HANDBOOK OF COMPUTATIONAL SOCIAL SCIENCE, VOLUME 1

Theory, Case Studies and Ethics

Edited by Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu, and Lars Lyberg

First published 2022

ISBN: 978-0-367-45653-5 (hbk)
ISBN: 978-0-367-45652-8 (pbk)
ISBN: 978-1-003-02458-3 (ebk)

7

DIGITAL TRACE DATA

Modes of data collection, applications, and errors at a glance

Florian Keusch and Frauke Kreuter

DOI: 10.4324/9781003024583-8

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

7

DIGITAL TRACE DATA

Modes of data collection, applications, and errors at a glance

Florian Keusch and Frauke Kreuter

Introduction

Digital traces, often defined as "records of activity (trace data) undertaken through an online information system (thus, digital)" (Howison, Wiggins, & Crowston, 2011, p. 769) or "behavioral residue [individuals leave] when they interact online" (Hinds & Joinson, 2018, p. 2), provide researchers with new opportunities for studying social and behavioral phenomena. These data come from a variety of technical systems, among others, business transaction systems, telecommunication networks, websites, social media platforms, smartphone apps, sensors built in wearable devices, and smart meters (Stier, Breuer, Siegers, & Thorson, 2019). Their analysis is a core part of computational social science (Edelmann, Wolff, Montagne, & Bail, 2020; Lazer et al., 2009). The excitement about digital trace data mainly stems from the fine-grained nature of the data that potentially allows researchers to observe individual and social behavior as well as changes in behavior at high frequencies and in real time. In addition, their measurement is non-intrusive; that is, the data collection happens without the observed person having to self-report. Removing human cognition and social interactions from the data collection process can mitigate their well-documented negative impacts on the quality of self-reports (e.g., Tourangeau, Rips, & Rasinski, 2000). However, the true potential of digital trace data to answer a broad range of social science research questions depends on the features of the specific type of data, that is, how they were collected and from whom. The original definition of digital trace data is limited to data that are found, that is, data created as a by-product of activities not stemming from a designed research instrument. We argue that digital traces can and should sometimes be collected in a designed way to afford researchers control over the data generating process and to expand the range of research questions that can be answered with these data.

Readers of this chapter will quickly realize that the use of digital trace data is in its early stages. We share the enthusiasm of many researchers to explore the capabilities of digital trace data, and enhance and systematize their collection. However, there is more research needed to tackle problems of privacy, quality assurance, and a good understanding of break-downs in the measurement process.

In this chapter, we introduce digital trace data and their use in the computational social sciences ("Use of Digital Trace Data to Study Social and Behavioral Phenomena"). In order to successfully use digital trace data, research goals need to be aligned with the available data, and

 DOI: 10.4324/9781003024583-8

researchers need to recognize that not all data are suitable to answer the most relevant questions ("What Is the Research Goal?"). Data quality also needs to be evaluated with the research goal in mind ("Quality Assessment – Quality Enhancement"). To ensure reproducibility and replicability, documentation of digital trace data collection and processing is necessary, and creating sufficient transparency might be even harder than it already is with traditional data sources ("Transparency and Reporting Needs").

Use of digital trace data to study social and behavioral phenomena

Digital trace data allow researchers to study a variety of social and behavioral phenomena and can be organized in a number of ways. Given the fast-paced development of digital technology and the concomitant emergence of novel forms of digital trace data, a mere taxonomy of the types of digital trace data (e.g., social media data, Internet search data, geolocation data from smartphones) might become outdated quickly. Instead, we organize this section along dimensions of their use in computational social science (i.e., what type of phenomenon is studied?) and the type of observation used when collecting the data (i.e., how obtrusive is the observation?). This broader perspective might help social researchers to detect new sources of digital trace data and assess properties of digital trace data that already exist and data sources that will emerge in the future.

Type of phenomenon to be studied

We use two dimensions to describe the types of phenomena that can be studied with digital trace data (see Figure 7.1). First, we differentiate between phenomena that pertain to *individual behavior* and those that represent *social interactions* involving multiple individuals. Second, we distinguish between digital and analog phenomena, building on and extending the classification of mobile sensing data by Harari, Müller, Aung, and Rentfrow (2017). Digital phenomena are types of behaviors and interactions that happen while using a digital device, such as browsing the Internet, posting a comment on a social media platform, or making a video call. These behaviors and interactions are inherently digital, as they could not happen without the use of digital technology. Analog phenomena are behaviors and social interactions that people encounter in their everyday lives and that existed well before the age of digital technology, including face-to-face communication, physical activity, mobility, and sleep. While the phenomena themselves happen without the use of digital technology, the ubiquity of smartphones, wearables, sensors, and other digital devices leaves a digital trace about them that researchers can leverage.

The combination of these two dimensions creates four broad categories of phenomena that can be measured using digital trace data: digital individual behavior (e.g., browsing the Internet, typing a query into an online search engine, using an app), analog individual behavior (e.g., sleeping, working out, doing chores), digital social interactions (e.g., video calling, text messaging), and analog social interactions (e.g., face-to-face conversations). While it is helpful to organize phenomena along these four categories, we acknowledge that there can be overlap between the groups. In particular, behaviors and interactions that used to be primarily analog have become increasingly digital over time. Consider, for example, driving; driving is inherently an analog individual behavior that does not necessarily require digital technology. However, increasingly, cars rely on a combination of traditional mechanics and digital technologies for navigation, safety, and autonomous driving (Horn & Kreuter, 2019), changing driving from a primarily analog behavior to a digital behavior in the near future.

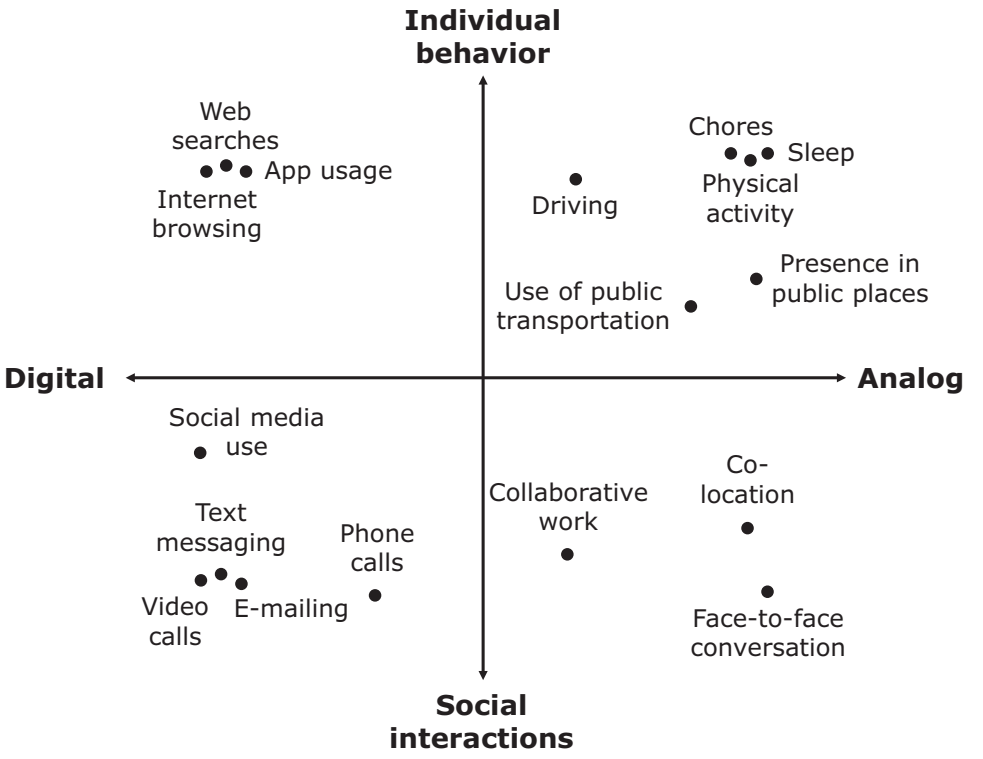

Figure 7.1 Examples of analog and digital behaviors and interactions that can be studied using digital trace data

Similarly, while individuals have always worked together on projects even without the help of digital technologies, collaborative work is increasingly done via platforms, such as Dropbox, Google Docs, and GitHub. Other phenomena can be considered as muddling the boundaries between individual behavior, social behavior, and social interaction. For example, while posting something on a social media platform is at first an individual behavior (in particular if the site is private and there are no followers), the post might trigger a conversation with other users, leading to social interactions. Figure 7.1 plots examples of analog and digital behaviors and interactions that can be measured using digital trace data in a two-dimensional space along our two dimensions.

