

# Big data and Belmont: On the ethics and research implications of consumer-based datasets

Big Data & Society  
July–December: 1–12  
© The Author(s) 2021  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/20539517211048183  
journals.sagepub.com/home/bds  
 SAGE

Remy Stewart

## Abstract

Consumer-based datasets are the products of data brokerage firms that agglomerate millions of personal records on the adult US population. This big data commodity is purchased by both companies and individual clients for purposes such as marketing, risk prevention, and identity searches. The sheer magnitude and population coverage of available consumer-based datasets and the opacity of the business practices that create these datasets pose emergent ethical challenges within the computational social sciences that have begun to incorporate consumer-based datasets into empirical research. To directly engage with the core ethical debates around the use of consumer-based datasets within social science research, I first consider two case study applications of consumer-based dataset-based scholarship. I then focus on three primary ethical dilemmas within consumer-based datasets regarding human subject research, participant privacy, and informed consent in conversation with the principles of the seminal Belmont Report.

## Keywords

Data ethics, big data, consumer-based data, computational social science, data brokers, Belmont Report

## Introduction

The use of big data within social science research has experienced notable increases in adoption by individual academics and subsequent wider institutionalization across disciplines within the last decade. This is closely linked to the greater availability of digital data sources, growth in computational power to clean and process large, messy datasets, and the rising methodological interest in machine learning and computer science techniques adapted for the social sciences (Chen, 2018; Lazer and Radford, 2017). Big data has provided unparalleled access to records of individual behavior at a scale and granularity once unimaginable to engage with research topics, such as personal networks, political polarization, and social movements (Hofstra et al., 2017; Lotan et al., 2011; Wang et al., 2018). These advances have also brought about new controversies regarding the proper use of big data as related to research ethics. This is represented in debates around infamous projects, such as the Facebook emotional contagion study or the OkCupid data scrape (Hallinan et al., 2019; Zimmer, 2018). The political contentiousness of these examples highlights the ethical ambiguities social scientists contend with regarding current best practices for research using big data sources.

A sizable portion of social science research involving big data has used datasets obtained from publicly accessible platforms such as social media websites, including Twitter and Reddit (Golder and Macy, 2014; Proferes et al., 2021). Although there is a common practice toward anonymizing and deidentifying individual records when analyzing big data, major ethical issues remain unsettled for these sources on key topics regarding consent, privacy, and the definition of “human subjects”-based research. There are also additional concerns regarding how big data is used outside of the academy such as within commercial and governmental sectors. There is increasing evidence regarding how the unchecked use of big data promotes inequality and discrimination, such as in realms of policing, social services, and employment (Brayne, 2017; Eubanks, 2018; Valentine, 2019). By using big data in research, scholars must also reflect on

Department of Sociology, Cornell University, USA

### Corresponding author:

Remy Stewart, Department of Sociology, Cornell University, Ithaca, NY, USA.

Email: rps256@cornell.edu



how said adoption residually condones other questionable practices that draw from the same data sources.

Certain traits associated with big data have limited its greater use within the social sciences, such as the computational skills and processing power required to acquire, clean, and manage large datasets. However, a variety of private companies have streamlined this labor by synthesizing a massive amount of personal information into purchasable dataset products. This work will consider one of these sources of big data, known as consumer-based datasets (CBDs).

Also known as customer behavior data, customer analytics, consumer profiles, and various other equivalent industry terms, CBDs refer to the subset of big data that is collected and sold by businesses, commonly known as data brokers. These datasets contain a staggering sum of personal information on individuals and households. One broker, Acxiom, claims to possess data on 700 million individuals worldwide, including 95% of American households (US Security and Exchange Commission, 2004, 2017). The “consumer” in CBDs refers to the common use of these datasets for marketing campaigns to consumers, but also to one of the primary sources of collected data stemming from consumption-based mediums, such as credit card transactions and public purchasing records. Data brokers collect information for CBDs from a variety of primary sources, which when further developed by within-industry transactions leads to exceptionally comprehensive records on the personal information of a large portion of the US population. CBDs are products of the expanding data brokerage industry that capitalizes on the growing ubiquity of individual surveillance, record keeping, and digitally profiling consumer behavior.

The recent examples of big data used within social science research often feature records obtained from social media or search engines (Flores, 2017; Golder and Macy, 2011; Wang et al., 2018) or from specialized online spaces such as dating sites or virtual marketplaces (Diekmann et al., 2013; Lin and Lundquist, 2013). There is an additional line of big data research that draws from public or private administrative sources such as tax filings or medical records (De Vaan and Stuart, 2019; Young et al., 2016). CBDs stand at the intersection of these different types of big data. They are massive datasets with wide-reaching coverage of the US population, and they incorporate both digital traces from online activity and public administrative records.

CBDs feature multiple defining characteristics of big data as noted by Salganik (2018)—they have the *volume* of the sheer amount of collected material, the *variety* of available information, and the *velocity* by which new information is made available. They are additionally managed as a for-profit commodity that makes large sums of what is often previously private knowledge accessible to buyers. This adds a novel layer of complexity to the ethical standing

of CBDs compared to data that is either publicly available or provided only to select parties. The ethical challenges inherent within CBDs are further aggravated by the opacity and questionable business practices of data brokers themselves.

