The Future of Computational Social Science

David Lazer and Katherine Ognyanova

Citation: Lazer, D., & Ognyanova, K. (2024). The future of computational social science. In J. M. Box-Steffensmeier, D. P. Christenson, & V. Sinclair-Chapman (Eds.), *Oxford Handbook of Engaged Methodological Pluralism in Political Science* (pp. 1–22). Oxford University Press.

In recent decades, quantitative social sciences have witnessed a gradual paradigm shift towards analysis of larger-scale, richer data, with an emphasis on connectivity and complexity. This development is part of a broader epistemic shift in science driven by enhanced instrumentation that collects massive amounts of data, combined with matching increases in computational power. As a result, we have seen the rise of computational natural sciences (biology, physics, astronomy, etc.) and, a step behind, computational social science. The objective of this chapter is to review some of the advances in computational social science, as well as discuss the distinct intellectual opportunities and obstacles the field faces.

What is Computational Social Science?

Computational social science (CSS) is an interdisciplinary research field combining tools and techniques from computer science with theories and methods from the social sciences. CSS develops and applies computational methods to complex data about human behavior in order to understand social phenomena (Lazer et al. 2020)

In recent decades, technological advances have given us access to vast amounts of digital information about human actions and attitudes. We have records of behaviors ranging from communication and social interactions, information consumption and production, location and physical movement, purchasing and financial transactions, to a wide variety of day-to-day activities that people engage in. Traditional social scientific methods designed for small and relatively homogeneous samples were not suited for data of this volume or variety. Computational social science emerged as a way of allowing researchers

to collect and process large-scale, complicated, and interconnected data to gain insights into social phenomena. The field is especially well suited for the study of complex systems and dynamic processes, as its tools allow us to examine factors at the individual, community, and societal levels (Lazer et al. 2021).

This chapter outlines several key aspects of social scientific research using computational methods. We begin by providing a brief overview of computational social science, including its emergence and evolution. The next section explores types of data that may require (or benefit from) the use of computational methods. We follow up by discussing the role of theory in computational social science. The final part of the chapter focuses on current trends, challenges, and opportunities for CSS.

Background

The last 20 years have represented an important epistemological turning point for the social sciences from the behavioral approaches that emerged in the 1950s, that emphasized an agent by variable framework. This behavioral revolution, in turn, was driven in part by the marriage of computational power and survey research in the 1950s and 60s (Lepore 2020).

The current moment of computational social science is a synthesis of many threads. It builds on: long standing work using archival data for social science research; detailed ethnographic observational research; spatially-oriented research; hand coding of text and images; social network analysis; and agent-based modeling that explored the complexity of human behavior (Epstein and Axtell 1996). It mixes a theoretical starting point of social systems as essentially complex (Macy and Willer 2002), with rich data on human behavior, and computational methods that render those data tractable.

The quality of quantity

When it comes to study samples, social science research since the 1930s has largely emphasized *quality* over *quantity*. Quality has typically been linked to sample representativeness, which provides the important benefit of generalizability. However, the focus on quality has neglected the benefits of having large quantities of data, which are often available with large scale observational efforts.

Quantity creates its own inferential challenges, but also offers unprecedented opportunities for researchers. Importantly, large quantities of data allow for the careful examination of heterogeneity within populations. For example, in many cases, the same conditions or events may produce different (sometimes diametrically opposed) reactions across different groups of people. These nuances are difficult to capture in studies with a small sample size which tend to focus on average effects. For instance, a lab experiment with a few hundred participants may tell us how people in general would respond to a treatment. The experiment can also help us find out whether the effects on men would differ from those on women. In contrast, a dataset with a sample size in the millions might allow us to understand the effects on a much more precisely defined social group — such as young African-American mothers living in Southern states, or Hispanic medical professionals working in large cities. At the same time, these large-scale digital data have their own challenges, largely lacking the controlled conditions of a lab experiment.

Computational methods allow us to sift through massive data to find the scientifically relevant pearls. Chetty et al. (2016) is exemplary, using contemporary administrative tax data containing hundreds of millions of records to identify the causal effects of an experiment on residential relocation on thousands of children in 1970s households. Finally, having many observations at the individual level (e.g., real time monitoring) allows powerful within-individual inferences, as we can see in large-scale studies of social media data exploring variables like mood and sleep (Golder and Macy 2011).

Simulation studies

Computational social science research often seeks to model and predict the behavior of complex systems. Simulation techniques provide a flexible way to examine how collective behavior emerges from individual decisions, transactions, and structural factors. *Agent-based simulations*, which model larger structures arising from the actions and interactions of autonomous agents (Namatame and Chen 2016), are becoming increasingly popular in health, economics, and the social sciences. Many of the strengths and weaknesses of the approach in social science stem from its relation to the real-world dynamics an application represents (Marshall 2017). On one hand, simulations can help explain how individual decisions drive the system-level patterns we observe. On the other hand, it is often difficult to evaluate the extent to which the built-in assumptions of an agent-based model reflect the reality of human behavior. Large-scale detailed observational data can mitigate that challenge, allowing for a more accurate representation of the rules

governing our system and the behavior of actors within it. More generally, it is clear that there are strong synergies between large-scale data on behavior, and building simulations informed by those data.

Simulation studies are also gaining a wider use for policy recommendations. During the COVID-19 pandemic, for instance, simulations were used not only to model the spread of the virus (Castro, Reis, and Salles 2022), but also to evaluate vaccination strategies (Wrigley-Field et al. 2021), and preventive behaviors (Ku et al. 2021).

The *data* behind CSS

Data used in computational social science research come from a multitude of sources in a variety of formats. We can distinguish among them based on their *origin*, *format*, and their *structure*. In this section, we mention a few data types found "in the wild", often created for purposes other than research: text, multimedia, location, and other sensor data. We also discuss more structured data typically produced after imposing some organization, such as networks, archives, and large-scale surveys. Finally, we note that computational methods can also be usefully applied to small and qualitative datasets.

An important format that we do not specifically address here involves the digital records of human behavior. Those include server logs, user activities on social media, and platform data from a variety of APIs, among others. Given their considerable format diversity, we do not discuss them separately in this section. We note, however, that they are a key source of large-scale digital data, and refer to them throughout this chapter.

Networks

Many of the emerging data sources, especially ones recording patterns of communication, are intrinsically relational. Their focus is not only on individual characteristics and activities, but also on connections. Indeed, there are many modalities of data that do have some network component. Cell phone (Onnela et al. 2007), email (Kossinets and Watts 2009), and social media (Park, Blumenstock, and Macy 2018) data all involve connections from one individual to another. Tax, census, and genealogical data involve kinship ties. Locational and media consumption involve connections of individuals to physical and virtual loci (Yang et al. 2020). Individual location data can connect people based on their proximity to others, while aggregate movement data connect locations (Woudenberg et

al. 2020). Political contribution data involve ties from contributors to recipients (Bonica 2014). Video, image, and textual information often include individuals and/or organizational entities linked through shared appearances in the data (Dietrich 2016).