One important piece that we will discuss in "What Is the Research Goal?" in more detail is the absence of digital trace data in all of these quadrants for certain people and certain behaviors through selective use of digital devices. It is very easy to get blindsided by the vast amount of data available and to overlook what is not there. Results of research projects can be easily biased.

Type of observation

Similar to traditional observational methods in the social sciences, the observation of aforementioned individual behaviors and social interactions using digital trace data can be more or less obtrusive, depending on how aware the individuals are of the fact that they are being observed and that their data are used for research. For example, a form of unobtrusive collection of digital

trace data happens when private companies utilize technologies such as cookies and browser fingerprinting to collect information about the browsing behavior of Internet users (Lerner, Simpson, Kohno, & Roesner, 2016). Data brokers (e.g., Acxiom, LexisNexis Risk Solutions, Experian) provide vast amounts of these data back to interested parties, mainly with the goal to infer user attributes (e.g., sociodemographics, personal and political interest) from the contents visited for targeted marketing and political campaigning (Duhigg, 2012; Kruschinski & Haller, 2017; Nickerson & Rogers, 2014).

As a consequence of the introduction of the EU General Data Protection Regulation (GDPR) in May 2018, website providers have started asking Internet users to agree to the terms and conditions and accept cookies upon entering their website. However, only a small fraction of users seem to be actively reading and understanding what information they are agreeing to share with the website and third parties (Obar & Oeldorf-Hirsch, 2020). While in some contexts, for example, when using an online shop such as Amazon, users might expect their data to be used for various purposes, in other cases, for example, when scientists and researchers use the platform ResearchGate[1] to share papers, users might be surprised about the amount of data that is collected about them and with whom they are shared.

Digital trace data can also be collected in an unobtrusive manner from social media platforms where users post comments, share content, and interact with each other. These data can usually be scraped or accessed via an application programming interface (API). While APIs are usually not primarily designed for a research purpose but for software systems to communicate with each other, social scientists have explored the use of, for example, Twitter data to study political communication (Jungherr, 2015); Reddit data to measure strength of attitudes on politics, immigration, gay rights, and climate change (Amaya, Bach, Keusch, & Kreuter, 2020); and Facebook data to study friendship networks (Cheng, Adamic, Kleinberg, & Leskovec, 2016; Ugander, Karrer, Backstrom, & Marlow, 2011). While posts, for example, on Twitter are public by default, only few users are aware that their tweets are used by researchers (Fiesler & Proferes, 2018).

Several other forms of unobtrusive digital trace data have been used to study behavior and other social phenomena, for example:

- Researchers have used aggregated data from online search engines and the queries users post there to study consumer trends (Vosen & Schmidt, 2011), tracking of disease outbreaks such as influenza (Ginsberg et al., 2009), tracking of economic crises (Jun, Yoo, & Choi, 2018), political polarization (Flaxman, Goel, & Rao, 2016), and migration (Böhme, Gröger, & Stöhr, 2020; Vicéns-Feliberty & Ricketts, 2016).
- Blumenstock, Cadamuro, and On (2015) used anonymized mobile phone metadata from cellular network operators to predict poverty and wealth in Africa.
- Göbel and Munzert (2018) studied how German politicians enhance and change their appearance over time based on traces of changes to biographies on the online encyclopedia Wikipedia.
- Edelman, Luca, and Svirsky (2017) used Airbnb postings to understand racial discrimination.
- The Billion Price Project scrapes online prices to measure consumption and inflation across countries (Cavallo & Rigobon, 2016).
- Przepiorka, Norbutas, and Corten (2017) studied reputation formation in a cryptomarket for illegal drugs using price and buyers' ratings data of finished transactions.
- Philpot, Liebst, Levine, Bernasco, and Lindegaard (2020) analyzed bystander behavior, that is, whether and how individuals intervene during an emergency when in the presence of others or alone, using footage from closed-circuit television (CCTV) in public spaces.

- Social epidemiologists increasingly use electronic health record data to study, for example, the impact of built and social environment, for example, poverty rates in certain geographic areas, on health outcomes (Adler, Glymour, & Fielding, 2016).
- Several large-scale projects have deployed connected environmental sensors (Internet of Things [IoT]), measuring, for example, temperature, humidity, air quality, noise levels, and traffic volume, in so-called "smart cities" allowing researchers access to urban measurements with greater spatial and temporal resolution (Benedict, Wayland, & Hagler, 2017; Catlett, Beckman, Sankaran, & Galvin, 2017; Di Sabatino, Buccolieri, & Kumar, 2018; English, Zhao, Brown, Catlett, & Cagney, 2020).

In contrast, some collection of digital trace data is much more obtrusive in that the individuals who produce the data are made explicitly aware of the fact that their data are used for research purposes. That is, they have to consent to the data collection and install a designated research app to their smartphone, download a meter and install it as a plugin to their Internet browser, or wear a sensor on their body. Smartphones in particular have become popular data collection tools among social and behavioral scientists (Harari et al., 2016; Link et al., 2014; Raento, Oulasvirta, & Eagle, 2009), because many users carry their phones around with them throughout the day, allowing for real-time, in situ data collection using the growing number of sensors built into these devices (see Figure 7.2). Using designated research apps, researchers can get access to log files that are automatically generated by a device's operating system, enabling the collection of information about the usage of the device for tasks like texting, making and receiving phone calls, browsing the Internet, and using other apps on smartphones (i.e., digital behaviors and interactions). These data allow researchers to study, among others, social interactions (e.g., Keusch, Bähr, Haas, Kreuter, & Trappmann, 2020c), and even infer personality based on how users interact with the smartphone and what apps they use (e.g., Stachl et al., 2020). The native sensors built into smartphones and other wearable devices enable the measurement of users' current situation and their behavior outside of the generic functions of the phone, where the device is merely present in a given context (i.e., analog interactions and behaviors). For example, researchers have collected information about smartphone users' location and movements via global navigation satellite systems (GNSS), Wi-Fi, and cellular positioning, proximity to others using Bluetooth, and physical activity through accelerometer data. In addition, a combination of sensors (e.g., microphone, light sensor, accelerometer) can be used to capture information about the smartphone's and – by extension – the participant's ambient environment, inferring frequency and duration of conversation and sleep (e.g., Wang et al., 2014), as well as levels of psychological stress (Adams et al., 2014).

To provide context to the passively collected sensor and log data, researchers often administer in-app survey questions that inquire about phenomena such as subjective states (e.g., mood, attitudes) that require self-report (Conrad & Keusch, 2018). This combined approach of self-report and passive measurement on smartphones has been used to study, among others, mobility patterns (Elevelt, Lugtig, & Toepoel, 2019; Lynch, Dumont, Greene, & Ehrlich, 2019; Scherpenzeel, 2017), the influence of physical surroundings and activity on psychological well-being and health (Goodspeed et al., 2018; Lathia, Sandstrom, Mascolo, & Rentfrow, 2017; MacKerron & Mourato, 2013; York Cornwell & Cagney, 2017), student well-being over the course of an academic term (Ben-Zeev, Scherer, Wang, Xie, & Campbell, 2015; Harari, Gosling et al., 2017; Wang et al., 2014), integration efforts of refugees (Keusch et al., 2019), job search of men recently released from prison (Sugie, 2018; Sugie & Lens, 2017), the effects of unemployment on daily life (Kreuter, Haas, Keusch, Bähr, & Trappmann, 2020), and how students interact with