I examine CBDs and their future role within social science research by first providing an overview of what these datasets are composed of and how they are created. I investigate key issues surrounding these sources to highlight the multiple ethically questionable aspects of CBDs. I offer a case study of one CBD data broker and the recent use of their data in research published in high-impact social science journals to argue for the eminent significance of CBDs. I then engage with the core ethical dilemmas that such data sources raise, guided by the principles outlined in the seminal National Commission for the Protection of Human Subjects’ Belmont Report.

## Overview of CBDs and brokers

CBDs and their brokers offer a fundamentally different framework for data obtainment compared to current research practices that collect disarranged yet publicly accessible digital traces from online sources. Although brokers do not make CBDs for academics as their target customer base, the data is still purchasable by approved clients that have previously included scholars.

Collected information by data brokers includes full names, predicted household income, racial and ethnic identity, education level, family composition, lifestyle interests, consumption patterns, residential addresses, social security numbers, likely medical conditions, and beyond (Federal Trade Commission, 2014a). Individual consumer profiles are often further classified into subgroups representing commonly associated characteristics that can be used in targeted marketing. This categorization can be based on sensitive or potentially uncomplimentary associations, such as regarding chronic medical conditions or classifying individuals as likely low income and less educated based on their consumption habits.

The sources behind the collected data within CBDs are convoluted and often undisclosed. Information is known to be gathered from both available public records, such as state licensing and property transactions as well as from mediums with murkier privacy standards such as social media sites and credit card transactions. When solicited for greater transparency regarding the direct channels where personal information is acquired, firms have been repeatedly reluctant to disclose this information on the grounds of industry trade secrets. The data broker industry is also strikingly lucrative. The broker Acxiom alone made \$917 million in 2017 (US Security and Exchange Commission, 2017).

The Federal Trade Commission (FTC) released a seminal report in 2014 addressing CBDs and data brokers

that provided one of the first comprehensive overviews of the industry. The nine featured organizations within the investigation specialize in data agglomeration for product marketing, anti-fraud background checks, or the searchable registries of individuals and their personal information. The report emphasized the lack of transparency regarding firms' data collection practices and recommended enhancing consumer protection by adopting federal regulations toward CBD brokers.

The FTC's recommendations have been limited in implementation due to brokers' initiatives to preserve the ongoing concealment of their data collection protocols. Two recent state-based measures have offered limited gains toward enforcing procedural transparency within the CBD broker industry. The California Consumer Privacy Act (CCPA) of 2018 grants state residents the right to full disclosure of collected personal information when opting out of their data's inclusion within CBDs (California Civil Code, 2018). Vermont legislation passed in 2018 requires brokers to register under the Attorney General, effectively providing one of the first views regarding the relative size of the industry in the United States with 361 companies listed as of August 2021 (Vermont Secretary of State, 2021). However, these policies only apply to a select state. Brokers receive virtually no other federal or state oversight specific to this industry's practices and the distinct ethical concerns that arise from their commodification of personal consumer data.

CBD brokers highlighted within the FTC investigation have emphasized in their public relations messaging that individuals may "opt-out" of the collection of their personal information. This has served as a means to preserve the minimal regulation of the industry through firms claiming that opt-out policies facilitate consumer choice regarding privacy protections (Crain, 2018). Two primary weaknesses undermine the legitimacy of opt-out policies as a substitute for more stringent industry regulations. First, individuals must separately opt out with each brokerage firm through isolated requests to each company. Brokers often acknowledge within opt-out policies that the removal of information from their specific database does not include a complete retraction of their records for the broker's internal reference, or that revoking permission for further data collection does not delete previously sold information or change the behavior of any third party that may also own the consumer's personal records. Second, CBD brokers will often retain information on the consumer that may no longer be sold for business purposes but is kept for identifying the individual as a consumer who had "opted out." This practice is critiqued in how it transcends from the widespread concern toward individual privacy rights to the violated "right to be forgotten" (Newman, 2015). CBDs must retain individual records for their identification services to remain presumably valid, causing their business model to never truly be able to honor complete anonymity.

Once CBDs possess information on an individual, it cannot be retroactively rescinded.

This is one example of the variety of data broker practices that preserve the general lack of industry regulation at the expense of individual consumer autonomy. However, these questionable industry proceedings have not prevented the successful use of CBDs within academic research. There are unavoidable risks of harm raised toward study "participants"—referring to the individual consumers featured within CBDs—both directly within research that uses these datasets and indirectly by the actions of the data broker industry that academics purchase CBDs from. Previous studies that feature CBDs have often depended on personally identifiable information provided by data brokers such as home addresses and demographic characteristics for their substantive findings. This information is highly sensitive in nature, and even best practices around anonymization and data aggregation do not eliminate risks associated with data leakages or the reidentification of participants within these datasets. Researchers themselves may not be conducting ethical malpractice that would actively harm participants, but by patronizing data brokers they inadvertently condone the ethically questionable actions of the brokers themselves and the third parties they contract with. Past infractions that brokers have perpetuated include selling CBDs to predatory financial firms that target low-income households, promoting inequality in medical insurance access based on "risky" consumption patterns, and contracting with law enforcement agencies to support the surveillance of immigrants and communities of color (Allen, 2018; Federal Trade Commission, 2014b; Selbst, 2017). Furthermore, there are high-level ethical concerns toward participant harm via violating their ability to dictate who has access to their personal information. This touches on issues surrounding active consent regarding whether consumers would approve of having their personal information sold to academics and subsequently used within research. I further consider these dimensions of potential harm to subjects featured within CBDs through my following case study of two academic publications that use CBDs acquired from the same data broker.

