decisions. Given the layered nature of social media data, the amount and varied nature of missing data, and the potential identifying information present, some form of data aggregation seems to be needed. Decisions for data aggregation need to be guided by the research question, as well as available resources as this process can be very labor intensive. A useful direction would be 1) creating a mapping of different types of possible aggregation, allowing researchers to map their research question to those levels that are relevant, and 2) developing rules for approaches to imputing the missingness in the data as aggregation decisions are made.

The importance of contextualizing these considerations is particularly pronounced given the ever-changing nature of social media data and the evolving toolkit researchers have to collect, prepare, and examine them. Absent the formal approach to considering these choices that we embark upon here, there is reason to worry that the fields that engage with these data may reach incommensurate conclusions due to different norms of data collection and preparation. As we continue to update the menu of options underlying each of these decisions, we need an agreed upon system for articulating what methods were chosen to remain simultaneously flexible and clear. We will also need to maintain the interdisciplinary dialogue begun in these meetings to ensure that the processes we consider continue to define the scope of social media research.

Acknowledgements

We would like to thank the National Science Foundation and the McCourt School's Massive Data Institute (MDI) at Georgetown University for supporting this collaborative meeting. This white paper is an output of that meeting and is co-authored by those who participated in the meeting. This series of white papers is funded by National Science Foundation awards #1934925 and #1934494 as a collaboration between Georgetown University and University of Michigan. A special thanks to Rebecca Vanarsdall from MDI for her support in planning the meeting and helping prepare this white paper. We also want to thank all the students and staff who helped facilitate and take notes during the meeting. Those notes were invaluable when putting this document together.

References

Al Baghal, T., Sloan, L., Jessop, C., Williams, M. L., & Burnap, P. (2020). Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. *Social Science Computer Review, 38*(5), 517-532.

Andridge, R. R., West, B. T., Little, R. J., Boonstra, P. S., & Alvarado-Leiton, F. (2019). Indices of non-ignorable selection bias for proportions estimated from non-probability samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 68*(5), 1465-1483.

Bajak, A., Guynn, J., & Thorson, M. (2021, February 1). When Trump started his speech before the Capitol riot, talk on Parler turned to civil war. *USA Today.* Retrieved from https://www.usatoday.com/in-depth/news/2021/02/01/civil-war-during-trumps-pre-riot-speech-parler-talk-grew-darker/4297165001/

Barberá, P., & Rivero, G. (2015). Understanding the political representativeness of Twitter users. *Social Science Computer Review, 33*(6), 712-729.

Beauchamp, N. (2017). Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science, 61*(2), 490-503.

Benton, A., Coppersmith, G., & Dredze, M. (2017). Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing* (pp. 94–102). Valencia, Spain.

Berent, M. K., Krosnick, J. A., & Lupia, A. (2016). Measuring voter registration and turnout in surveys: Do official government records yield more accurate assessments?. *Public Opinion Quarterly, 80*(3), 597-621.

Berzofsky, M. E., McKay, T., Hsieh, Y. P., & Smith, A. (2018). Probability-based samples on Twitter: Methodology and application. *Survey Practice, 11*(2), 1-12.

The Big Data Innovation Hubs (2021). The Big Data Innovation Hubs. Retrieved from https://bigdatahubs.org/

Blackwell, M., Honaker, J., & King, G. (2017). A unified approach to measurement error and missing data: overview and applications. *Sociological Methods & Research, 46*(3), 303-341.

Bode, L., Davis-Kean, P., Singh, L., Berger-Wolf, T., Budak, C., Chi, G., Guess, A., Hill, J., Hughes, A., Jensen, J. B., Kreuter, F., Ladd, J. M., Little, M., Mneimneh, Z., Munger, K., Pasek, J., Raghunathan, T., Ryan, R., Soroka, S., & Traugott, M. (2020). Study Designs for Quantitative Social Science Research Using Social Media. PsyArXiv: https://psyarxiv.com/zp8q2/

Brandwatch. (2021). Brandwatch. Retrieved from https://www.brandwatch.com/

Baumgartner, J. M. (2019). Reconstructing Twitter's Firehose. Retrieved from https://docs.google.com/document/d/1xVrPoNutyqTdQ04DXBEZW4ZW4A5RAQW2he7qIpTmG-M/edit?usp=sharing

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81.

Chancellor, S., Birnbaum, M. L., Caine, E. D., Silenzio, V., & De Choudhury, M. (2019). A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 79–88). Atlanta, GA, USA.