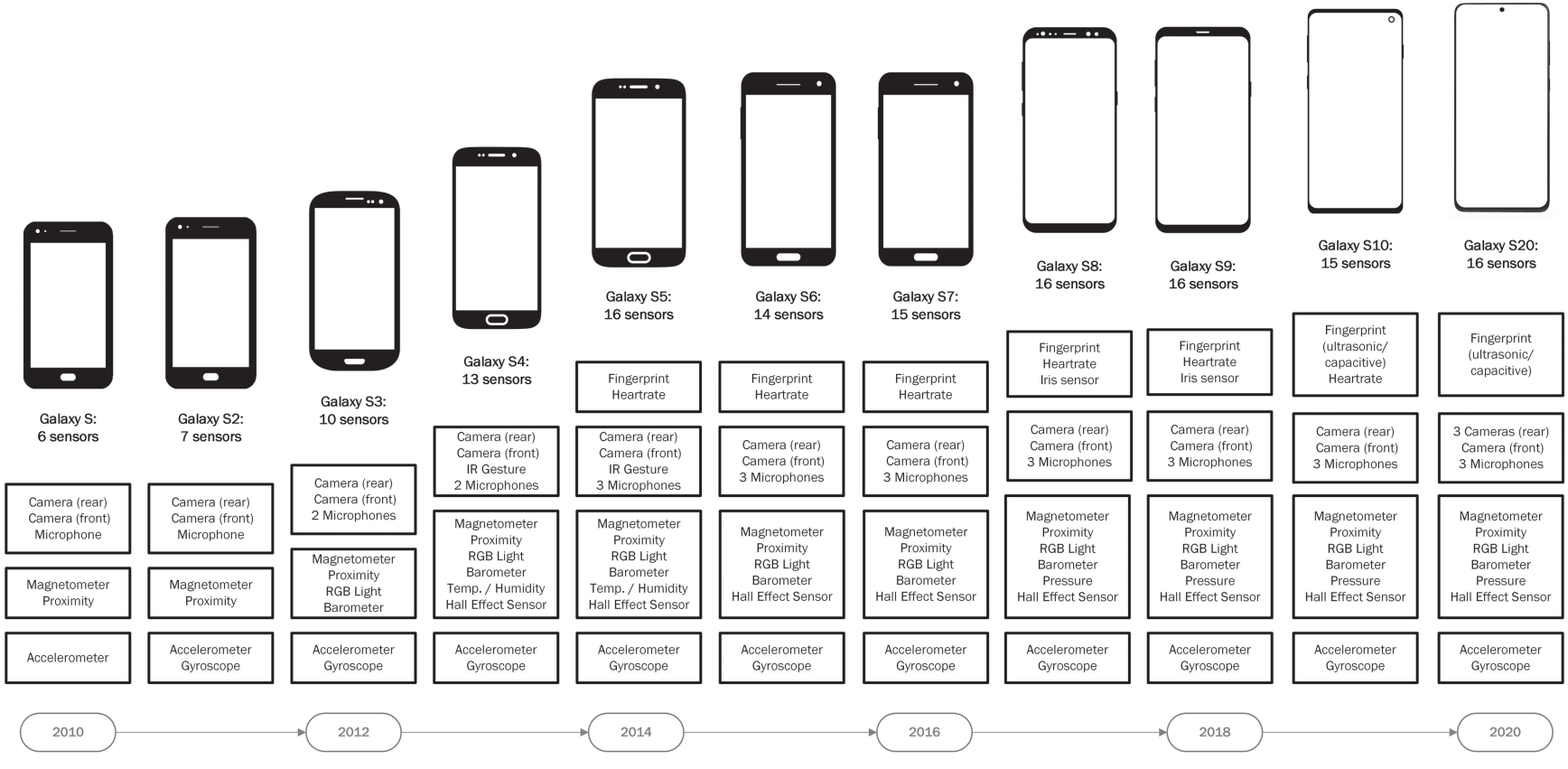

Figure 7.2 The growing number of native smartphone sensors provides researchers access to even more digital trace data (Struminskaya, Lugtig, Keusch, & Höhne, 2020)

each other across a variety of communication channels (Sapiezynski, Stopczynski, Lassen, & Lehmann, 2019; Stopczynski et al., 2014).

Digital traces can also be collected from wearable devices such as wrist- or waist-worn trackers that measure physical activity and, depending on the device type, additional information such as heart rate and geolocation. Some studies have recruited existing users of consumer-grade fitness trackers (e.g., Fitbit, Garmin) and smartwatches (e.g., Apple Watch) to share their data with researchers (Ajana, 2018). For example, over 500,000 German volunteers donated data collected from their fitness wristbands and smartwatches to the Robert Koch Institute (RKI) during the COVID-19 pandemic.[2] Some studies have used platforms such as Fitabase[3] to get access to the wearable data of people recruited into their studies (Phillips & Johnson, 2017; Stück, Hallgrímsson, Ver Steeg, Epasto, & Foschini, 2017). In population studies, another option is to equip all participants with the same research-grade wearable device (e.g., Actigraph, Geneactive) and then collect the devices at the end of the field period (Harris, Owen, Victor, Adams, & Cook, 2009; Kapteyn et al., 2018; Troiano et al., 2008).

Another approach of obtrusive digital terrace data collection is the use of online tracking applications ("meters") that users need to actively install to their Internet browsers and/or mobile devices to allow the collection of browsing histories and app usage. This approach allows the researcher to trace individual online behavior, for example, news consumption via social media websites (Scharkow, Mangold, Stier, & Breuer, 2020), across time. Linking behavioral meter data with self-reports from web surveys allows researchers to study, for example, the relationship between passively measured online news consumption and self-reported voting behavior (Bach et al., 2019; Guess et al., 2020) or online news consumption and political interest (Möller, van de Velde, Merten, & Puschmann, 2019).

Depending on the design of the study, the measurement might be perceived as being less obtrusive over time because participants forget or get used to the presence of the measurement device. If the only active task for the participant is, for example, to download a research app or a meter that collects data in the background on their smartphone or Internet browser, then participants might soon forget that their behavior is even being observed. However, wearing a research-grade device on the body will potentially serve as a constant reminder that the individual is part of a study. Similarly, if the study design involves a combination of passive measurement of digital traces and repeated collection of self-reports (e.g., ecological momentary assessment [EMA] questions multiple times a day), it will probably make participants more aware of the observational part of the study.

What is the research goal?

Given that digital trace data are often collected incidentally and are reused for research purposes, it helps to take a step back and examine the goal of the research. Different research questions place different requirements on data and might create the need to go above and beyond readily available (digital trace) data. To simplify the discussion, we differentiate between three very general research goals irrespective of the data types: *description*, *causation*, and *prediction*. Of course, any given research project might combine several of these aspects or include variants not spelled out here in detail. This is not the first time we have discussed these issues. Readers interested in our presentations in other contexts can refer to Foster, Ghani, Jarmin, Kreuter, and Lane (2020) for general big data methods and privacy topics and Kohler, Kreuter, and Stuart (2019) for more detailed thoughts on causality and prediction.

When social scientists aim at describing the state of the society or a special population within the society, they typically seek to report a mean, a median, or a graphical distribution of

a variable of interest. A first decision has to be made when interpreting the descriptive statistic. Researchers need to be clear if their aim is to describe a population or only report on the data at hand. In the case of a census data collection, where by definition all units of the population are covered, the two aspects overlap. In all other cases, an extra step is needed, which is often difficult when dealing with digital trace data. Say, for example, a researcher is scraping job postings in Germany in the first week of May in a given year. She can then describe the percentage of data scientists sought in her scraped set of posts and only in those. Such restrictions need to be communicated clearly when presenting and publishing the results. A much harder task is to estimate the percentage of data scientists searched for by all German companies in that year with such data at hand.

When data are not available for the entire population but accessed via samples, the goal of inferring to the population is solved by taking a sample with *known* selection probabilities and ensuring that everybody from the population of interest has a *positive selection probability*. Doing so requires a sampling frame that ideally covers the entire population. In the scraping example, this could be achieved by having a complete list of companies and being able to acquire all the job postings of a sample of companies selected from such a frame. In such a setting, standard errors would be used to express uncertainty due to the sampling procedure. When totals are reported (i.e., the absolute number of postings for data scientists), getting the selection probabilities right is particularly important (Lohr, 2009). In practice, one will often face a situation in which not all elements in the sample (companies) post all the data scientist positions, or, if the method of data collection is a survey, they refuse to respond to the survey request. The survey methodology literature has decades of publications on this topic and suggestions for adjustments for situations in which the mechanism leading to the missing values is well understood (see, for example, Bethlehem, Cobben, & Schouten, 2011; Groves & Couper, 1998; Schnell, 1997; Valliant, Dever, & Kreuter, 2018; Willimack, Nichols, Elizabeth, & Sudman, 2002). Starting with a probability sample has the strong advantage that sampling errors can be estimated; nonresponse error can be adjusted for known covariates; and with sufficient information on the sampling frame, the coverage errors are also known.