### *Case study: Infutor and housing*

The data broker featured in the proceeding works is Infutor Data Solutions, a brokerage firm founded in 2003 and headquartered in the Chicago metropolitan area. The company specializes in consumer data analytics for marketing. Infutor advertises its data inventory as including over 260 million individual Americans with 97 million daily updates on consumer profiles that delineate over 155 types of personal characteristics (Infutor Data Solutions, 2021). Available information within Infutor's datasets includes full names, home addresses, email and phone

numbers, demographic traits, lifestyle interests, and purchasing behavior. Although it appears that Infutor's primary clients are private companies, the firm additionally contracts with academic researchers. Universities that Infutor has conducted business with include Stanford, Notre Dame, and Pennsylvania State.

The diversity of personal information within CBDs makes these sources well-suited for a variety of research interests across social science disciplines. One topic featured in recent scholarship using CBDs is housing dynamics. CBDs address previous limitations within housing research regarding longitudinal datasets and missing data. Due to their dual access to both property ownership records and purchases linked to residential addresses, CBDs offer comprehensive overviews of population-level housing trajectories. They are well-suited for research investigating individual household mobility patterns as well as aggregate housing trends within a given geographic region over time. Using CBDs as a primary data source is also relatively cost-effective for the magnitude and specificity of housing information they provide compared to alternative means of data acquisition.

These features of CBDs are all likely reasons behind why they have since emerged as a growingly popular data source for studying housing dynamics. I feature two recent empirical articles that both use Infutor data to conduct research on housing that was published in top-ranked journals within their respective disciplines. First is Diamond et al. with their 2019 article "The effects of rent control expansion on tenants, landlords, and inequality: Evidence from San Francisco" published in *American Economic Review*. The second is Phillips' 2020 piece "Measuring housing stability with consumer reference data" in *Demography*.

Diamond and colleagues investigate how a change in a preexisting rent control policy—which commonly implements limitations on rental price increases within a given region—impacts the housing tenure of renters and the leasing interests of landlords in San Francisco. The authors use difference-in-difference regression analyses with the natural experiment of a 1994 policy revision in San Francisco that implemented expanded rent control eligibility for smaller rental units. They find a housing retention effect for tenants occupying rent-controlled units but also a movement toward the non-rent-controlled condominium conversions of properties after the 1994 policy change. They argue that the total reduction in available units incentivized by the expansion of rent control eligibility has contributed to increased rental prices in San Francisco over time.

The findings of this study were widely publicized for their controversial conclusions on a politically contentious topic within housing policy. However, my particular interest in this work relates to the data behind how the authors obtained their results. Diamond et al. (2019: 3369) establish

that their sample sourced from Infutor consumer-based data "provides the entire address history of individuals who resided in San Francisco at some point between the years of 1980 and 2016." In practice, this provides the authors with a sample of 1.43 million tenant-to-building records. They effectively claim coverage of all tenants between 20 and 65 years of age living in the city starting from 1993, with their dataset additionally connecting personal identifiers to trace the housing mobility of individual families.

Diamond et al.'s article appears to have been a significant milestone for introducing CBDs within social science research. Unsurprisingly, the sheer prospects that such data offers for studying housing did not go without notice beyond the specific focus on rent control. Moving to my second empirical example, Phillips acknowledges the precedent set by Diamond et al. regarding the use of CBDs in his publication. He then expands on their work to consider further applications of Infutor data on a variety of topics related to housing mobility.

Phillips validates the accuracy of his particular Infutor dataset by comparing its coverage of address changes to a selection of relevant housing mobility events. These include moves out of New Orleans after Hurricane Katrina, the closure of the Robert Taylor Homes public housing complex in Chicago, and moves linked to a particular apartment complex in Washington, DC. The Infutor data robustly captures residential movement for individuals identified as having lived in the specific location of all three of these place-and-time-based examples at a cross-national level. The data additionally replicates well to reliability checks that refer to the Census' American Community Survey.

Phillips highlights many of the previously identified strengths of CBDs within his work, such as the data's size, general comprehensiveness, and their relatively low cost compared to other data collection methods. Phillips (2020: 1340) briefly touches on the ethics of the data by identifying how the selling of CBDs implies that these datasets can be considered publicly available which he claims "reduces [the data's] ethical sensitivity in institutional review and other ethical discussions." The author explains his process of deidentifying individuals in his featured housing investigations as well as highlighting how the Institutional Review Board (IRB) at his home institution approved his study. Phillips refers to various ethical "dilemmas" and "questions" throughout his article's discussion on the ethics of Infutor data without explicitly defining said questions or dilemmas. Issues such as informed consent, consumer privacy, and the opaque data management practices of brokers such as Infutor are not directly addressed by either Phillips or Diamond et al. Furthermore, Diamond et al. do not engage with any ethical concerns involving their Infutor data within their article and do not provide information regarding their practices for handling sensitive consumer records. Their article's accompanying disclosure

statement explains that they did not submit their project for IRB approval since the Infutor data was available for purchase and was not collected by the research team directly.