Analyzing networks can be highly computationally intensive. Methodologies for relational data have to account not only for individual characteristics, but also for all existing and possible connections among the involved actors. In the early days of network science, scholarship often focused on identifying key actors, detecting connected groups, and describing the structure of ties at various levels (Wasserman & Faust, 1994). Technological innovations in the last few decades have facilitated the development of new analytical methods going beyond the descriptive (Light & Moody, 2020). Research can now investigate the complex mechanisms of network evolution, as well as identify the drivers behind tie formation and dissolution. Network science, previously perceived by some as atheoretical, has also developed useful explanatory frameworks that support theory building in a variety of disciplines.

Text

While network analysis captures the structure of social interactions, understanding their content requires additional methodological and theoretical tools. A large proportion of the digital data recording interpersonal and mass communication is stored in the form of unstructured text. Available textual information includes digitized archives, books, news stories, academic publications, web pages, emails, social media posts, search terms and results, ads, and many other formats.

With the increasing amounts of textual data, industry and academic researchers have developed more advanced tools for its computational processing (Grimmer and Stewart 2013). Extracting meaning from a document (e.g., identifying objects and the relations among them) remains a considerable challenge. At the same time, computational methods have gotten better at classifying large sets of documents based on their source, topic, sentiment, ideology, and other characteristics. At present, computers remain less accurate at such tasks compared to human coders and manual annotations are still the gold standard (van Atteveldt, van der Velden, and Boukes 2021). Allowing for some errors, however, the vast scale and coding consistency possible with automated methods are changing the way social scientists think about content analysis.

Images, audio, video

New digital technologies for multimedia recording, editing, and distribution have vastly increased the amount of images, audio, and video we consume on a daily basis. The widespread use of smartphones with high-resolution cameras and the growing storage capacity of digital platforms made multimedia formats ubiquitous in the sharing and accessing of information about our social environment.

The social sciences and humanities have a long tradition in studying visual communication (Barthes 1964). More recently, scholars have examined the viral diffusion of images and videos (Milner 2016) and their role in political persuasion. Casas et al. (2019), for instance, examine how emotions produced by images on social media affect political mobilization.

Yet, social science is only now beginning to take advantage of computational tools that can classify images, video, and audio directly, without relying on accompanying text provided by the users or platforms. Researchers examining visual data often use small samples and rely on human coding (Chen et al. 2021). A lot of our understanding of the role played by multimedia content comes from experimental research (Gerber et al. 2011; Powell et al. 2015). Image recognition and object detection are likely to be crucial in the toolkit of future computational social science researchers (Webb Williams, Casas, and Wilkerson 2020).

Audio data remains similarly underused, though interesting applications in research (beyond just transforming speech into text) are becoming more common. Political scientists, for instance, have examined the link between candidate voice and election outcomes (Klofstad 2016), as well as studying the pitch of congressional speeches to measure their emotional intensity (Dietrich, Hayes, and O'brien 2019). This approach opens up possibilities for a more thorough understanding of the way audiovisual cues affect the outcomes of interpersonal and mass communication.

Location data

The mobile devices most of us constantly carry along generate large amounts of data on human location and movement. That information is accessible to applications and services which use it to make real-time decisions. The uses of geodata go beyond just finding where we are on a map or giving us driving directions — our position affects the information, recommendations, and advertising we see on our screens. Location is often

shared during ordinary transactions we make on a daily basis — like posting on social media or searching the internet.

For researchers, location data allows studying human behavior at a number of levels: from individual movement and proximity to others, to international flows of human mobility around the globe. The most striking recent example was the use of geolocation from mobile devices in public health scholarship during the COVID-19 pandemic. Researchers and governments in the United States, China, Japan, Italy, Germany, UK and other countries collaborated with mobile operators to examine the spread of the disease and the effectiveness of social distancing measures (Oliver et al. 2020). While there are considerable privacy concerns with that approach, some governments used individual phone data to control the spread of the virus and track the contacts of infected citizens. Aggregated cell phone data was used to identify mobility patterns (Levin et al. 2021) that could predict where the disease would spread next. Geodata could enable the effective use of travel restrictions to control the pandemic, as well as help in assessing the health and economic consequences of adherence to social distancing recommendations. It is singularly helpful in allowing scholars to examine the outcomes of different approaches to regulation at the national and local level.

Location-related data is used in a number of research areas in the social sciences, including the study of social networks where proximity of mobile devices might indicate that their owners are spending time together. Network research has, for instance, explored the ways in which spatial and social behavior are interconnected (Alessandretti, Lehmann, and Baronchelli 2018). In political science, aggregated geodata made possible the tracking of participation in mass protests (Rotman and Shalev 2022) and citizen displacement during violent conflicts (Tai, Mehra, and Blumenstock 2022).

In addition to challenges in processing and interpreting location information, key issues remaining in this space include obtaining access from mobile phone companies and protecting the privacy of individuals in the data.

Sensor data

In addition to sensor data about location, movement, and proximity, the ubiquitous devices in our pockets and in the built environment can track a wide variety of metrics, capturing anything from individual behavior to environmental conditions. A wearable

tracker can record not only your daily activity levels, but also biometric data such as heart rate, blood pressure, or blood oxygen.

With the advent of "the internet of things", "smart homes", and "smart cities", many of our devices have acquired built-in sensors which upload information to the cloud. Those sensors can be used to measure an increasing number of physical phenomena, including temperature, light intensity, sound levels, touch or pressure, air humidity, or presence of dangerous gasses. Smart cities aim to deploy sensors that monitor the environment, including the weather, air pollution, water quality, and traffic.

Much of the extant research using sensors focuses on applications in health, environmental studies, transportation, and urban planning (e.g., Pentland et al. 2009). Yet those data also offer important opportunities to study human behavior in the social sciences. They can be used to evaluate how our actions change in response to different stimuli, how we interact with each other, with the environment, and with technology. Biometric information, for instance, has provided insights in the fields of psychology, communication, and political science. Heart rate, blood pressure, and skin conductance have been used to evaluate people's attentiveness and arousal when exposed to political messages (Soroka et al. 2019). Studies have also used eye tracking to examine how users process visual information and what information on the screen draws their attention (Vraga et al. 2019).

Archival data

In addition to vast records of people's current behavior, technology has made it possible to go back in time and digitize historical information and artifacts. Archives maintained by government agencies, libraries, publishers, museums, and galleries are becoming available in new formats. Those often include not only digital versions of the original objects, but also important metadata that describes them.

One of the best known archival projects, *Google Books*, offers an example of digitizing analogue data at a massive scale. In 2019, Google announced that they had digitized more than 40 million books in over 400 languages, even though the initiative was somewhat slowed down by copyright lawsuits. While there are still errors and limitations in the scanned text and included metadata, the project offers an extraordinary opportunity to study the evolution of language and culture. The *Google Ngram* service allows users to track the frequency with which sequences of words appear in books over

time. The ngram data has been used in studies of language (Montemurro and Zanette 2016) and cultural history (Schich et al. 2014). Interestingly, scholars have also used Google Ngram to study broader questions, including the weakening of US social norms over time (Jackson et al. 2019).