Of course, even if sampling and nonresponse error are adjusted for, assumptions about the measurement process still have to be made. Mislabeling might occur, for example, if a job is classified as "data scientist" even if the activities do not match the label (false positive) or, conversely, if a job entails what is commonly understood as data science but is not explicitly labeled as such in the ad (false negative).

The issue of overinterpreting the results is not new to digital trace data. We saw and still see this happening in the context of traditional data collections via (sample) surveys. The classic example where a data generating process was not understood or ignored is the Literary Digest poll, which incorrectly called the 1936 election (Squire, 1988). The Literary Digest went for volume and overlooked issues of selective access to phones and magazine subscriptions when assembling its mailing lists. Many of the data collection efforts in the COVID-19 pandemic show a similar tendency (see Kohler, 2020 and the associated special issue). Likewise, using Twitter data as a source to identify areas in need of support after natural disasters (e.g., hurricanes) my misguide policy makers. Resources and attention would likely flow towards the younger population, people with easy Internet access, or those generally well connected (Shelton, Poorthuis, Graham, & Zook, 2014).

When the main research goal is the establishment of a *causal relationship*, the situation is a bit different (Kohler et al., 2019). If a treatment is applied with a proper randomized experiment or a strong non-experimental study design, then statements about such causal relationship can be made for anyone who had a chance to be treated. Knowing the selection probability of the

cases is then much less important, though it is very important that all elements have a positive selection probability to be assigned to the treatment and control conditions (Imbens & Rubin, 2015).

One example of an experiment done in a controlled fashion with digital trace data as the outcome is the Facebook emotional contagion study (Kramer, Guillory, & Hancock, 2014), where the number and type of posts seen on the users' wall were manipulated for a random sample of Facebook users. Differences in posting behavior (i.e., number of posts, sentiment) between users who were exposed to the treatment and those who were not can be interpreted as the causal effect of the treatment. Similarly, in a study with 193 volunteer Japanese smartphone owners who downloaded a research app, a random half of participants received on-screen reminders designed to stimulate interaction with communication weak ties during the two-month study period. The researchers compared the average number of phone calls, text messages, and emails from the smartphone log files to estimate the causal effect of the reminder messages (Kobayashi, Boase, Suzuki, & Suzuki, 2015).

Interesting causal claims can also be made in quasi-experimental settings where external shocks create the treatments, and regression discontinuity or similar designs can be used. During the COVID-19 pandemic, digital trace data from mobile devices were used to assess the (causal) effects of lock-down restrictions or other interventions designed to slow the spread of the virus. However, in the digital trace data setting – just as with traditional data collection – detailed knowledge about what is measured always needs to be available to interpret a "treatment" correctly. The Google mobility data in Figure 7.3 show how difficult it can be to differential signal and noise and how much pre-processing and data cleaning is still necessary.

While these examples have internal validity (albeit to different degrees), they lack external validity: inference to the population at large is not possible without further assumptions. In the Google mobility data example, not all units in the population have mobile devices that feed into this analysis. Thus any causal claim made from the data is generalizing to

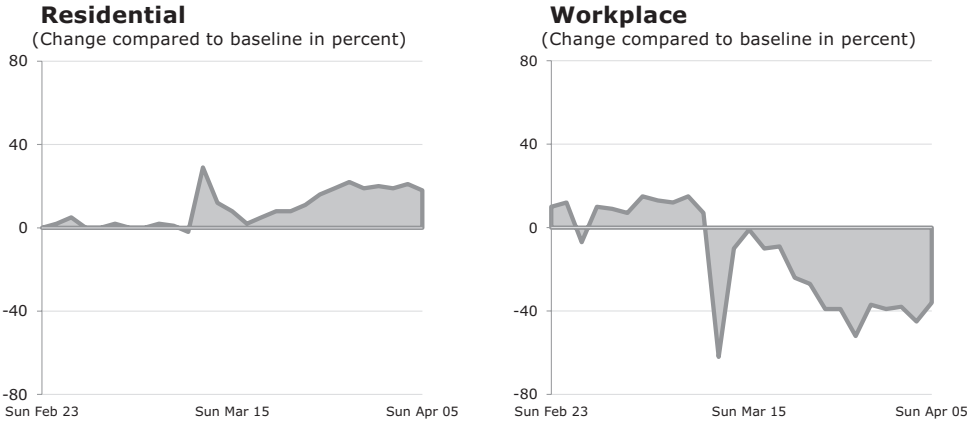

Figure 7.3 Google mobility data. Google points out "March 12 was a widely celebrated public holiday in this region. Workplace and residential changes are a little different from the community's response to COVID-19 but give an idea of the scale of the change. You'll need to apply your local knowledge, but holidays provide a very specific point of comparison" (Adapted from https://support.google.com/covid19-mobility/answer/9825414?hl=en-GB Accessed: 7/21/2020)

the population without mobile devices. In general, should the research question involve a causal claim or causal inference about a larger population than what is covered by the data, then the same issues arise as in the descriptive setting described earlier. The causal relationship will only hold if the causal effect is the same for the people that had the chance to be randomized to the treatment and those that did not (effect homogeneity) or, in the case of the Google mobility data, those for whom measures can be obtained and those for whom they cannot.

Not being able to randomize into treatment and control on a random sample of the population is a known problem in medical research. There the assumption of effect homogeneity has often been made. More recently, medical and public health researchers have increasingly scrutinized these assumptions and created statistical methods that help to generalize causal effects to the general population (DuGoff, Schuler, & Stuart, 2014). We would not be surprised if similar efforts will take place for digital trace data, with researchers thinking hard about the behavior of people not contributing to datasets (at all or at the same rate).

Prediction tasks are common in a data science pipeline when dealing with digital trace data. While social scientists are usually more interested in description of a population or causal effects, in doing either, they use predictions when specific measurements cannot be designed for the covariates of interest. Examples are Fitbit converting accelerometer sensor data into steps, using algorithms that differentiate between different motions, settings, and movements;[4] predicting voting behavior based on online news consumption (Bach et al., 2019); predicting personality traits based on smartphone usage (Stachl et al., 2020); or predicting the level of gentrification in a neighborhoods based on data about business activities from Yelp (Glaeser, Kim, & Luca, 2018). The reason prediction models are popular for such tasks is that they "often do not require specific prior knowledge about the functional form of the relationship under study and are able to adapt to complex non-linear and non-additive interrelations between the outcome and its predictors while focusing specifically on prediction performance" (Kern, Klausch, & Kreuter, 2019, p. 73).

Prediction tasks can work very well when large amounts of data are available, ideally for the exact same situation, person, or setting, and the prediction is made in close temporal proximity to the observation. However, the further the predicted outcome is from the data the prediction is based on, both temporally (e.g., predicting a "like" or "click" three months down the road) and conceptually (e.g., predicting an election outcome) or both (e.g., predicting an election outcome three months down the road), the lower the prediction success.

The potential for (massive amounts) of digital trace data to be used in prediction tasks is undeniable due to its unprecedented scope and variety. However, knowing who is covered by the data, for which settings, and which circumstances or time frames is just as important here as it is in the description and causal inference setting. Without knowing the ins and outs of the data generating process, there are real risks of biases due to unknown or unobserved systematic selection with respect to a given research question. This will increasingly be an issue with automated decision systems used in the societal context. For example, while predictive policing can be used to allocate police resources, it could harm society if the data do not represent the population at large and predictions are biased (Rodolfa, Saleiro, & Ghani, 2020).