Both of these articles highlight the potentials of CBDs within academic research. They are methodologically rigorous studies tackling important topics through their respective datasets. They also provide applied examples of ethical uncertainties that surround CBDs. While I overall believe that both authors incorporated best research practices around ethical data use as experienced scholars, it does not stop the ongoing presence of potential harm toward their research subjects. The reidentification of participants is one of these risks—even if personally identifiable information was removed from either study—because geographic mobility patterns may be a uniquely identifiable behavior. Even when explicit steps were taken to disaggregate movement trends to minimize individual risks of reidentification, there is an ongoing concern toward whether the individuals featured in these projects would have actively consented to be included within these studies. Consumers may very well conceptualize having records of their residential moves both available for purchase and then subsequently used within academic research as itself a form of perpetuated harm against them. There is an additional broader level of potential harm toward the individuals featured within CBDs regarding what unknown ethically questionable behavior researchers are inadvertently condoning through patronizing a data broker such as Infutor due to the industry's widespread lack of business practice transparency.

## The ethics of CBDs

The undeniable promise that CBDs offer for social science research calls forward a need for further consideration of the ethical standing of CBDs before, rather than after, their greater adoption within academia. I ground my following analysis of the ethical dilemmas within these datasets from an original source regarding ethical standards via the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research's (1979) publication, commonly known as the Belmont Report. The Belmont Report brought the three key terms of respect for persons, beneficence, and justice into the research ethical canon. I consider each further as applied to CBDs and data brokers below.

Respect for persons emphasizes honoring research participants as autonomous individuals. This principle encompasses the additional key themes of informed consent and protecting vulnerable groups. Groups classically defined as "vulnerable" such as people who are pregnant or individuals who have cognitive disabilities are presumably included within CBDs. CBDs do not appear to have coverage of certain vulnerable groups such as minors, and the inclusion of other groups remains overall uncertain such

as individuals who are currently incarcerated. There is also a need to reexamine which personal characteristics should be classified as "vulnerable" beyond the traditional examples listed above. The Belmont Report suggests that "racial minorities" and "the economically disadvantaged" can also be considered as vulnerable groups, particularly within the context of research ethics (Stone, 2003).

There are previous incidents where CBD brokers have actively harmed individuals through their data transactions that disproportionately impact low-income households or people of color. For example, the broker LeapLab was charged by the FTC for malpractice for their targeted data aggregation of individuals who applied for payday loans. These rapidly approved but notoriously high-interest loans are primarily used by disadvantaged households experiencing financial hardship and are linked to predatory lending practices (Federal Trade Commission, 2014b; Mayer, 2003). LeapLab sold their loan-based CBD to outside firms including Ideal Financial Solutions, which then used the acquired sensitive information that included names and social security numbers to withdraw money directly from applicants' bank accounts. This is one instance where explicit harm was perpetuated toward vulnerable groups by CBD brokers, with other potential cases left unknown within the widespread secrecy that characterizes broker business proceedings.

The Belmont principle of beneficence underscores the need to maximize personal and wider social gain and minimize harms for both individual participants and the broader good of all parties potentially impacted by a given research project. Although CBD-powered research does undeniably contain the eminent scientific potential to support innovative and socially positive findings, it would do so with a source that has questionable ethical validity produced by firms operating under a high level of methodological secrecy. Furthermore, boyd and Crawford (2012) rightly caution against the notion that big data is inherently superior data for research despite widespread excitement around its adoption within academia. This is further applicable when the potential of CBDs for research is undermined by a lack of critical discourse to assess data validity and reliability compared to traditional collaborative and peer-reviewed scholarly practices. There have been multiple exposures regarding the inaccuracy of CBDs when representing individual consumer characteristics, which undermines the potential gains of CBD research if study findings are drawn from fundamentally incorrect information (Lucker et al., 2017).

The theme of justice within research ethics calls for "fair distribution of burden and benefits" between research participants, project facilitators, and in reference to the larger social world (Rawls, 1971). There are prominent concerns related to distributive justice and CBDs when considering the unequal burden of having data classified and then sold without direct compensation for the

“participants” within CBDs contrasted to brokerage firms that profit from said transactions. This is further magnified by consumer’s limited ability to opt-out of having their information collected and preexisting records permanently deleted. The financial benefits brokers obtain contrast various burdens individuals currently must bear regarding having their personal information commodified and the resulting risks of harm that stem from the use of CBDs.

Returning to the empirical examples featured in the case studies, I honor Phillip’s engagement with the ethical murkiness of his Infutor data in its own section of his publication. His argument in support of using this particular CBD despite concerns regarding consent is driven by two primary claims. The first refers to his own best research practices of minimizing the ability to reidentify participants in his analysis by aggregating individual locations by ZIP code and not including detailed map visuals. The second is that the IRB within his home institution approved his study where it is implied that classifying the data as “public information” dilutes the ethical dilemmas at hand.

IRBs are the administrative representatives that regulate research following the principles of the Belmont Report within the American academic context. They contribute to managing the ethical standards of human subject research conducted by scholars in their respective institutions. IRBs are not the focus of the ethical dilemmas inherent within research using CBDs, but they have and will continue to play a major role regarding how ethical precedents for CBD-based studies will be established. The fact that Phillips sought out the guidance of his home IRB and specified so in a section of his study addressing ethical considerations exemplifies the author’s diligence to the issues at hand. However, as research and publicity on both IRBs and the critiqued studies they have overseen (Flick, 2016), demonstrates *one IRB approval does not a clear ethic*.

The works featured in the Infutor case studies emphasize the need to engage with the ethics of CBDs before these data sources become more widely adopted within social science research. Using the guiding frameworks of the Belmont Report, I detail three key ethical tensions within CBDs. The first is whether CBDs can be classified as human subject-based research. The second point considers issues around confidentiality and privacy in CBDs. The third is the obligation toward informed consent in research and how both CBDs themselves and researchers who use the data sources may mishandle this core ethical practice.