Newly produced online information also presents a record of our history that might disappear if not preserved. The Internet Archive, a non-profit organization founded in the 1990s, has taken up the mission of archiving the world wide web. Scholars have also used the Internet Archive to extract network data, focusing on hyperlinks among websites. Research using these data has examined the evolution of online organizations (Weber and Monge 2017) and the transition of journalism to the web (Weber, Ognyanova, and Kosterich 2017). As we note below, a key gap in archiving the web is any effort to capture the ephemeral internet, including search, recommendations, feeds, and other types of data that are omnipresent on the internet.

Self-reported data

Surveys are not typically mentioned as a source of big data requiring computational methods. A traditional social science approach, the gold standard for surveys involves a small probability sample with each member of the population having an equal chance of being selected.

Today, the landscape of polling has changed. With landlines no longer in wide use and low mobile phone response rates, probability surveys have become more difficult and expensive to conduct. At the same time, the internet now allows for fast and inexpensive polling with large-scale online nonprobability samples. Working with such data can present statistical and computational challenges, yet it is likely to play an important part in the future of polling.

We have also learned a lot more about adjusting non-probability samples to match the characteristics of the population. One well-known work seeking to show that properly adjusted non-representative data can produce accurate forecasts comes from Wang et al. (2015). Their study uses polling data from over three hundred thousand users of the Xbox gaming platform (a group in no way expected to be representative of likely voters) to correctly predict the 2008 presidential election results.

In recent years, large-scale online surveys have been conducted to track the spread of COVID-19. A group of Carnegie Mellon researchers collaborating with Facebook polled the social media platform's users, collecting over 20 million responses in their first year of operation (Salomon et al. 2021). Taking a somewhat different approach, the COVID States Project[1] focused on quota-based sampling in each of the 50 US states and DC. From 2020 to 2022, the project launched 25 survey waves, collecting close to 500 thousand responses from 350 thousand unique online participants. After quality control and adjustments, that data performed well against a variety of benchmarks, including infection and vaccination rates as well as political behavior (Green et al, 2023).

In addition to adjusting nonprobability samples to match population characteristics, challenges with large-scale online surveys include ensuring respondent quality and preventing automated responses (Kennedy et al. 2020). This is a work in progress, unlikely to be completely resolved in the foreseeable future, but necessary to minimize survey error and bias.

Combining survey data with digital records (e.g. social media posts) is another especially productive avenue for research, as it allows us to match reported attitudes and online behavior (Wojcik et al. 2021). An interesting emerging form of data collection presents questions to respondents that pop up on their device at a particular time, when they are at a particular location, or when they engage in a predefined activity (e.g. reading a news article, purchasing a product, or posting on social media).

Small and qualitative data

While computational methods are most often applied to large datasets of digital records, that is not the only context in which they are useful. In some cases, the approaches provided by CSS are productively employed to explore very small data or support qualitative research. In network science, for instance, scholars have applied exponential random graph models to examine networks among a small number of wealthy Florentine families, a few dozen karate club members, a handful of sociable dolphins, or the students in a single high-school class (Caimo and Friel 2011).

[1] A consortium of researchers from Harvard, Northeastern, Northwestern, and Rutgers university, including the two chapter authors (see covidstates.org).

In the realm of digital humanities, machine learning methods for natural language processing can be used to examine a single work of literature or a set of related poems. In those cases, the computational approach is often helpful in support of a qualitative text analysis applied to historical records, literary works, or other cultural objects.

Automated text analysis tools can oversimplify some results and overcomplicate others, missing nuances that an experienced scholar can capture. Digital humanities researchers, for instance, have spent several years debating the merits of automated sentiment analysis methods seeking to extract emotional plot arcs from classics like *Oliver Twist* or *Pride and Prejudice* (Kim 2022).

At the same time, computational tools perform very well at some important tasks, such as examining text features to identify the author of a document. For example, algorithms designed for such identification tasks have been tested on *The Federalist Papers* to identify the most probable author of 12 disputed papers.

Overall, computational methods can typically achieve higher accuracy and usefulness when applied to (or trained on) larger datasets. Yet, we find it important to emphasize that they can be successfully applied to smaller or qualitative datasets (Tanweer et al, 2021).

The *theory* behind CSS

Combining theory and found data

While computational social science is often seen as a methodologically driven area, it does have an important role to play in the development, modification, and testing of theory. The core mission of CSS is to advance our understanding of complex social systems. One way to do that is to build on existing theoretical frameworks seeking to explain human behavior, using larger scale data to capture nuances and temporal dynamics that may have been previously unknown. A key challenge in that regard, however, is that theory and available data are not always perfectly aligned.

The mapping of what is measured to theoretically driven constructs is essential to all science. Much of existing social science is based on constructs built to evaluate a particular theory and operationalized specifically for the purposes of research. In contrast, computational social science largely relies on “found” data. That includes user-generated

content, media content, digitized offline information, and a variety of digital records created as part of the functioning of an online system.

To serve the goals of computational social science, found data need to be repurposed for research. This does not make them intrinsically less valid than intentionally designed constructs, but it often presents a tension. For example, using IRS tax-related data (Chetty et al. 2014) may well be more accurate than asking people how much they earn; but measuring friendship using Facebook ties would be problematic. Leaving aside the fact that defining "friendship" is challenging on its own, it seems unlikely that either the academic or the colloquial definition of "a friend" will fit everyone we connect to on social media.

Signals and constructs

In order to produce theoretically robust computational social science scholarship, we need to understand the relationship between the signals from our data and the constructs from our theory. To do that, the CSS field has to dedicate efforts to two basic (but not simple!) tasks.

The first task involves ***mapping signals to known constructs***. Researchers need to understand how they can operationalize existing theoretical constructs using new types of data. We have developed measures of social capital in an offline setting. How can we measure social capital using digital behavior records? We know how to evaluate depression using a nine-item survey scale. Can we extract reliable metrics of depression from unstructured text posted on social media? Scholars are already making steps in that direction, trying to predict self-reported measures from digital metrics. One good example of this is the literature discussing ways to identify political ideology from social media data (Barbera 2015).

Another important aspect of CSS scholarship has to do with **mapping constructs to known signals**. The idea in this case is to identify the theoretically relevant meanings of observable digital behavior. Does "liking" a tweet suggest approval? Does clicking on media links correspond to classic measures of news use? Can we use the same signal to measure political interest? Is there a need to develop an entirely new theoretical construct that fits better? Quite often, such questions may not have a simple answer. It may be challenging, for instance, to find a construct that cleanly maps onto the signal of "Facebook friendship." In these cases, we can also advance scholarship by doing important

descriptive work.[2] We might ask questions such as "What is the distribution of tie strength among Facebook friends?" or "How does the composition of Facebook friends differ across social groups?" The descriptive details can help us when we need to answer more complicated questions, such as "Can Facebook friends influence a person's political behavior?" or "How do friendship ties differ across various social media platforms?"