Chen, X., Wang, Y., Agichtein, E. & Wang, F. (2015). A comparative study of demographic attribute inference in Twitter. In *Proceedings of AAAI Conference on Weblogs and Social Media*. (pp. 590-593). Oxford, England, UK.

Coleridge Initiative. (2021). Coleridge Initiative. Retrieved from https://coleridgeinitiative.org/

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., ... & Wenz, A. (2020). A review of conceptual approaches and empirical evidence on probability and nonprobability sample survey research. *Journal of Survey Statistics and Methodology, 8*(1), 4-36.

CrowdTangle. (2021). CrowdTangle. Retrieved from https://www.crowdtangle.com/

De Choudhury, M., Gamon, M., Counts, S., Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the International Conference on Weblogs and Social Media.* (pp. 128-137). Boston, MA, USA.

Diaz, F., Gamon, M., Hofman, J. M., Kıcıman, E., & Rothschild, D. (2016). Online and social media data as an imperfect continuous panel survey. *PloS one*, *11*(1), e0145406.

Druckman, J. N., & Kam, C. D. (2011). Students as Experimental Participants: A Defense of the 'Narrow Data Base. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge Handbook of Experimental Political Science.* Cambridge University Press.

Dutwin, D., & Buskirk, T. D. (2017). Apples to oranges or gala versus golden delicious? Comparing data quality of nonprobability internet samples to low response rate probability samples. *Public Opinion Quarterly*, *81*(S1), 213-239.

Franzke, A. S., Bechmann, A., Zimmer, M., Ess, C., & the Association of Internet Researchers (2020). Internet Research: Ethical Guidelines 3.0. https://aoir.org/reports/ethics3.pdf

Freelon, D. (2018a). Computational research in the post-API age. *Political Communication, 35(*4), 665-668.

Freelon, D. (2018b). Inferring individual-level characteristics from digital trace data: Issues and recommendations. In *Digital Discussions* (pp. 96-110). Routledge.

Gelman, A., & Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology, 23,* 127–135.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science, 22*(2), 153-164.

Google (2021). Google Trends. Retrieved from https://trends.google.com/trends/?geo=US

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science, 363*(6425), 374-378.

Guess, A., Munger, K., Nagler, J., & Tucker, J. (2018). How accurate are survey responses on social media and politics?. *Political Communication, 36*(2), 241-258.

Haenschen, K. (2020). Self-reported versus digitally recorded: Measuring political activity on Facebook. *Social Science Computer Review, 38*(5), 567-583.

Harter, R., Battaglia, M. P., Buskirk, T. D., Dillman, D. A., English, N., Fahimi, M., Frankel, M. R., Kennel, T., McMichael, J. P., McPhee, C. B., Montaquila, J., Yancey, T., & Zukerberg, A. L. (2016). Address-Based Sampling. *AAPOR Task Force on Address-based Sampling.*

Harvard Dataverse. (2021). Harvard dataverse. Retrieved from https://dataverse.harvard.edu/

Harvard's Institute for Quantitative Social Science (2021). Social science one. Retrieved from https://socialscience.one/contact-us

Hemphill, L., Hedstrom, M. L., & Leonard, S. H. (2021). Saving social media data: Understanding data management practices among social media researchers and their implications for archives. *Journal of the Association for Information Science and Technology, 72*(1), 97-109.

Henderson, M., Jiang, K., Johnson, M., & Porter, L. (2019). Measuring Twitter use: Validating survey-based measures. *Social Science Computer Review,* 0894439319896244.

HiQ Incorporated v. LinkedIn Corporation, D.C. No. 3:17-cv-03301-EMC (2019). https://www.eff.org/document/hiq-v-linkedin-ninth-circuit-decision

Hughes, A., McCabe, S., Hobbs, W., Remy, E., Shah, S., & Lazer D. (forthcoming). Using administrative records and survey data to construct samples of Tweeters and Tweets. *Public Opinion Quarterly*.

IETF OAuth Working Group. (2021). OAuth 2.0. Retrieved from https://oauth.net/2/

Jungherr, A. (2015). *Analyzing political communication with digital trace data.* Cham, Switzerland: Springer.

Jürgens, P., Stark, B., & Magin, M. (2020). Two half-truths make a whole? On bias in self-reports and tracking data. *Social Science Computer Review, 38*(5), 600-615.

Krupnikov, Y., & Levine, A. S. (2014). Cross-sample comparisons and external validity. *Journal of Experimental Political Science*, *1*(1), 59.