Quality assessment – quality enhancement

The quality of digital trace data is relative to the research goal. Or, to put it differently, without knowing the research goal, it is only in very specific circumstances possible to make an overall

claim about the quality of the data – or assess the fit of the data for a given research goal. Several frameworks exist that can help characterize data quality (see summaries in Christen, 2012; National Academies of Sciences, 2017). Typical elements are *accuracy, completeness, consistency, timeliness*, and *accessibility*. Ultimately, a researcher has to ask herself whether the data can support the inference she is planning to make. This situation is no different for digital trace data than for any other data source, for example, more traditional survey data (Schnell, Hill, & Esser, 2018).

Klingwort and Schnell (2020) show the difficulty in using digital trace data related to COVID-19 data collection. They convincingly question the use of the volunteer Fitbit app data donations in the previously described Robert Koch Institut effort. Not only was the number of people installing the app insufficient to cover all the variability in the population (as of early May 2020), but it also suffered from sources of nonresponse due to lack of knowledge about the app, privacy concerns, willingness to participate, and regular device use, as well as sources of coverage error due to owning the appropriate device and having the necessary technical skills (see Figure 1 in Klingwort & Schnell, 2020).

In addition, digital trace data can be subject to quality challenges less common in traditional data sources. For example, easily overlooked are problems of de-duplication, with units appearing in found or donated digital trace data multiple times, or records representing multiple units without being noticed as such. Schober, Pasek, Guggenheim, Lampe, and Conrad (2016) describe the former in their assessment of the use of social media data for social research and state that individual posts can represent. The latter is a problem that easily appears when devices such as computers, tablets, or smartphones are used by multiple people (Hang, von Zezschwitz, De Luca, & Hussmann, 2012; Matthews et al., 2016; Silver et al., 2019) and likely occurs more often when data are collected via smart devices in households. Another problem in analyzing social media data is the presence of bots that would be treated as human posters when computing the summary statistics. Data from search engines also illustrate novel challenges to data quality in digital trace data. Search engines might change in terms of how they are designed, who uses them, and how users engage with them over time in ways that are out of the researcher's control (Lazer, Kennedy, King, & Vespignani, 2014).

In a more abstract way, a *design feature* is the possibility to gain access to data in an *organic, found, or ready-made* way (Groves, 2011; Japec et al., 2015; Salganik, 2017). While this distinction between data found in the wild and data collected by design highlights an important feature, the data collection itself (say digital trace vs. survey questions) is independent of the found vs. design distinction.

Designed measurement of digital traces would mean participants are selected into the study and a specific technology such as an online meter, a mobile app, or a wearable specifically designed for a research study is used for data collection. Found data are byproducts of interactions with the world that leave digital traces; or, put differently, they arise organically. With found data, researchers have no control over who provides the data and how. A typical example for found data would be credit card transactions, postings in online search engines, or interactions with and on social media.

In practice, we often see a mix of designed and organic data. Sometimes, when researchers collaborate closely with the primary entities that collect the data, they might have the chance to provide input into what information is captured. For example when working closely with government agencies, researchers might have some input into how the measurement is taken (i.e., fields on a form of digital health records or unemployment insurance notices). Likewise, one can select respondents carefully (design the sample), but collect "found/organic" data that were not designed for the purpose of the research study but metered through already existing devices. An example is the IAB-SMART study (Kreuter et al., 2020), where existing

measurement instruments in smartphones (accelerometer, pedometer, GPS) are used, but measurements are taken at specific intervals or in response to an event, bringing a design element into the mix. One major advantage of collecting data through designated research apps is that this allows researchers to specifically design all aspects of the data collection process (e.g., field period, participants' characteristics, particular sensors used) with a specific research question in mind. The controlled environment, potentially in conjunction with a probability sample, allows the researcher to not only to assess coverage (Keusch, Bähr, Haas, Kreuter, & Trappmann, 2020a), nonresponse (Keusch, Bähr, Haas, Kreuter, & Trappmann, 2020b), and measurement error (Bähr, Haas, Keusch, Kreuter, & Trappmann, 2020) but also to address these issues through weighting techniques known from survey research that would allow inference of the results to a larger population.

It can be useful to ask prior to any applied research the following questions:

- Which population is covered by the data? If not, which groups are missing? Is it even known which groups are missing?
- Can the sample represent the population? If not, are certain units entirely missing, or are they just not represented in the proportion needed? Are the reasons they are missing known, and can they be measured (in which case weighting might be an option)?
- Do I know what the measurements represent? Or do I need to generate new features from the digital trace data to answer the research question? How accurate are the attribute values in the data? Are all variables needed for the analysis in the data?
- How timely are the data?
- Are there data available that can be used to assess the quality of the generated features on a small scale? Can the small-scale assessment be generalized to the entire data?

For computational social science to be successful in using digital trace data, we foresee that in most (if not all) cases, data from different sources need to be combined, either to overcome the problem of unknown populations of inference or to overcome the problem of missing covariates and overall unclear measurement properties (Christen, Ranbaduge, & Schnell, 2020; Couper, 2013; Schnell, 2019).

Transparency and reporting needs

As tempting as the use of (easily available) digital trace data is, one has to keep in mind that there is a long path between the raw data and insights derived from the data. Because many of the digital trace data are by-products of processes with a purpose different from the researcher's intent, many pre-processing steps are needed before the analyses can begin. In a complex fast-moving world, where platforms and processes change, digital traces will by definition be inconsistent and noisy (Foster et al., 2020) and be filled with missing data, not the very least because of different terms and conditions for data access and use that the different platforms exhibit (Amaya, Bach, Keusch, & Kreuter, 2019).

How sensitive results are to such preprocessing steps and the accompanying decisions was demonstrated by Conrad et al. (2018) for studies trying to create alternative indicators for consumer confidence and consumer sentiment from Twitter data. The volatility of the results raised skepticism among the authors and prompted them to call for best practices in generating features and documenting results when using Twitter data for these purposes. Such desire for best practices and standards in reporting can be seen in many other communities as well, very prominently among statistical agencies around the world. The United National Statistics Division[5]

lists principles governing international statistical activities, including a call for transparency of "concepts, definitions, classifications, sources, methods and procedures employed". Growing adoption of FAIR data principles – Findability, Accessibility, Interoperability, and Reusability – by funding agencies, journals, and research organizations further increases the need to acquire sufficient information about the data generating process, as well as subsequent steps in preprocessing the data. It is important to realize that FAIR principles also apply to algorithms, tools, and workflows generating the analytic dataset, not just to the raw data or, in our case, the raw digital traces.[6] As Wilkinson et al. (2016) state, "all scholarly digital research objects – from data to analytical pipelines – benefit from application of these principles, since all components of the research process must be available to ensure transparency, reproducibility, and reusability" (p. 1).

This said, researchers intending to use digital trace data should be aware that even if data collection is cheap, there are substantial costs associated with cleaning, curating, standardizing, integrating, and using the new types of data (Foster et al., 2020). Novices in using digital trace data might benefit from reading Amaya et al. (2019) to get a sense of challenges and opportunities when working with digital trace data from platforms, in this case Reddit. While specific to Reddit, the paper lists types of information one may seek to acquire prior to conducting a project that uses any type of social media data.

Conclusion

We are, without a doubt, excited about the possibilities digital trace data provide to social science research. The direct and often unobtrusive observation of individual behaviors and social interactions through digital systems produces data in breadth and depth that cannot be generated using traditional methods. However, for digital trace data to become a mainstream data source, there is still a long way to go. The initial hype has leveled off, and more and more research papers appear showing the challenges and limits of using digital trace data. At the same time, research studies that use clever designs to combine multiple data sources are on the rise.

The combination of multiple sources is not without risk. Multiple streams of data from different sources can create detailed profiles of users' habits, demographics, or well-being that carry the risk of unintentionally de-identifying previously anonymous data providers (Bender, Kreuter, Jarmin, & Lane, 2020; Deursen & Mossberger, 2018). We are hopeful that the parallel efforts going on right now with respect to privacy preserving record linkage (Christen et al., 2020) and encrypted computing (Goroff, 2015) will help mitigate those risks. While both areas are heavily dominated by computer scientists and statisticians, we encourage social scientists to inject themselves into this discussion so that the solutions work not just theoretically but also in practice (see Oberski and Kreuter, 2020, for the controversy around the use of differential privacy).