### *CBDs and human subject research*

Is big data scholarship that uses CBDs as its primary data source featuring human subjects as its study participants? Referring to the standard definitions of human subject-based research would likely lead to the conclusion that

CBD-powered research does qualify, but that the data is exempted from IRB review. According to the Code of Federal Regulations for Protection of Human Subjects (i.e. “Common Rule”) as applied to social science research, human subjects are living individuals with whom researchers either directly interact with or collect their identifiable private information (Federal Policy for the Protection of Human Subjects, 1991). CBDs are primarily composed of individuals’ records, and the data contains extensive amounts of identifiable private information. However, research can be understood as including human subjects but still be exempted from undergoing a full IRB review. The secondary nature of this data source for a principal investigator—drawing from the fact that they are not the primary collector of the data but rather obtain it from commercial brokers—confounds the resulting review standards that are applied to CBDs. A research type that proponents have effectively argued CBDs as qualifying under is “secondary research using identifiable information... if publicly available, or recorded such that subjects cannot be re-identified,” which falls underneath Category 4 of exempted research types in the Common Rule (Federal Policy for the Protection of Human Subjects, 1991). Brokers can provide selections of datasets that avoid personally identifiable information entirely. CBDs can be further argued to qualify as public data since they are commercially available for purchase by any party a given broker decides to conduct business with.

IRBs are heavily involved in regulating studies that often include active personal risk to participants—such as biomedical clinical trials or psychological experiments—that are quite different from research conducted by social scientists using big data. A history of serious ethical infractions within research led to the Belmont Report and the resulting Common Rule to emphasize the need to protect participants (Adashi et al., 2018). Once the participant has become a publicly accessible record of information within a dataset rather than physically partaking in a study, concerns regarding harm overall diminish. These secondary data sources are often considered to be far enough removed from the direct research participant to not pose any imminent risk to their wellbeing should their information be used in research.

However, the unique features of CBDs call forward a need to reconsider how to conceptualize harm toward human subjects. The dividing line for IRBs between expedited reviews versus full reviews of human subject research is often a matter of the immediate level of risk a study poses toward participants. In social and behavioral research, *informational risk* is often a more prominent concern beyond direct physical or psychological harm. Informational risk refers to “the potential harm [that can occur] from disclosure of information about an identified individual” (National Research Council, 2014: 112). Accounting for informational risk requires identifying

what types of disclosure risks are liable to occur from a research project and how specific individuals may be more likely to be identified and therefore experience a greater risk of harm. Using publicly available data, not including sensitive personal information, and not featuring vulnerable groups are all viable practices to mitigate informational risk. CBDs can incorporate all these risk reduction strategies with due diligence between both brokers and researchers.

However, even these best practices do not absolve all concerns regarding potential harm toward participants within CBDs. Metcalf and Crawford (2016: 7) critique the idea that “the risk to research subjects depends on *what kind* of data is obtained and *how* it is obtained, not *what is done* with the data after it is obtained.” There are ethically controversial aspects to all three of these factors as applied to CBDs, but for understanding its place within human subjects research the foremost issue is regarding use after CBDs are obtained by a given party. CBDs may be generated from a combination of public and semi-public sources, but that does not prevent either the collected information itself or the deductions that stem from the gathered information from being private and sensitive in nature for a given participant.

The underlying assumption to be questioned when considering whether or not standards of human subject research apply to CBDs is the idea that there is no potential risk of harm for participants due to the data being classified as public and secondarily sourced. Brokers that can provide information as vulnerable as social security numbers and residential addresses to a variety of third parties with minimal regulation directly imply a significant level of informational risk. Individuals are classified within these datasets on measures that identify vulnerable group membership such as belonging to a socially marginalized identity or having a medical condition. Even massive sample sizes using aggregate records do not completely diminish fears of data breaches or individual reidentification. It is likely that the majority of social scientists would anonymize their CBD samples quite effectively, particularly within the large *N* studies that these sources are well suited for. However, it does not excuse researchers from having to address the hazards associated with handling what is often deeply private information that was originally collected on questionable ethical grounds.

Another key consideration is the risk scholars may be condoning outside of academia when they support brokerage firms through purchasing CBDs and promoting their use through their research. There are a variety of dangers related to external third parties that researchers cannot actively control but also cannot be passively tolerant of while following the Belmont Report’s ethical principles. Data breaches from brokered CBDs have occurred multiple times, such as in 2017 with 147 million Americans having records released from the credit broker Equifax or in 2018

with 340 million individuals’ information exposed from leaked Exactis data (Greenberg, 2018; Puig, 2019). Brokers that specialize in the data behind people search services profit from the unscrupulous actions of third parties that search for individuals for reasons linked to domestic violence, identity fraud, or other maleficent purposes. Data mismanagement can lead to the false representations of credit histories or criminal records that may lead to severe consequences for individuals such as not being hired for a potential job or approved for housing when background checks draw from CBDs. Brokers have been previously charged with representing said information inaccurately, while there is concurrently limited recourse to ensure future data validity due to the lack of influence individual consumers can leverage toward the brokerage industry (Hurley and Adebayo, 2016; Weiss, 2012). The sheer amount of personal information that is collected for CBDs is, therefore, vulnerable to multiple types of misuse. This is likely to disproportionately harm the already marginalized and socially disempowered as well. A final underlying assumption to question is that the risk of harm—physical, emotional, or otherwise—is the most valid metric to center the ethical framing of human subject-based research on. Participant agency and a more equal footing between subjects and researchers are also desirable standards that support regulating the use of CBDs despite the secondary and “public” nature of the data itself.