Kun, I. (2020). Python-YouTube 0.6.4. Retrieved from: https://pypi.org/project/python-youtube/

Ladd, J., Ryan, R., Singh, L., Bode, L., Budak, C., Conrad, F., Cooksey, E., Davis-Kean, P., Dworak-Fisher, K., Freelon, D., Hopkins, D., Jensen, J. B., Kelley, K., Miller, R., Mneimneh, Z., Pasek, J., Raghunathan, T., Gresenz, C. R., Roy, S., Soroka, S., & Traugott, M. (2020). Measurement Considerations for Quantitative Social Science Research Using Social Media Data. [White paper]. PsyArXiv.

Lane, J., Stodden, V., Bender, S., & Nissenbaum, H. (Eds.). (2014). *Privacy, big data, and the public good: Frameworks for engagement.* Cambridge University Press.

Lazer, D. M., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., ... & Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science, 369*(6507), 1060-1062.

Lepkowski, J. M., Tucker, C., & Brick, J. M. (2008). *Advances in telephone survey methodology.* E. de Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, & R. L. Sangster (Eds.). Hoboken (NJ): John Wiley & Sons.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data (Vol. 793*). John Wiley & Sons.

Little, R. J., West, B. T., Boonstra, P. S., & Hu, J. (2020). Measures of the degree of departure from ignorable sample selection. *Journal of Survey Statistics and Methodology*, *8*(5), 932-964.

Loux, T., Nelson, E. J., Arnold, L. D., Shacham, E., & Schootman, M. (2019). Using multilevel regression with poststratification to obtain regional health estimates from a Facebook-recruited sample. *Annals of Epidemiology, 39,* 15-20.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 39, pp. 234-265). Cambridge: Cambridge University Press.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., & Gabrielli, L. (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics, 31*(2), 263-281.

Mebane Jr, W. R., Pineda, A., Woods, L., Klaver, J., Wu, P., & Miller, B. (2017). Using Twitter to observe election incidents in the United States. In *Annual Meeting of the Midwest Political Science Association.* (pp. 1-64). Chicago, IL, USA.

Meng, X. L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics, 12*(2), 685-726.

Mislove, A., Lehmann, S., Ahn, Y. Y., Onnela, J. P., & Rosenquist, J. (2011). Understanding the demographics of Twitter users. In *Proceedings of the International AAAI Conference on Web and Social Media.* (pp. 554-557). Barcelona, Spain.

Mneimneh, Z. N., McClain, C., Bruffaerts, R., & Altwaijri, Y. A. (2020). Evaluating survey consent to social media linkage in three international health surveys. *Research in Social and Administrative Pharmacy*.

Moore, W. B., Wei, Y., Orshefsky, A., Sherr, M., Singh, L., & Yang, H. (2013). Understanding site-based inference potential for identifying hidden attributes. In *Proceedings of the IEEE International Conference on Social Computing (*pp. 570-577). Washington, DC, USA.

O'Connor, B., Balasubramanyan, R., Routledge, B., & Smith, N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Web and Social Media (*pp. 122-129). Washington, DC, USA.

O'Connor, D. (2013). The apomediated world: regulating research when social media has changed research. *The Journal of Law, Medicine & Ethics, 41*(2), 470-483.

Paparrizos, J., White, R., & Horvitz, E. (2016). Screening for Pancreatic Adenocarcinoma using signals from web search logs: Feasibility study and results. *Journal of Oncology Practice, 12*(8), 737-744.

Pasek, J., & Dailey, J. (2019). Why don't tweets consistently track elections? Lessons from linking Twitter and survey data streams. In *Digital discussions: How big data informs political communication*, 68-93.

Pasek, J., Yan, H. Y., Conrad, F. G., Newport, F., & Marken, S. (2018). The stability of economic correlations over time: identifying conditions under which survey tracking polls and Twitter sentiment yield similar conclusions. *Public Opinion Quarterly*, *82*(3), 470-492.

Pasek, J., McClain, C. A., Newport, F., & Marken, S. (2019). Who's tweeting about the president? What big survey data can tell us about digital traces?. *Social Science Computer Review*, *38*(5), 633-650.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12,* 2825-2830.

Perrin, A., & Anderson, M. (2019). Share of US adults using social media, including Facebook, is mostly unchanged since 2018. *Pew Research Center.*

Pew Research Center (2018). Social Media Fact Sheet. *Pew Research Center Internet & Technology.* Retrieved from http://www.pewresearch.org/fact-sheet/social-media/.