Conducting research that unpacks the links between theoretical constructs and digital signals can be a challenge. Such works require extensive data collection including both the new digital signal and the classically measured construct. In many cases, this involves a combination of observational data with self-reported survey metrics (e.g., Eagle, Pentland, and Lazer 2009). The work on validating a construct can thus be fairly resource-intensive, while also appearing less appealing than testing exciting new hypotheses. Yet, it is crucial in a field where many works still take for granted the untested equivalence between measures and constructs.

The *challenges* of CSS

The substantial reliance of computational social science on "found" data created for nonscientific purposes is a considerable challenge. As we discuss above, this poses a measurement problem: how do we turn messy raw material into a highly polished scientific product? Even before we ask that question, however, we run into another major obstacle: the messy raw material we need is rarely freely available. As is often the case with valuable materials, it is largely the property of governments and corporations.

Access

Access to data presents a major difficulty for scholars working in the area of computational social science. Most datasets of interest are proprietary and virtually all of them pose considerable privacy and ethical challenges (see more in following sections).

Powerful companies are often the ones who decide what digital data will be accessible and what aspects of their platforms and products will be visible to the public. This is a

[2] Similar considerations have motivated the recent launch of a new journal, *Journal of Quantitative Description: Digital Media*. Both authors are members of its editorial board.

particular concern with respect to analyses published by those corporate entities (Lazer 2015), but it may also apply to any research from data extracted from their systems. Companies may, for instance, hold back data which might reveal information that would be useful to competitors or provide fuel for regulation and policy intervention demands.

Such issues are not unique to computational research — they are endemic in the interplay of power and science. The 17th-century Church, for instance, did not encourage investment in telescopes; and fields of science are sometimes held captive by cadres of scholars raised in old paradigms (per Planck's quip that science advances one funeral at a time). In the case of CSS, the entities that produce data are not neutral with respect to the insights that they might yield. And, even when those entities are oblivious — consider, for instance, the Elon Musk-led changes at Twitter — this does not obviously play to the advantage of science. The particular reliance of the field on Twitter — with its generous APIs and well-structured data, with rich textual, visual, relational, and geographic dimensions — suddenly looks like a risky bet for CSS in 2023. It is as if the lenses of half of the telescopes of the world may have cracked.

Ethics

Many of the data pose particular challenges for the current self-regulatory approaches in research in academia. Vast archives of human data — typically collected for non-research purposes — have emerged as foci of CSS research. It is far more feasible to observe "public" behaviors of individuals today than it was a generation ago. For example, whether people voted and their political contributions were, in principle, public a generation ago; but in practice, quite hard to access. In contrast, social media are intrinsically visible to third parties, and Twitter, in particular, has been a mainstay of CSS research for the last decade. There are, furthermore, many proprietary archives of human behavior that have been utilized for research purposes. Very often, because they are pre-existing, these archives are considered "exempt" from human subject research guidelines. At the same time, it is clear that scholars using such datasets without careful consideration of ethics can harm individuals and groups in the data. Unsurprisingly, an ethical framework for research built in the 1970s did not anticipate the issues of computationally powered research on massive third-party data archives.

Another set of ethical issues surround the utilization of large-scale experiments on virtual platforms. The malleability of digital spaces provides enormous possibilities for testing

social science theories. Prominent examples include manipulations to encourage people to vote on Facebook (Bond et al. 2012; Jones et al. 2017); an experiment to evaluate the spread of emotions on Facebook (Kramer, Guillory, and Hancock 2014); and a recent test of the "strength of weak ties" in helping people find jobs (Rajkumar et al. 2022). These experiments must, by definition, affect people's experience on the platform, yet they rarely involve consent (except, perhaps, for broadly phrased language in the terms of service).

Online experiments have also typically been considered exempt from human subjects review since they are conducted by the platform and not by academic researchers. That has been the case even when an academic has been involved in the design of the data collection. While Meyer et al (2019) persuasively argue that an experiment with multiple conditions is not intrinsically problematic, the fact remains that there is a major gap in the systems that regulate the ethics of human subjects research when the data collection is conducted by corporate entities. Note that this is not relevant to experiments that are directly conducted by academic researchers; the de-activation experiment in Allcott et al. (2020) is exemplary in this regard.

Standards

Conducting computational research to examine human behavior is still a relatively new practice. Academic researchers and companies alike are developing standards and best practices for the collection, storing, and sharing of information. One of the most common tasks of computational scholars, combining datasets from different sources, becomes daunting when those sources have different (and often non-transparent) systems for measuring and recording social phenomena. Clear data standards are also crucial for conducting replications studies and data reanalyses, both of which are integral to a robust process of knowledge production in science.

Things are similarly complicated with regards to ethical standards and practices. Policies about consent, information privacy, and participant protection vary widely not only between academic and industry research, but also between scholars in different disciplines, universities, or countries. Developing a common vocabulary and a shared understanding of key ethical principles is a key challenge ahead of the field.

Data quality

Even when scholars have clear and robust ways of matching theoretical constructs to digital signals, good research may be hindered by low data quality. The "found" data designed for nonscientific purposes is often messy, inconsistent, biased, and generally problematic in ways that may be difficult to predict. Individual data points may contain considerable error, and researchers have to ensure that the carefully processed and aggregated data reveal meaningful patterns.

Different sources of data vary in the levels of curation they offer. While the Smithsonian may carefully select, describe, and label objects in their archives, online forum data scraped from the internet is rarely that clean and organized. Understanding dataset provenance, quality, and collection methods is crucial if we want to make sure our findings are genuine and not the result of faulty information processing and misinterpretation.

Representation

Who is represented in the data has always been a central challenge for the social sciences. Random digit dialing surveys, for example, did not capture people without landlines; and overrepresented those with multiple landlines. Even the US census, one of the most rigorous and expensive national measurement efforts, largely misses the homeless despite recent attempts to capture that group. Any data collection requires a focal point — a frame within which individual units will be studied (e.g., landlines, homes). This is an issue for CSS methods as well, and for many contemporary digital data collections, the focal point involves activity on an online platform. Lurking (consuming information without engaging in other activities) is often invisible, even though it is an important behavior. For instance, Twitter is likely the most common source of computational social science data, and yet until recently research only captured the active use of the platform. The fairly common practice of just using Twitter to find information remains invisible to external researchers. Scholars often rely on activity as a measure for attention (Lazer 2019), which is surely distorting research results in many ways. We are bound to miss the people who pay attention to a piece of information without hitting the "like" or "share" buttons, while our research only records the non-representative group of users who do press those buttons.