### *CBDs and privacy*

CBDs regularly include information that due to its personal sensitivity for participants IRBs mandate active privacy protocols for handling within research studies. However, brokers operate outside of a regulatory framework that enforces protective measures such as those facilitated by IRBs. Concerns regarding breaches of sensitive data lead researchers and IRBs to practice due diligence toward keeping information protected, preserved, and anonymized. In comparison, what has often been closely regulated research material regarding sensitive data is now a publicly purchasable commodity through CBDs. Concerns toward the exposure of personal information that has historically driven privacy standards must instead reckon with the widespread availability of the same information as now publicly for sale via data brokers.

Understanding the relationship between privacy and research ethics requires engaging with two key concepts that assist in classifying what is and is not private information. The first is that private information is what individuals can reasonably expect to not be monitored or collected by third parties, while the second connects to reasonable expectations toward information that is not usually publicly available (Zimmer, 2018: 3). The first principle addresses privacy concerns toward information that could otherwise be considered public knowledge. Activity on social media

sites and public administrative records are information types within CBDs that fall into this category. Many consumers are aware that their behavior becomes public information for the organizations that facilitate these services, but they are likely not as knowledgeable about how the same information is then made available to third parties. Privacy ethics are immediately questionable based on this principle regarding what individuals do or do not expect to happen with their information when applied to CBDs. An individual can post about themselves on a public digital platform and not expect it to be included in a CBD and subsequently used within research. All because it is possible to extract publicly posted information does not mean that third parties are now condoned to do so by the individuals whose said records they represent or that these practices are morally or ethically sound. As Boyd and Crawford (2012: 672) succinctly argue, “just because content is publicly accessible does not mean that it was meant to be consumed by just anyone.”

The second concept refers to private information that individuals know is recorded but likely do not expect to be made public to third parties whatsoever. Medical histories are a definitive example within brokered data. Although CBDs do not receive direct medical documentation from providers due to HIPAA protections, they are still able to identify highly sensitive health conditions through recording the secondary consumption patterns of health-related products. Another context where this principle applies is when consumers “consent” to the collection of their private information as required for obtaining certain goods and services. An individual can consent to a degree of personal disclosure within a specific transaction without actively agreeing to have the provided information distributed to outside parties. This theoretical understanding of privacy is often incongruous with the contemporary applications of privacy policies. Individuals must often agree to both the collection and distribution of their private information without differentiating one from the other, with the only viable alternative being the consumer avoiding using services that require these permissions altogether.

An implied assumption through defining private information by these two principles is that there is a common standard of what is “reasonable” for delineating what information should or should not be protected. Data within CBDs that align more with the second principle is likely easier to defend as “reasonably” private than the first. Private information that is not expected to become public is often more normatively seen as protected information than what is public but not assumed to be collected by actors such as data brokers. The act of classifying public versus private information cannot be separated from the larger social contexts in which information occurs or from the different actors that either provide or receive the information.

Nissenbaum’s (2004) theory of contextual integrity offers a robust framework for engaging with the dilemmas around classifying private information. She explains how contextual integrity calls for understanding privacy through the twin influences of *norms of appropriateness* and *norms of distribution*. Appropriateness addresses how information can or cannot be reasonably expected to be revealed. Distribution concentrates on standards of how information flows or does not flow to other parties. For Nissenbaum, infractions toward either or both norms constitute a contextually informed violation of privacy. CBDs are formed with what appears to be transgressions toward both standards of contextual integrity. They are composed of personal information that many consumers likely do not believe is normatively appropriate to collect, nor would they find it normatively defensible to have said information be distributed in the private market to third parties such as researchers.

Privacy concerns regarding CBDs also extend beyond what personal information individuals can reasonably expect to be observed by strangers. It continues into expectations around classification such as when CBDs place individuals into demographic and lifestyle clusters that could be perceived as defamatory, as well as the commodification of information that is then sold to outside parties. This moves beyond individual stakeholder interests to instead the influence of defining identity groups with potential social and political consequences, as well as the taking of private knowledge that is then developed into a market commodity. These uses of personal data create more complex circumstances for what individuals would or would not be comfortable with regarding their private information. These factors also heighten the risk of privacy breaches due to the wider range of involved actors beyond the “providers” of individual consumers and the “receivers” of brokers that develop CBDs.

A final perspective to aid with understanding the ethical murkiness of CBDs and consumer privacy is the *privacy asymmetry* between consumers and data brokers. As specified by Crain (2018: 91), privacy asymmetry recognizes how “people are opened up to increasingly extensive forms of monitoring, while the institutions doing the monitoring and the information they collect remain hidden from view.” Privacy concerns regarding sensitive information become even more pertinent when additionally considering the lack of transparency behind the creation and use of CBDs. Consumers have limited leverage to protect their private information, but they also have little ability to request less privacy from the firms that create the CBD marketplace. Following Belmont guidelines, this unequal division between the personal exposures of private information with concurrent secrecy from brokerage firms implies a prominent ethical injustice regarding respect for participants or equivalent benefits gained between two parties.