Pfeffer, J., Mayer, K., & Morstatter, F. (2018). Tampering with Twitter's sample API. *EPJ Data Science*, *7*(1), 50.

Pushshift.io (2021). Pushshift.io: Learn about Big Data and Social Media Ingest and Analysis Retrieved from: https://pushshift.io/

The Python-Twitter Developers. (2019). Python-Twitter. Retrieved from: https://github.com/bear/python-twitter/wiki

Reitz, K. (2020). Requests 2.25.1. Retrieved from: https://pypi.org/project/requests/#history

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V. A., & Boyd-Graber, J. (2015). Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology* (pp. 99-107). Denver, CO, USA.

Resnik, P., Foreman, A., Kuchuk, M., Musacchio Schafer, K., & Pinkham, B. (2020). Naturally occurring language as a source of evidence in suicide prevention. *Suicide and Life-Threatening Behavior.*

Richardson, L. (2020). Beautiful Soup 4.9.3. Retrieved from: https://pypi.org/project/beautifulsoup4/

Roesslein, J. (2020). Tweepy: Twitter for Python! Retrieved from: https://github.com/tweepy/tweepy.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*(3), 515–530.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (1997). *Database system concepts* (Vol. 5). New York: McGraw-Hill.

Singh, L., Yang, G. H., Sherr, M., Hian-Cheong, A., Tian, K., Zhu, J., & Zhang, S. (2015). Public information exposure detection: Helping users understand their web footprints. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 153-161). Paris, France.

Singh, L., Traugott, M., Bode, L., Budak, C., Davis-Kean, P. E., Guha, R., Ladd, J., Mneimneh, Z., Nguyen, Q., Pasek, J., Raghunathan, T., Ryan, R., Soroka, S., Wahedi, L. (2020a). *Data blending: Haven't we been doing this for years?* [White paper]. Georgetown Massive Data Institute Report. https://live-guwordpress-mccourt.pantheonsite.io/wp-content/uploads/2020/05/MDI-Data-Blending-White-Paper-April2020.pdf

Singh, L., Polyzou, A., Wang, Y., Farr, J., & Gresenz, C. R. (2020b). Social Media Data-Our Ethical Conundrum. *IEEE Data Engineering,* 23.

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2020). Linking survey and twitter data: Informed consent, disclosure, security, and archiving. *Journal of Empirical Research on Human Research Ethics, 15*(1-2), 63-76.

Small, M. L. (2009). How many cases do I need?' On science and the logic of case selection in field-based research. *Ethnography, 10*(1), 5-38.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review.*

Sysomos. (2021). Sysomos: Now meltwater social. Retrieved from https://sysomos.com/

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 505-514). Ann Arbor, MI, USA.

University of Maryland (UMD) & NORC. (2021). The Mental Health Data Enclave. Retrieved from https://enclave.umd.edu/

U. S. Bureau of Labor Statistics (2021). Current Population Survey. Retrieved from https://www.bls.gov/cps/home.htm

U.S. Census Bureau (2021). American Community Survey (ACS). Retrieved from https://www.census.gov/programs-surveys/acs

Waksberg, J. (1978). Sampling methods for random digit dialing. *Journal of the American Statistical Association, 73*(361), 40-46.

Wang, W., Rothschild, D., Goel, S., & Gelman, A. (2015). Forecasting elections with non-representative polls. *International Journal of Forecasting*, *31*(3), 980-991.

Wang, Q., & Taylor, J. E. (2014). Quantifying human mobility perturbation and resilience in Hurricane Sandy. *PLoS one, 9*(11), e112608.

Wienberg, C. (2017). *Demographic bias correction for social media data* [Doctoral dissertation, University of Southern California].

Wojcik, S. & Hughes, A. (2019). Sizing up Twitter users. *Pew Research Center.*

Wojcik, S., Bijral, A. S., Johnston, R., Ferres, J. M. L., King, G., Kennedy, R., Vespignani, A. & Lazer, D. (2021). Survey data and human computation for improved flu tracking. *Nature Communications, 12(*1), 1-8.

Zhang, J., Hu, X., Zhang, Y., & Liu, H. (2016). Your age is no secret: Inferring microbloggers' ages via content and interaction analysis. In *Proceedings of the International AAAI Conference on Web and Social Media* (pp. 476-485). Cologne, Germany.

Zimmer, M. (2010). "But the data is already public": on the ethics of research in Facebook. *Ethics and Information Technology, 12*(4), 313-325.