More generally, CSS scholarship requires a careful consideration of the ways in which data collection affects the applicability of our findings. Are results from Twitter data generalizable to the entire world's population? To Americans? To social media users? To people on Twitter? To those who actively posted during the period under observation? To users who posted with specific hashtags? Many works overestimate or leave unclear the group to which their findings are applicable. This can be particularly problematic if we design policy based on social media data which happens to exclude people with low socioeconomic status, elderly people, or any other relevant social group.

Algorithmic confounding / temporal drift

Many sources of large-scale digital data suffer from algorithmic confounding. We define algorithmic confounding as the ephemeral computational structures that often guide human decisions, yielding human behaviors that cannot be properly interpreted as being purely the result of individual preferences. For example: Twitter routinely nominates accounts for users to follow. Ultimately, it is the choice of the user to follow an account; however, it is very likely that the nomination process strongly affects those choices. Research that relies on Twitter follower patterns may thus be vulnerable to changes in the Twitter algorithm (or changes in the input data to that algorithm).

The issue of structural confounding is, of course, endemic in social science; everything from where we travel, to what we purchase in the grocery store, potentially confounds interpretation. However, the particular challenge of online research is the fact that these momentary influences on decisions are typically not recorded and thus not available for research. As a result, much — perhaps the vast majority — of research that purports to be on online humans is really on cyborg behavior. In essence, we are looking at the behaviors of technology-powered and algorithm-influenced humans. The algorithm may change overnight, resulting in dramatic changes in the "cyborg", while the underlying human behavior in response to particular stimuli may have remained constant.

Adjacent to the issue of algorithmic confounding is the idea of *temporal drift*, which suggests that patterns of behavior observed today may not generalize to patterns we will see tomorrow. This is, more generally, a concern with respect to all of science, and especially to the social sciences. The malleability of human society means that the generalizability of scientific propositions and even of our measures must constantly be critically evaluated. Changes in human mobility, for instance, have radically altered the

rate and speed with which diseases spread. The dollar in 2022 means different things than it did in 2000, with some metrics fairly translatable over time (the cost of bread), and others not (the bitcoin exchange rate). The issue of temporal drift, however, is likely more acute with online research, in part because of its inherent dynamism (Munger 2019). Comparing the dollar of 2000 to 2022 poses some challenges; but that is a small leap relative to the comparison of the social media of 2000 to 2022. Or, to take on non-internet example; research positing a certain relationship between mobility and unemployment rates (Toole et al. 2015) would likely have greatly overestimated the surge in unemployment around the world at the beginning of the COVID-19 pandemic, when the locus of work for many people shifted to the home.

The fracturing of data silos

In a narrow (and ironic) sense, the opportunity to create good samples from digital trace data came before the incipient rise of the field. The very move of people to the online realm meant that theoretically similar behaviors may now fall into widely divergent siloes of proprietary technology. Consider, for instance, *synchronous, non-physically proximate, voice mediated communication*. There was a point in time, not too long ago, when this construct would simply have been described as a "phone call". At that time, all of the relevant US data would have been limited to phone records held by a single, state-sanctioned monopolist. This initial pattern gave way to an oligopoly, later joined by few internet-based voice-over-IP providers (e.g. Skype), until finally in 2023 there are a countless number of ways to talk with people over long distances. This shift has been good in many ways: long distance calls were very expensive 50 years ago, and are nearly free today. The emergence of countless providers, however, presents considerable challenges for research, which now may be based on odd, single-silo samples of behaviors. Consider, for example, work looking at call detail record data from a single cell phone carrier (Onnela et al. 2007). The resulting data set is massive, it includes detailed call information for millions of people, and yet it is weirdly biased. The data does not include landlines (which were still common when it was collected), it does not include other carriers, or various internet voice providers. It is, therefore, capturing the idiosyncratic sample of people who subscribed to a particular carrier, and a set of behaviors for that population using one among many available tools to reach others. It is the work of researchers to consider whether the findings of single-silo research are limited

to a particular company, to companies using similar technology, or to all long-distance voice communication.

The opportunities of CSS

Diversity and inclusion

The scale of many of the data used in CSS offer particular opportunities in studying subpopulations. For a long time, the social and medical sciences have had a strong majoritarian bias. Both for historical and for statistical reasons, the focus was on the larger social groups within the population. Minorities are often scientifically neglected, and smaller minorities even more so (Knight, Roosa, and Umana-Taylor 2009). This issue becomes especially problematic when there is heterogeneity in observed relationships across subpopulations.

Even today, social science has a strong tendency to study common individual-level outcomes (income, voting, employment). However, many important behaviors reside in small fractions of the population. For instance, there are vastly more academic papers in US politics on voting than on contributing to campaigns; and even those papers that look at contributing tend to examine median contributors rather than the median dollars. This is so even though we know the top few hundred contributors have an outsized impact on campaign funding and thus on political elections in the US.

In health and social policy research, we have underprivileged minority groups whose specific needs or behaviors may get lost when we focus on the entire population (Yancey, Ortega, and Kumanyika 2006). Large-scale digital data can give us the chance to study small groups or unusual events, escaping the traditional focus on majorities that comes with smaller research samples.

New metrics

The instrumentation of modern life makes it feasible to develop new measures of things, including constructs that have long been theoretically central to the social sciences. For example, while news consumption has been crucial to the study of mass communication since the beginning of the field, it has only recently become possible to unobtrusively measure what stories people actually read. Today, this is eminently doable — it is possible

to monitor digital media consumption, capture browsing data, record TV viewing, or get access to app usage (Watts, Rothschild, and Mobius 2021).

A variety of other behaviors have become possible or easier to measure. Physical activity and sleep patterns can be captured through personal devices. Network connections can be captured through device and platform records. New measures of previously unanticipated phenomena (*zoom fatigue*, anyone?) are also emerging daily.

Rebuilding and developing theory

Computational social science can play an important role in advancing existing theories and developing new models of social phenomena. Many classic frameworks need to be carefully reexamined as their premises reflect a long-gone social and information environment. Established ideas in communication, sociology, and political science, for instance, may have new implications today — or even be entirely outdated. The new spaces where socialization, communication, and information consumption happen require new methods of study. Computational research offers the necessary tools to replicate or evolve knowledge in a variety of domains.

Reconstructing history

One of the distinctive opportunities of computational social science is to reach back in time thanks to digitally tractable large corpora of analog information. It is not possible to go back in time to survey people regarding their beliefs and preferences; but it is possible to analyze the content of millions of books, as evident from the Google Ngram research we discuss in previous sections. Key sampling and measurement issues remain — for instance, how can we evaluate whether the sampling criteria have remained constant over time? How reliable is the software we use to digitize printed text over generations of typesets used in documents across multiple centuries? Yet, the key point remains: the opportunity for CSS to shed light on important phenomena stretches backwards in time, not just forward.

Conclusion: The future of CSS

There is every reason to expect that computational social science will grow in the next few years. The field is driven forward by an ever increasing capacity to instrument human

behavior and computationally make sense of the resulting massive stream of information. Over time, it will be possible to record more data about humans, and more of that data will be computationally accessible for analysis. Current advances in using text as data will be followed by increased use of images (and later video) as data. The epistemological advantage that these formats often offer, the analytic power of large scale data, the leverage on understanding context and complexity, should continue to attract scholars to the field.