Whether multiple data sources are combined or single sources of digital traces are used, ethical challenges arise when the digital traces are the result of organic processes with a different original purpose (*found data*). As Helen Nissenbaum (2018) clearly lays out in her framework of contextual integrity, one cannot or should not ignore the question of the appropriateness of data flows. Appropriateness is a function of conformity with contextual informational norms. To give a brief example: A bouncer at a nightclub might see a woman's address as he checks her age to allow entrance into the club. If he later shows up at her house using the piece of information acquired during his job, he violates contextual informational norms. Re-purposing of digital trace data can violate contextual informational norms in similar ways. While the unanticipated secondary use constitutes the "crown jewels" of passively collected digital trace data (Tene & Polonetsky, 2013), users are increasingly concerned about the privacy of their data and how

much they can control how their personal information is used (Auxiere et al., 2019). For the researcher, the use creates a challenge in how to balance the risk to the participants with the utility of the collected data, the so-called privacy-utility trade-off (Bender et al., 2020).

Notes

1 At the time of writing this chapter, ResearchGate asks the users for permission to share their personal data (e.g., IP address, cookie identifiers) with almost 500 external partners (www.researchgate.net/privacy-policy, June 30, 2020).
2 https://corona-datenspende.de/science/en/
3 https://www.fitabase.com/
4 https://help.fitbit.com/articles/en_US/Help_article/1136
5 https://unstats.un.org/unsd/methods/statorg/Principles_stat_activities/principles_stat_activities.asp
6 Including code leading to data as well as metadata, data describing the data, has already been part of the data management plan requirements of the U.S. National Science Foundation, for example. https://nsf.gov/eng/general/ENG_DMP_Policy.pdf

References

Adams, P., Rabbi, M., Rahman, T., Matthews, M., Voida, A., Gay, G., Choudhury, T., & Voida, S. (2014). *Towards personal stress informatics: Comparing minimally invasive techniques for measuring daily stress in the wild*. Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare, pp. 72–79. https://doi.org/10.4108/icst.pervasivehealth.2014.254959

Adler, N. E., Glymour, M. M., & Fielding, J. (2016). Addressing social determinants of health and health inequalities. *JAMA*, *316*(16), 1641–1642. https://doi.org/10.1001/jama.2016.14058

Ajana, B. (2018). Communal self-tracking: Data philanthropy, solidarity and privacy. In B. Ajana (Ed.), *Self-tracking: Empirical and philosophical investigations* (pp. 125–141). Springer International Publishing. https://doi.org/10.1007/978-3-319-65379-2_9

Amaya, A., Bach, R. L., Keusch, F., & Kreuter, F. (2019). New data sources in social science research: Things to know before working with Reddit data. *Social Science Computer Review*. https://doi.org/10.1177/0894439319893305

Amaya, A., Bach, R. L., Keusch, F., & Kreuter, F. (2020). Measuring attitude strength in social media data. In C. A. Hill, P. Biemer, T. D. Buskirk, L. Japec, A. Kirchner, S. Kolenikov, & L. E. Lyberg (Eds.), *Big data meets survey science* (pp. 163–192). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118976357.ch5

Auxiere, B., Rainie, L., Anderson, M., Perrin, A., Kumar, M., & Turner, E. (2019). *Americans and privacy: Concerned, confused and feeling lack of control over their personal information*. PEW Research Center. Retrieved from www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/

Bach, R. L., Kern, C., Amaya, A., Keusch, F., Kreuter, F., Hecht, J., & Heinemann, J. (2019). Predicting voting behavior using digital trace data. *Social Science Computer Review*. https://doi.org/10.1177/0894439319882896

Bähr, S., Haas, G.-C., Keusch, F., Kreuter, F., & Trappmann, M. (2020). Measurement quality in mobile geolocation sensor data. *Social Science Computer Review*. https://doi.org/10.1177/0894439320944118

Bender, S., Kreuter, F., Jarmin, R. S., & Lane, J. (2020). Privacy and confidentiality. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science* (2nd ed.). Chapman and Hall and CRC Press. https://textbook.coleridgeinitiative.org/

Benedict, K., Wayland, R., & Hagler, G. (2017, November). Characterizing air quality in a rapidly changing world. *Air and Waste Management Association's Magazine for Environmental Managers*. Retrieved from https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&dirEntryId=338745

Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., & Campbell, A. T. (2015). Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal*, *38*(3), 218–226. https://doi.org/10.1037/prj0000130

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys*. John Wiley & Sons.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420

Böhme, M. H., Gröger, A., & Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics*, *142*, 102347. https://doi.org/10.1016/j.jdeveco.2019.04.002

Catlett, C. E., Beckman, P. H., Sankaran, R., & Galvin, K. K. (2017). *Array of things: A scientific research instrument in the public way: Platform design and early lessons learned*. Proceedings of the 2nd International Workshop on Science of Smart City Operations and Platforms Engineering, pp. 26–33. https://doi.org/10.1145/3063386.3063771

Cavallo, A., & Rigobon, R. (2016). The billion prices project: Using online prices for measurement and research. *Journal of Economic Perspectives*, *30*(2), 151–178. https://doi.org/10.1257/jep.30.2.151

Cheng, J., Adamic, L. A., Kleinberg, J. M., & Leskovec, J. (2016). *Do cascades recur?* Proceedings of the 25th International Conference on World Wide Web – WWW'16, pp. 671–681. https://doi.org/10.1145/2872427.2882993

Christen, P. (2012). *Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer-Verlag. https://doi.org/10.1007/978-3-642-31164-2

Christen, P., Ranbaduge, P., & Schnell, R. (2020). *Linking sensitive data: Methods and techniques for practical privacy-preserving information sharing*. Springer.

Conrad, F. G., Gagnon-Barsch, J., Ferg, R., Hou, E., Pasek, J., & Schober, M. F. (2018, October 25). *Social media as an alternative to surveys of opinions about the economy*. BigSurv18.

Conrad, F. G., & Keusch, F. (2018, October 25). *Emergent issues in the combined collection of self-reports and passive data using smartphones*. BigSurv18, Barcelona, Spain.

Couper, M. P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods*, *7*(3), 145–156. https://doi.org/10.18148/srm/2013.v7i3.5751

Deursen, A. J. A. M. van, & Mossberger, K. (2018). Any thing for anyone? A new digital divide in internet-of-things skills. *Policy & Internet*, *10*(2), 122–140. https://doi.org/10.1002/poi3.171

Di Sabatino, S., Buccolieri, R., & Kumar, P. (2018). Spatial distribution of air pollutants in cities. In F. Capello & A. V. Gaddi (Eds.), *Clinical handbook of air pollution-related diseases* (pp. 75–95). Springer International Publishing. https://doi.org/10.1007/978-3-319-62731-1_5

DuGoff, E. H., Schuler, M., & Stuart, E. A. (2014). Generalizing observational study results: Applying propensity score methods to complex surveys. *Health Services Research*, *49*(1), 284–303. https://doi.org/10.1111/1475-6773.12090

Duhigg, C. (2012, October 13). Campaigns mine personal lives to get out vote – The New York times. *New York Times*. Retrieved from www.nytimes.com/2012/10/14/us/politics/campaigns-mine-personal-lives-to-get-out-vote.html

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics*, *9*(2), 1–22. https://doi.org/10.1257/app.20160213

Edelmann, A., Wolff, T., Montagne, D., & Bail, C. A. (2020). Computational social science and sociology. *Annual Review of Sociology*, *46*(1). https://doi.org/10.1146/annurev-soc-121919-054621

Elevelt, A., Lugtig, P., & Toepoel, V. (2019). Doing a time use survey on smartphones only: What factors predict nonresponse at different stages of the survey process? *Survey Research Methods*, *13*(2), 195–213. https://doi.org/10.18148/srm/2019.v13i2.7385

English, N., Zhao, C., Brown, K. L., Catlett, C., & Cagney, K. (2020). Making sense of sensor data: How local environmental conditions add value to social science research. *Social Science Computer Review*. https://doi.org/10.1177/0894439320920601

Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media + Society*, *4*(1). https://doi.org/10.1177/2056305118763366

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, *80*(S1), 298–320. https://doi.org/10.1093/poq/nfw006