### *CBDs and informed consent*

Obtaining informed consent from participants is a key ethical practice within human subject research. Even within projects that raise ethical concerns due to involving participants from vulnerable groups or researching sensitive topics, having participants provide actively informed consent to be studied permits otherwise ethically controversial projects to move forward for greater social beneficence. However, the standards of informed consent are problematic throughout CBDs and data broker practices.

A telltale signal to justify questioning the ethics of consent within CBDs is the documented discomfort consumers have toward their data being collected and marketed. A 2019 Pew Research survey representative of the American population investigated participants' experiences with data privacy, online surveillance, and the practices of companies that include CBDs brokers (Auxier et al., 2019). The results emphasize how consumers feel disempowered of personal agency toward protecting their own information. 81% of the survey participants report believing that the potential risks outweigh any benefits toward having their personal information be collected by data brokers. Additionally, 64% of participants would "not feel comfortable with companies sharing their personal data with outside groups doing research that might help them improve society." It is difficult to persuasively claim that adequate consent is occurring for research that features CBDs when potentially almost two-thirds of the general population may not have provided such consent if they had more leverage to actively decide whether they would like their data to be included within a given study.

Claiming adequate standards of consent through consumers "agreeing" with often lengthy and jargon-filled privacy policies is morally dubious. IRBs often require informed consent requests to be presented in accessible language with clear definitions of participant rights in studies that fall under their jurisdiction. The ethical unsoundness of privacy policies is further heightened by the sheer magnitude of policies individuals encounter including from a variety of sources that brokers collect records from. McDonald and Cranor (2008) estimate that it would take individuals around 201 hours to read through all of the privacy policies that they encounter in a given year. Individuals often cannot participate in these services unless they agree to a privacy policy, and they often have little leverage to debate components of these policies with the companies that mandate them.

Another common ethical practice that CBD-based research does not incorporate is the ability for participants to rescind consent at any time within a study. This is a customary option within research that actively collects data such as interviews and experiments as a core procedure for retaining participant agency and promoting respect for persons. The researcher-researched dichotomy has been

critiqued for its inherent power imbalance (Plesner, 2011), and participants having the option to rescind consent mitigates harm that could result from this inherently hierarchical relationship. The ability to rescind consent is commonly applied to projects that actively collect data on participants over time. CBDs are created and maintained through comparable longitudinal data collection practices but under significantly murkier consent standards. If a participant in a CBD dataset rescinds the consent they "provided" through opting out of having their data collected, not only do their records remain in the original brokered data universe, but parties who already possess the dataset are neither obligated to delete said records nor are presumably notified of any opt-out requests. Once the data has been dispersed on already weak consent standards it is even more difficult for participants to rescind consent regarding their information to not be used by outside agents. An individual may avoid further commodification of their information within CBDs through opt-out policies, but they cannot truly obtain full anonymity with data brokers or outside parties who already own their records.

A counterargument to the idea that only human subject research grounded in active consent is ethically permissible can be found within the American Sociological Association Code of Ethics. According to ASA standards, consent can be waived in research when there is "no more than minimal risk for research participants" and "the research could not practically be carried out if informed consent were required" (American Sociological Association, 2018: 13). Comparing CBDs to standards within conventional human subject research already calls forward a need to question the traditional understandings of participant risk, but this guideline introduces the additional factor of practicality in the obligation toward obtaining informed consent. It would certainly be unreasonable to contact the sheer number of individuals included within CBDs to ask if they consent to be participants in a study, and even attempting to connect with subpopulations of these datasets would still likely place an unreasonable burden on researchers. It is also ethically questionable to contact potential participants through personal information provided through CBDs to begin with.

At a more fundamental level, researchers should not be the ones obligated to pursue obtaining informed consent from the "participants" of CBDs as they are the customers of brokers and are secondary users of the data. The need for informed consent begins at the primary data collection level. Actively provided consent from users is not prioritized by brokers likely due to potentially reducing a primary selling point of CBDs regarding wide-range population coverage if consumers are provided more leverage to dictate the use of their data. This disregard toward active consent is an inherently disempowering practice and one that does not align with Belmont principles similar to current privacy standards within CBDs.

## Discussion

In the age of big data and its incorporation within academic research, CBDs stand as a topic of rising significance and scholarly potential. CBDs consolidate a variety of big data sources into a market commodity form where private data brokers facilitate the transaction of large sums of information regarding a broad range of the US population. The design of CBDs that feature a wide range of personal records connected to consumer profiles tracked over time can facilitate a variety of academic studies using these sources, such as those highlighted in the case study regarding housing dynamics. However, CBDs also have distinct issues regarding privacy, consent, risk of harm, transparency, lack of regulation, and the potential to promote social inequities. They stand at the crossroads of strong empirical opportunity contrasted with deep-set ethical challenges. Such data sources are a commercialized manifestation of what Salganik (2018: 302) notes as “one of the most fundamental tensions in research ethics: using potentially unethical means to achieve ethical ends.”

The featured case study of two articles that use Infutor CBDs is well-suited for conceptualizing this tension within recent empirical scholarship. These works are both theoretically and methodologically rigorous pieces with important findings that depend on the information available through CBDs. Both articles also demonstrate the ethical murkiness involved with using CBDs within social science research. Diamond et al. spend minimal time engaging with ethics after claiming their data qualifies as public information. Phillips follows common practices around deidentification and record aggregation aligned with an ethical harm-reduction strategy to minimize the risks inherent in CBDs. However, I argue that the risks of harm to participants within CBDs cannot be ameliorated through traditional risk mitigation strategies alone. By following common standards for managing private and sensitive data as featured by these two studies, the authors can only address a select portion of the ethical concerns found within CBDs. Information risk linked to data breaches and reidentification, the lack of active participant consent, and the questionable business practices of both data brokers and the third parties they contract with remain as deeper ethical issues underlying both studies.