How much of the data will be accessible to scholars is a more difficult question. It is likely that some uses of corporate data for research will decrease due to concerns about the economic interest of companies contaminating the science. We can point to recent examples from Facebook which illustrate typical corporate behavior in that regard. After the public relations disaster of publishing a controversial emotion contagion paper (Kramer, Guillory, and Hancock 2014), the company instituted careful review of externally facing scholarship to evaluate the impact of that scholarship on perceptions of Facebook before sharing it with the public. In 2015, perhaps, in anticipation of the public relations disaster of Cambridge Analytica, Facebook sharply restricted the platform's API (Lazer 2015).

Unfortunately, we can also anticipate the decay of free research APIs for access to data streams from online platforms (Freelon 2018). There are (at least) three reasons why: (1) reducing potential privacy impingements on users; (2) reducing the use of those data against the corporate interest; and (3) the recognition that those data are a core profit generator for the firm. One can also expect increased ethical concerns within the academy regarding such data. That prediction is at least partly due to the fact that the implicit consent of users to participate in research rests on their understanding of visibility their information has on social media. That said, it is likely that there will be countervailing pressures on companies to come up with viable models for providing research access to data. We do see that Facebook has attempted various innovations in providing increased analytic access, both through sharing aggregated data (King and Persily 2019), and through the Facebook election integrity project.[3] Soon these efforts would have do be expanded further thanks to new regulations. An important recent example, the Digital

[3] See https://research.facebook.com/2020-election-research. Note that Lazer is an investigator.

Services Act adopted by the European Union in 2022, requires large digital companies to share data with researchers.

We anticipate the rise of scientifically oriented data collection platforms such as the National Internet Observatory and Mozilla's Project Rally.[4] These research platforms represent a compromise among the potential strengths of CSS: well-consented subjects, carefully and deeply measured, with behaviors captured across multiple internet platforms; but at a much smaller scale and less representative than "all of a social media platform" approaches to the field.

At present, simulation research and empirical data analysis are rarely used together in the realm of social science. We expect that to change, as future work takes advantage of the flexibility of agent-based models combined with the representativeness of observational data. This shift should be helped along by the growing use of simulation in social science, as well as by the ever-increasing computational power of technology which makes complex modeling methods more viable.

We conclude by emphasizing the importance of including computational methods in the standard toolkit that social scientists acquire during their academic training. The relevant programming languages and software tools may change over time, but this training should reliably allow researchers to collect, process, and analyze large-scale digital data in theoretically meaningful and ethical ways. And, while not every social scientist will use computational methods, they need to be both strongly represented and well-understood across disciplines. Innovation enters a field as emerging scholars see new opportunities and acquire skills aligned with those opportunities. Some social science doctoral programs have integrated computational methods; and some information science programs are attempting a fusion of computer science and social science. It is in these efforts to create a synthesis in training that we will build a new generation of researchers who can combine robust theoretical understanding with large-scale data analysis in order to examine a complex new information environment. In turn, computational social science offers a shared vocabulary that can strengthen the bridges among disciplines in social, natural, and computer sciences from which the field draws.

[4] Note that Lazer is PI on the grant associated with the NIO and Ognyanova on an advisory committee.

References

Alessandretti, Laura, Sune Lehmann, and Andrea Baronchelli. 2018. “Understanding the Interplay between Social and Spatial Behaviour.” *EPJ Data Science* 7 (1): 36. https://doi.org/10.1140/epjds/s13688-018-0164-6.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. “The Welfare Effects of Social Media.” *American Economic Review* 110 (3): 629–76. https://doi.org/10.1257/aer.20190658.

Atteveldt, Wouter van, Mariken A. C. G. van der Velden, and Mark Boukes. 2021. “The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms.” *Communication Methods and Measures* 15 (2): 121–40. https://doi.org/10.1080/19312458.2020.1869198.

Barbera, Pablo. 2015. “Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data.” *Political Analysis* 23 (01): 76–91. https://doi.org/10.1093/pan/mpu011.

Barthes, Roland. 1964. *The Responsibility of Forms: Critical Essays on Music, Art, and Representation*. Translated by Richard Howard. Berkeley: University of California Press.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. 2012. “A 61-Million-Person Experiment in Social Influence and Political Mobilization.” *Nature* 489 (7415): 295–98. https://doi.org/10.1038/nature11421.

Bonica, Adam. 2014. “Mapping the Ideological Marketplace.” *American Journal of Political Science* 58 (2): 367–86.

Caimo, Alberto, and Nial Friel. 2011. “Bayesian Inference for Exponential Random Graph Models.” *Social Networks* 33 (1): 41–55. https://doi.org/10.1016/j.socnet.2010.09.004.

Casas, Andreu, and Nora Webb Williams. 2019. “Images That Matter: Online Protests and the Mobilizing Role of Pictures.” *Political Research Quarterly* 72 (2): 360–75. https://doi.org/10.1177/1065912918786805.

Castro, Brenno Moura, Marcelo de Miranda Reis, and Ronaldo Moreira Salles. 2022. “Multi-Agent Simulation Model Updating and Forecasting for the Evaluation of COVID-19 Transmission.” *Scientific Reports* 12 (1): 1–18. https://doi.org/10.1038/s41598-022-22945-z.

Chen, Yan, Kate Sherren, Michael Smit, and Kyung Young Lee. 2021. “Using Social Media Images as Data in Social Science Research.” *New Media & Society*, August, 1–23. https://doi.org/10.1177/14614448211038761.

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. “The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment.” *American Economic Review* 106 (4): 855–902. https://doi.org/10.1257/aer.20150572.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. 2014. “Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility.” *American Economic Review* 104 (5): 141–47. https://doi.org/10.1257/aer.104.5.141.

Dietrich, Bryce J. 2016. “If a Picture Is Worth a Thousand Words, What Is a Video Worth?” In *Exploring the C-SPAN Archives*, edited by Robert X. Browning, 241–64. Advancing the Research Agenda. Purdue University Press. https://doi.org/10.2307/j.ctv15wxr41.15.

Dietrich, Bryce J., Matthew Hayes, and Diana Z. O’brien. 2019. “Pitch Perfect: Vocal Pitch and the Emotional Intensity of Congressional Speech.” *American Political Science Review* 113 (4): 941–62. https://doi.org/10.1017/S0003055419000467.

Eagle, Nathan, Alex (Sandy) Pentland, and David Lazer. 2009. “Inferring Friendship Network Structure by Using Mobile Phone Data.” *Proceedings of the National Academy of Sciences* 106 (36): 15274–78. https://doi.org/10.1073/pnas.0900282106.

Epstein, Joshua M., and Robert L. Axtell. 1996. *Growing Artificial Societies: Social Science From the Bottom Up*. First edition. Washington, D.C: Brookings Institution Press & MIT Press.