Foster, I., Ghani, R., Jarmin, R. S., Kreuter, F., & Lane, J. (Eds.). (2020). *Big data and social science* (2nd ed.). Chapman and Hall and CRC Press. https://textbook.coleridgeinitiative.org/

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014. https://doi.org/10.1038/nature07634

Glaeser, E. L., Kim, H., & Luca, M. (2018). Nowcasting gentrification: Using Yelp data to quantify neighborhood change. *AEA Papers and Proceedings*, *108*, 77–82. https://doi.org/10.1257/pandp.20181034

Göbel, S., & Munzert, S. (2018). Political advertising on the Wikipedia marketplace of information. *Social Science Computer Review*, *36*(2), 157–175. https://doi.org/10.1177/0894439317703579

Goodspeed, R., Yan, X., Hardy, J., Vydiswaran, V. V., Berrocal, V. J., Clarke, P., . . . Veinot, T. (2018). Comparing the data quality of global positioning system devices and mobile phones for assessing relationships between place, mobility, and health: Field study. *JMIR MHealth and UHealth*, *6*(8), e168. https://doi.org/10.2196/mhealth.9771

Goroff, D. L. (2015). Balancing privacy versus accuracy in research protocols. *Science*, *347*(6221), 479–480.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5), 861–871. https://doi.org/10.1093/poq/nfr057

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in household interview surveys*. John Wiley & Sons.

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*, *4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

Hang, A., von Zezschwitz, E., De Luca, A., & Hussmann, H. (2012). *Too much information! User attitudes towards smartphone sharing*. Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design, pp. 284–287. https://doi.org/10.1145/2399016.2399061

Harari, G. M., Gosling, S. D., Wang, R., Chen, F., Chen, Z., & Campbell, A. T. (2017). Patterns of behavior change in students over an academic term: A preliminary study of activity and sociability behaviors using smartphone sensing methods. *Computers in Human Behavior*, *67*, 129–138. https://doi.org/10.1016/j.chb.2016.10.027

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science*, *11*(6), 838–854. https://doi.org/10.1177/1745691616650285

Harari, G. M., Müller, S. R., Aung, M. S., & Rentfrow, P. J. (2017). Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences*, *18*, 83–90. https://doi.org/10.1016/j.cobeha.2017.07.018

Harris, T. J., Owen, C. G., Victor, C. R., Adams, R., & Cook, D. G. (2009). What factors are associated with physical activity in older people, assessed objectively by accelerometry? *British Journal of Sports Medicine*, *43*(6), 442–450. https://doi.org/10.1136/bjsm.2008.048033

Hinds, J., & Joinson, A. N. (2018). What demographic attributes do our digital footprints reveal? A systematic review. *PLoS One*, *13*(11), e0207112. https://doi.org/10.1371/journal.pone.0207112

Horn, C., & Kreuter, F. (2019). *Die digitale Herausforderung: Tipping Points, die Ihr Unternehmen verändern werden* (1. Auflage 2020). Haufe.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, *12*(12), 768–797. https://doi.org/10.17705/1jais.00282

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., . . . Usher, A. (2015). Big data in survey research AAPOR task force report. *Public Opinion Quarterly*, *79*(4), 839–880. https://doi.org/10.1093/poq/nfv039

Jun, S.-P., Yoo, H. S., & Choi, S. (2018). Ten years of research change using Google trends: From the perspective of big data utilizations and applications. *Technological Forecasting and Social Change*, *130*, 69–87. https://doi.org/10.1016/j.techfore.2017.11.009

Jungherr, A. (2015). *Analyzing political communication with digital trace data: The role of Twitter messages in social science research*. Springer.

Kapteyn, A., Banks, J., Hamer, M., Smith, J. P., Steptoe, A., Soest, A. van, . . . Wah, S. H. (2018). What they say and what they do: Comparing physical activity across the USA, England and the Netherlands. *Journal of Epidemiol Community Health*, *72*(6), 471–476. https://doi.org/10.1136/jech-2017-209703

Kern, C., Klausch, T., & Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, *13*(1), 73–93. https://doi.org/10.18148/srm/2019.v1i1.7395

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., & Trappmann, M. (2020a). Coverage error in data collection combining mobile surveys with passive measurement using apps: Data from a German national survey: *Sociological Methods & Research*. https://doi.org/10.1177/0049124120914924

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., & Trappmann, M. (2020b, June 11). *Participation rates and bias in a smartphone study collecting self-reports and passive mobile measurements using a research app*. AAPOR 75th Annual Conference, Virtula Conference.

Keusch, F., Bähr, S., Haas, G.-C., Kreuter, F., & Trappmann, M. (2020c, July 17). *Social networks on smartphones. Congruence of online and offline networks and their effect on labor market outcomes*. 6th International Conference on Computational Social Science (IC^2C^2), Virtual Conference.

Keusch, F., Leonard, M. M., Sajons, C., & Steiner, S. (2019). Using smartphone technology for research on refugees: Evidence from Germany. *Sociological Methods & Research*. https://doi.org/10.1177/0049124119852377

Klingwort, J., & Schnell, R. (2020). Critical limitations of digital epidemiology: *Survey Research Methods*, *14*(2), 95–101. https://doi.org/10.18148/srm/2020.v14i2.7726

Kobayashi, T., Boase, J., Suzuki, T., & Suzuki, T. (2015). Emerging from the cocoon? Revisiting the tele-cocooning hypothesis in the smartphone era. *Journal of Computer-Mediated Communication*, *20*(3), 330–345. https://doi.org/10.1111/jcc4.12116

Kohler, U. (2020). Survey research methods during the COVID-19 crisis. *Survey Research Methods*, *14*(2), 93–94. https://doi.org/10.18148/srm/2020.v14i2.7769

Kohler, U., Kreuter, F., & Stuart, E. A. (2019). Nonprobability sampling and causal analysis. *Annual Review of Statistics and Its Application*, *6*(1), 149–172. https://doi.org/10.1146/annurev-statistics-030718-104951

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kreuter, F., Haas, G.-C., Keusch, F., Bähr, S., & Trappmann, M. (2020). Collecting survey and smartphone sensor data with an app: Opportunities and challenges around privacy and informed consent. *Social Science Computer Review*. https://doi.org/10.1177/0894439318816389

Kruschinski, S., & Haller, A. (2017). Restrictions on data-driven political micro-targeting in Germany. *Internet Policy Review*, *6*(4). Retrieved from https://policyreview.info/articles/analysis/restrictions-data-driven-political-micro-targeting-germany

Lathia, N., Sandstrom, G. M., Mascolo, C., & Rentfrow, P. J. (2017). Happier people live more active lives: Using smartphones to link happiness and physical activity. *PLoS One*, *12*(1), e0160589. https://doi.org/10.1371/journal.pone.0160589

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google flu: Traps in big data analysis. *Science*, *343*(6176), 1203–1205. https://doi.org/10.1126/science.1248506

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., . . . Van Alstyne, M. (2009). Computational social science. *Science*, *323*(5915), 721–723. JSTOR.

Lerner, A., Simpson, A. K., Kohno, T., & Roesner, F. (2016). *Internet Jones and the raiders of the lost trackers: An archaeological study of web tracking from 1996 to 2016*. Proceedings of the 25th USENIX Security Symposium, pp. 997–1013.

Link, M. W., Murphy, J., Schober, M. F., Buskirk, T. D., Hunter Childs, J., & Langer Tesfaye, C. (2014). Mobile technologies for conducting, augmenting and potentially replacing surveys executive summary of the AAPOR task force on emerging technologies in public opinion research. *Public Opinion Quarterly*, *78*(4), 779–787. https://doi.org/10.1093/poq/nfu054

Lohr, S. L. (2009). Sampling and survey design. In D. Pfeffermann & C. R. Rao (Eds.), *Sample surveys: Design, methods and applications* (pp. 3–8). Elsevier.