Freelon, Deen. 2018. "Computational Research in the Post-API Age." *Political Communication*, October, 1–4. https://doi.org/10.1080/10584609.2018.1477506.

Gerber, Alan S., James G. Gimpel, Donald P. Green, and Daron R. Shaw. 2011. "How Large and Long-Lasting Are the Persuasive Effects of Televised Campaign Ads? Results from a Randomized Field Experiment." *The American Political Science Review* 105 (1): 135–50.

Golder, Scott A., and Michael W. Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures." *Science* 333 (6051): 1878–81. https://doi.org/10.1126/science.1202775.

Green, Jon, James N Druckman, Matthew A Baum, Katherine Ognyanova, Matthew D Simonson, Roy H Perlis, and David Lazer. 2023. "Media Use and Vaccine Resistance." *PNAS Nexus* 2 (5): 1–13. https://doi.org/10.1093/pnasnexus/pgad146.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (03): 267–97. https://doi.org/10.1093/pan/mps028.

Jackson, Joshua Conrad, Michele Gelfand, Soham De, and Amber Fox. 2019. "The Loosening of American Culture over 200 Years Is Associated with a Creativity–Order Trade-Off." *Nature Human Behaviour* 3 (3): 244–50. https://doi.org/10.1038/s41562-018-0516-z.

Jones, Jason J., Robert M. Bond, Eytan Bakshy, Dean Eckles, and James H. Fowler. 2017. "Social Influence and Political Mobilization: Further Evidence from a Randomized Experiment in the 2012 U.S. Presidential Election." Edited by Christopher M. Danforth. *PLOS ONE* 12 (4): e0173851. https://doi.org/10.1371/journal.pone.0173851.

Kennedy, Courtney, Nick Hatley, Arnold Lau, Andrew Mercer, Scott Keeter, Joshua Ferno, and Dorene Asare-Marfo. 2020. "Assessing the Risks to Online Polls From Bogus Respondents." Washington, D.C.: Pew Research Center.

Kim, Hoyeol. 2022. “Sentiment Analysis: Limits and Progress of the Syuzhet Package and Its Lexicons.” *Digital Humanities Quarterly* 016 (2).

King, Gary, and Nathaniel Persily. 2019. “A New Model for Industry–Academic Partnerships.” *PS: Political Science & Politics*, 1–7. https://doi.org/10.1017/S1049096519001021.

Klofstad, Casey A. 2016. “Candidate Voice Pitch Influences Election Outcomes: Voice Pitch Influences Voters.” *Political Psychology* 37 (5): 725–38. https://doi.org/10.1111/pops.12280.

Knight, George P., Mark W. Roosa, and Adriana J. Umana-Taylor. 2009. *Studying Ethnic Minority and Economically Disadvantaged Populations: Methodological Challenges and Best Practices*. 1st edition. American Psychological Association.

Kossinets, Gueorgi, and Duncan J. Watts. 2009. “Origins of Homophily in an Evolving Social Network.” *American Journal of Sociology* 115 (2): 405–50. https://doi.org/10.1086/599247.

Kramer, A. D. I., J. E. Guillory, and J. T. Hancock. 2014. “Experimental Evidence of Massive-Scale Emotional Contagion through Social Networks.” *Proceedings of the National Academy of Sciences* 111 (24): 8788–90. https://doi.org/10.1073/pnas.1320040111.

Ku, Donggyun, Chihyung Yeon, Seungjae Lee, Kyuhong Lee, Kiyeon Hwang, Yuen Chong Li, and Sze Chun Wong. 2021. “Safe Traveling in Public Transport amid COVID-19.” *Science Advances* 7 (43): eabg3691. https://doi.org/10.1126/sciadv.abg3691.

Lazer, David. 2015. “The Rise of the Social Algorithm.” *Science* 348 (6239): 1090–91. https://doi.org/10.1126/science.aab1422.

———. 2019. “Studying Human Attention on the Internet.” *Proceedings of the National Academy of Sciences*, December, 201919348. https://doi.org/10.1073/pnas.1919348117.

Lazer, David, Eszter Hargittai, Deen Freelon, Sandra Gonzalez-Bailon, Kevin Munger, Katherine Ognyanova, and Jason Radford. 2021. “Meaningful Measures of Human

Society in the Twenty-First Century." *Nature* 595 (7866): 189–96. https://doi.org/10.1038/s41586-021-03660-7.

Lazer, David, Alex Pentland, Duncan Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, et al. 2020. "Computational Social Science: Obstacles and Opportunities." *Science* 369 (6507): 1060–62. https://doi.org/10.1126/science.aaz8170.

Lepore, Jill. 2020. *If Then: How the Simulmatics Corporation Invented the Future*. Illustrated edition. New York, N.Y: Liveright.

Levin, Roman, Dennis L. Chao, Edward A. Wenger, and Joshua L. Proctor. 2021. "Insights into Population Behavior during the COVID-19 Pandemic from Cell Phone Mobility Data and Manifold Learning." *Nature Computational Science* 1 (9): 588–97. https://doi.org/10.1038/s43588-021-00125-9.

Light, Ryan, and James Moody, eds. 2020. *The Oxford Handbook of Social Networks*. Oxford University Press.Macy, Michael W., and Robert Willer. 2002. "From Factors to Factors: Computational Sociology and Agent-Based Modeling." *Annual Review of Sociology* 28 (1): 143–66. https://doi.org/10.1146/annurev.soc.28.110601.141117.

Marshall, Brandon D. L. 2017. "Agent-Based Modeling." In *Systems Science and Population Health*, edited by Abdulrahman M. El-Sayed and Sandro Galea, 0. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190492397.003.0008.

Meyer, Michelle N., Patrick R. Heck, Geoffrey S. Holtzman, Stephen M. Anderson, William Cai, Duncan J. Watts, and Christopher F. Chabris. 2019. "Objecting to Experiments That Compare Two Unobjectionable Policies or Treatments." *Proceedings of the National Academy of Sciences*, May, 201820701. https://doi.org/10.1073/pnas.1820701116.

Milner, Ryan M. 2016. *The World Made Meme: Public Conversations and Participatory Media*. The MIT Press.

Montemurro, Marcelo A., and Damián H. Zanette. 2016. “Coherent Oscillations in Word-Use Data from 1700 to 2008.” *Palgrave Communications* 2 (1): 1–9. https://doi.org/10.1057/palcomms.2016.84.

Munger, Kevin. 2019. “The Limited Value of Non-Replicable Field Experiments in Contexts With Low Temporal Validity.” *Social Media + Society* 5 (3): 2056305119859294. https://doi.org/10.1177/2056305119859294.

Namatame, Akira, and Shu-Heng Chen. 2016. *Agent-Based Modeling and Network Dynamics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198708285.001.0001.

Oliver, Nuria, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Deletaille, Marco De Nadai, Emmanuel Letouzé, et al. 2020. “Mobile Phone Data for Informing Public Health Actions across the COVID-19 Pandemic Life Cycle.” *Science Advances* 6 (23): eabc0764. https://doi.org/10.1126/sciadv.abc0764.

Onnela, J.-P., J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. 2007. “Structure and Tie Strengths in Mobile Communication Networks.” *Proceedings of the National Academy of Sciences* 104 (18): 7332–36. https://doi.org/10.1073/pnas.0610245104.

Park, Patrick S., Joshua E. Blumenstock, and Michael W. Macy. 2018. “The Strength of Long-Range Ties in Population-Scale Social Networks.” *Science* 362 (6421): 1410–13. https://doi.org/10/gfrksw.

Pentland, Alex, David Lazer, Devon Brewer, and Tracy Heibeck. 2009. “Using Reality Mining to Improve Public Health and Medicine.” In *Strategy for the Future of Health*, 93–102. IOS Press.

Powell, Thomas E., Hajo G. Boomgaarden, Knut De Swert, and Claes H. de Vreese. 2015. “A Clearer Picture: The Contribution of Visuals and Text to Framing Effects.” *Journal of Communication* 65 (6): 997–1017. https://doi.org/10.1111/jcom.12184.

Rajkumar, Karthik, Guillaume Saint-Jacques, Iavor Bojinov, Erik Brynjolfsson, and Sinan Aral. 2022. “A Causal Test of the Strength of Weak Ties.” *Science* 377 (6612): 1304–10. https://doi.org/10.1126/science.abl4476.

Rotman, Assaf, and Michael Shalev. 2022. “Using Location Data From Mobile Phones to Study Participation in Mass Protests.” *Sociological Methods & Research* 51 (3): 1357–1412. https://doi.org/10.1177/0049124120914926.

Salomon, Joshua A., Alex Reinhart, Alyssa Bilinski, Eu Jing Chua, Wichada La Motte-Kerr, Minttu M. Rönn, Marissa B. Reitsma, et al. 2021. “The US COVID-19 Trends and Impact Survey: Continuous Real-Time Measurement of COVID-19 Symptoms, Risks, Protective Behaviors, Testing, and Vaccination.” *Proceedings of the National Academy of Sciences* 118 (51): e2111454118. https://doi.org/10.1073/pnas.2111454118.

Schich, Maximilian, Chaoming Song, Yong-Yeol Ahn, Alexander Mirsky, Mauro Martino, Albert-László Barabási, and Dirk Helbing. 2014. “A Network Framework of Cultural History.” *Science* 345 (6196): 558–62. https://doi.org/10.1126/science.1240064.

Soroka, Stuart, Patrick Fournier, Lilach Nir, and John Hibbing. 2019. “Psychophysiology in the Study of Political Communication: An Expository Study of Individual-Level Variation in Negativity Biases.” *Political Communication* 36 (2): 288–302. https://doi.org/10.1080/10584609.2018.1493008.

Tai, Xiao Hui, Shikhar Mehra, and Joshua E. Blumenstock. 2022. “Mobile Phone Data Reveal the Effects of Violence on Internal Displacement in Afghanistan.” *Nature Human Behaviour* 6 (5): 624–34. https://doi.org/10.1038/s41562-022-01336-4.Tanweer, Anissa, Emily Kalah Gade, P. M. Krafft, and Sarah Dreier. 2021. “Why the Data Revolution Needs Qualitative Thinking.” *Harvard Data Science Review* 3 (3). https://doi.org/10.1162/99608f92.eee0b0da.

Toole, Jameson L., Yu-Ru Lin, Erich Muehlegger, Daniel Shoag, Marta C. González, and David Lazer. 2015. “Tracking Employment Shocks Using Mobile Phone Data.” *Journal of The Royal Society Interface* 12 (107): 20150185. https://doi.org/10.1098/rsif.2015.0185.

Vraga, Emily K., Leticia Bode, Anne-Bennett Smithson, and Sonya Troller-Renfree. 2019. “Accidentally Attentive: Comparing Visual, Close-Ended, and Open-Ended Measures of Attention on Social Media.” *Computers in Human Behavior* 99 (October): 235–44. https://doi.org/10.1016/j.chb.2019.05.017.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. “Forecasting Elections with Non-Representative Polls.” *International Journal of Forecasting* 31 (3): 980–91. https://doi.org/10.1016/j.ijforecast.2014.06.001.

Watts, Duncan J., David M. Rothschild, and Markus Mobius. 2021. “Measuring the News and Its Impact on Democracy.” *Proceedings of the National Academy of Sciences* 118 (15). https://doi.org/10.1073/pnas.1912443118.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications*. New York, NY: Cambridge University Press.Webb Williams, Nora, Andreu Casas, and John D. Wilkerson. 2020. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification*. Cambridge University Press. https://doi.org/10.1017/9781108860741.

Weber, Matthew, and Peter R. Monge. 2017. “Industries in Turmoil: Driving Transformation During Periods of Disruption.” *Communication Research* 44 (2): 147–76. https://doi.org/10.1177/0093650213514601.

Weber, Matthew, Katherine Ognyanova, and Allie Kosterich. 2017. “Imitation in the Quest to Survive: Lessons from News Media on the Early Web.” *International Journal of Communication* 11: 1–25.

Wojcik, Stefan, Avleen S. Bijral, Richard Johnston, Juan M. Lavista Ferres, Gary King, Ryan Kennedy, Alessandro Vespignani, and David Lazer. 2021. “Survey Data and Human Computation for Improved Flu Tracking.” *Nature Communications* 12 (1): 194. https://doi.org/10.1038/s41467-020-20206-z.

Woudenberg, T. J. van, K. E. Bevelander, W. J. Burk, C. R. Smit, L. Buijs, and M. Buijzen. 2020. “Comparing the Measurement of Different Social Networks: Peer Nominations, Online Communication, and Proximity Data.” *Network Science* 8 (1): 62–78. https://doi.org/10.1017/nws.2019.65.

Wrigley-Field, Elizabeth, Mathew V. Kiang, Alicia R. Riley, Magali Barbieri, Yea-Hung Chen, Kate A. Duchowny, Ellicott C. Matthay, et al. 2021. “Geographically Targeted COVID-19 Vaccination Is More Equitable and Averts More Deaths than Age-Based Thresholds Alone.” *Science Advances* 7 (40): 1–14. https://doi.org/10.1126/sciadv.abj2099.

Yancey, Antronette K., Alexander N. Ortega, and Shiriki K. Kumanyika. 2006. “Effective Recruitment and Retention of Minority Research Participants.” *Annual Review of Public Health* 27: 1–28. https://doi.org/10.1146/annurev.publhealth.27.021405.102113.

Yang, Tian, Sílvia Majó-Vázquez, Rasmus K. Nielsen, and Sandra González-Bailón. 2020. “Exposure to News Grows Less Fragmented with an Increase in Mobile Access.” *Proceedings of the National Academy of Sciences* 117 (46): 28678–83. https://doi.org/10.1073/pnas.2006089117.