Lynch, J., Dumont, J., Greene, E., & Ehrlich, J. (2019). Use of a smartphone GPS application for recurrent travel behavior data collection: *Transportation Research Record*, *2673*(7), 89–98. https://doi.org/10.1177/0361198119848708

MacKerron, G., & Mourato, S. (2013). Happiness is greater in natural environments. *Global Environmental Change*, *23*(5), 992–1000. https://doi.org/10.1016/j.gloenvcha.2013.03.010

Matthews, T., Liao, K., Turner, A., Berkovich, M., Reeder, R., & Consolvo, S. (2016). *"She'll just grab any device that's closer": A study of everyday device & account sharing in households*. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 5921–5932. https://doi.org/10.1145/2858036.2858051

Möller, J., van de Velde, R. N., Merten, L., & Puschmann, C. (2019). Explaining online news engagement based on browsing behavior: Creatures of habit? *Social Science Computer Review*. https://doi.org/10.1177/0894439319828012

National Academies of Sciences, E. (2017). *Federal statistics, multiple data sources, and privacy protection: Next steps*. https://doi.org/10.17226/24893

Nickerson, D. W., & Rogers, T. (2014). Political campaigns and big data. *Journal of Economic Perspectives*, *28*(2), 51–74. https://doi.org/10.1257/jep.28.2.51

Nissenbaum, H. (2018). Respecting context to protect privacy: Why meaning matters. *Science and Engineering Ethics*, *24*(3), 831–852. https://doi.org/10.1007/s11948-015-9674-9

Obar, J. A., & Oeldorf-Hirsch, A. (2020). The biggest lie on the Internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, *23*(1), 128–147. https://doi.org/10.1080/1369118X.2018.1486870

Oberski, D. L., & Kreuter, F. (2020). Differential privacy and social science: An urgent puzzle. *Harvard Data Science Review*, *2*(1). https://doi.org/10.1162/99608f92.63a22079

Phillips, L. A., & Johnson, M. A. (2017). Interdependent effects of autonomous and controlled regulation on exercise behavior. *Personality and Social Psychology Bulletin*. https://doi.org/10.1177/0146167217733068

Philpot, R., Liebst, L. S., Levine, M., Bernasco, W., & Lindegaard, M. R. (2020). Would I be helped? Cross-national CCTV footage shows that intervention is the norm in public conflicts. *American Psychologist*, *75*(1), 66–75. https://doi.org/10.1037/amp0000469

Przepiorka, W., Norbutas, L., & Corten, R. (2017). Order without law: Reputation promotes cooperation in a cryptomarket for illegal drugs. *European Sociological Review*, *33*(6), 752–764. https://doi.org/10.1093/esr/jcx072

Raento, M., Oulasvirta, A., & Eagle, N. (2009). Smartphones: An emerging tool for social scientists. *Sociological Methods & Research*, *37*(3), 426–454. https://doi.org/10.1177/0049124108330005

Rodolfa, K. T., Saleiro, P., & Ghani, R. (2020). Bias and fairness. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science* (2nd ed.). Chapman and Hall and CRC Press. https://textbook.coleridgeinitiative.org/

Salganik, M. (2017). *Bit by bit: Social research in the digital age*. Princeton University Press.

Sapiezynski, P., Stopczynski, A., Lassen, D. D., & Lehmann, S. (2019). Interaction data from the Copenhagen networks study. *Scientific Data*, *6*(1), 315. https://doi.org/10.1038/s41597-019-0325-x

Scharkow, M., Mangold, F., Stier, S., & Breuer, J. (2020). How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences*, *117*(6), 2761–2763. https://doi.org/10.1073/pnas.1918279117

Scherpenzeel, A. (2017). Mixing online panel data collection with innovative methods. In S. Eifler & F. Faulbaum (Eds.), *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung* (pp. 27–49). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-15834-7_2

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmaß, Entwicklung und Ursachen*. Springer-Verlag.

Schnell, R. (2019). „Big Data"aus wissenschaftssoziologischer Sicht: Warum es kaum sozialwissenschaftliche Studien ohne Befragungen gibt. In D. Baron, O. Arránz Becker, & D. Lois (Eds.), *Erklärende Soziologie und soziale Praxis* (pp. 101–125). Springer Fachmedien. https://doi.org/10.1007/978-3-658-23759-2_6

Schnell, R., Hill, P. B., & Esser, E. (2018). Methoden der empirischen Sozialforschung. In *Methoden der empirischen Sozialforschung*. De Gruyter Oldenbourg.

Schober, M. F., Pasek, J., Guggenheim, L., Lampe, C., & Conrad, F. G. (2016). Social media analyses for social measurement. *Public Opinion Quarterly*, *80*(1), 180–211. https://doi.org/10.1093/poq/nfv048

Shelton, T., Poorthuis, A., Graham, M., & Zook, M. (2014). Mapping the data shadows of Hurricane Sandy: Uncovering the sociospatial dimensions of 'big data.' *Geoforum*, *52*, 167–179. https://doi.org/10.1016/j.geoforum.2014.01.006

Silver, L., Smith, A., Johnson, C., Jiang, J., Anderson, M., & Rainie, L. (2019). *Mobile connectivity in emerging economies*. PEW Research Center. Retrieved from www.pewresearch.org/internet/2019/03/07/mobile-connectivity-in-emerging-economies/

Squire, P. (1988). Why the 1936 literary digest poll failed. *Public Opinion Quarterly*, *52*(1), 125–133. https://doi.org/10.1086/269085

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., . . . Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1920484117

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*. https://doi.org/10.1177/0894439319843669

Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., & Lehmann, S. (2014). Measuring large-scale social networks with high resolution. *PLoS One*, *9*(4), e95978. https://doi.org/10.1371/journal.pone.0095978

Struminskaya, B., Lugtig, P., Keusch, F., & Höhne, J. K. (2020). Augmenting surveys with data from sensors and apps: Opportunities and challenges. *Social Science Computer Review*. https://doi.org/10.1177/0894439320979951

Stück, D., Hallgrímsson, H. T., Ver Steeg, G., Epasto, A., & Foschini, L. (2017). *The spread of physical activity through social networks*. Proceedings of the 26th International Conference on World Wide Web, pp. 519–528. https://doi.org/10.1145/3038912.3052688

Sugie, N. F. (2018). Utilizing smartphones to study disadvantaged and hard-to-reach groups. *Sociological Methods & Research*, *47*(3), 458–491. https://doi.org/10.1177/0049124115626176

Sugie, N. F., & Lens, M. C. (2017). Daytime locations in spatial mismatch: Job accessibility and employment at reentry from prison. *Demography*, *54*(2), 775–800. https://doi.org/10.1007/s13524-017-0549-3

Tene, O., & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. *Northwestern Journal of Technology and Intellectual Property*, *11*(5), xxvii–274.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.

Troiano, R. P., Berrigan, D., Dodd, K. W., Mâsse, L. C., Tilert, T., & Mcdowell, M. (2008). Physical activity in the United States measured by accelerometer. *Medicine & Science in Sports & Exercise*, *40*(1), 181–188. https://doi.org/10.1249/mss.0b013e31815a51b3

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the Facebook social graph. *ArXiv:1111.4503 [Physics]*. http://arxiv.org/abs/1111.4503

Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical tools for designing and weighting survey samples*. Springer International Publishing. https://doi.org/10.1007/978-3-319-93632-1

Vicéns-Feliberty, M. A., & Ricketts, C. F. (2016). An analysis of Puerto Rican interest to migrate to the United States using Google trends. *The Journal of Developing Areas*, *50*(2), 411–430. https://doi.org/10.1353/jda.2016.0090

Vosen, S., & Schmidt, T. (2011). Forecasting private consumption: Survey-based indicators vs. Google trends. *Journal of Forecasting*, *30*(6), 565–578. https://doi.org/10.1002/for.1213

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). *Student life: Assessing mental health, academic performance and behavioral trends of college students using smartphones*. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 3–14. https://doi.org/10.1145/2632048.2632054

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., . . . Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Willimack, D. K., Nichols, E., & Sudman, S. (2002). Understanding unit and item nonresponse in business surveys. In R. M. Groves, D. A. Dillman, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 213–242). John Wiley & Sons.

York Cornwell, E., & Cagney, K. A. (2017). Aging in activity space: Results from smartphone-based GPS-tracking of urban seniors. *The Journals of Gerontology: Series B*, *72*(5), 864–875. https://doi.org/10.1093/geronb/gbx063