CBDs can be critiqued in their ethical shortfalls through the Belmont Report principles of respect for persons, beneficence, and justice. The largest discrepancies between CBDs and ethical conventions are found within the core topics of what is classified as human subject research, dilemmas around privacy, and issues regarding informed consent. These quandaries do not have simple answers for best practices and often require considering the impacts of parties outside of academia. With proper discussion and institutionalization within preexisting ethical regulation

systems, social scientists can likely reach a larger consensus on best practices regarding CBDs.

A major limitation of this work is that its arguments are entirely contingent on the US context regarding big data. Although CBDs are certainly present and applicable within other countries, the ethical and legal arenas both firms and researchers operate under vary considerably from the US. Other nations and regions are setting precedents for data standards and privacy that offer indispensable guidance for American policy initiatives, even if the US does operate within a classically liberal political framework regarding industry deregulation and limited consumer privacy compared to other countries. Four of the seven principles of the EU’s General Data Protection Regulation (GDPR) Law enacted in 2018—minimizing the collection of data to what is necessary, protecting special categorical information, following a specific purpose for collection, and ensuring data accuracy (Andrew and Baker, 2021)—are all guidelines that scholars can advocate to be implemented toward CBDs within the US.

There are a variety of potential policy initiatives that scholars who conduct CBD-based research can support to promote greater transparency and ethical standards among data brokers. One policy includes federal-level broker registries where firms are required to provide public information regarding business proceedings such as how they obtain personal data, parties they contract with to sell said data, their privacy protection plans, their procedures for contacting affected individuals after data breaches, and guidelines for consumers to view, edit, and remove their collected data within CBDs. Another key initiative is the creation of non-partisan regulatory agencies tasked with ensuring policy compliance from brokers and addressing violations through fines similar to managing GDPR infractions within the EU (Runte and Kamps, 2021). Researchers can also proactively choose to patronize brokerage firms that practice greater consideration toward ethics such as by making their data collection procedures public information, demonstrating their prioritization of consumer rights, or by pursuing proactive “opt-in” policies regarding consumer consent against current “opt-out” industry standards. Finally, supporting the further training of IRBs regarding the unique ethical concerns of both CBDs and contemporary big data sources broadly will reinforce the institutionalization of ethical standards regarding CBDs within the academy. Subjects such as eminent informational risks within public online data, the ethically questionable actions of third parties that contract with brokers, and the changing landscape regarding threats of data breaches and participant reidentification are three distinct topics researchers can support the widespread consideration toward within IRB reviews.

Although CBDs may not be currently mainstream knowledge among social scientists as a source of quantitative data, their ongoing integration into disciplines such as

economics and demography as featured in the Infutor case study suggests that it is likely only a matter of time before more research features these datasets. It is essential that social scientists continue to grapple with the ethical issues within CBDs if and when their wider adoption becomes a reality within the academy. These matters are not ones that can be properly addressed without collective consideration of its multiple complexities by scholars from a diversity of interests, backgrounds, and opinions. This piece is, therefore, intended to serve as a subject introduction and a dialogue initiator regarding ethics in CBD-based scholarship and big data-powered research agendas in general.

### Acknowledgments

I am grateful to Cristobal Young, Karen Levy, and editors and reviewers for their comments and guidance on this manuscript. Additionally, I would like to acknowledge the helpful feedback provided during a presentation of this work at the American Sociological Association 2021 Annual Meeting.


### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Science Foundation Graduate Research Fellowship Program (Grant No. 1650441) and the Cornell Office of Diversity Dean's Excellence Fellowship.

### ORCID iD

Remy Stewart  <https://orcid.org/0000-0002-3193-8486>

### References

- Adashi E, Walters L and Menikoff J (2018) The Belmont report at 40: Reckoning with time. *American Journal of Public Health* 108(10): 1345–1348.
- Allen M (2018) Health insurers are vacuuming up details about you - and it could raise your rates. Available at: [www.propublica.org](http://www.propublica.org) (accessed 5 June 2021).
- American Sociological Association (2018) Code of ethics. Available at: [www.asanet.org/sites/default/files/asa\\_code\\_of\\_ethics-june2018a.pdf](http://www.asanet.org/sites/default/files/asa_code_of_ethics-june2018a.pdf) (accessed 5 June 2021).
- Andrew J and Baker M (2021) The general data protection regulation in the age of surveillance capitalism. *Journal of Business Ethics* 168: 1–14.
- Auxier B, Rainie L, Anderson M, et al. (2019) *Americans and privacy: Concerned, confused and feeling lack of control over their personal information*. Available at: [www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/](http://www.pewresearch.org/internet/2019/11/15/americans-and-privacy-concerned-confused-and-feeling-lack-of-control-over-their-personal-information/) (accessed 5 June 2021).
- Boyd D and Crawford K (2012) Critical questions for Big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication, & Society* 15(5): 662–679.
- Brayne S (2017) Big data surveillance: The case of policing. *American Sociological Review* 82(5): 977–1008.
- California Civil Code (2018) California Consumer Privacy Act of 2018. Available at: [www.leginfo.ca.gov](http://www.leginfo.ca.gov) (accessed 5 June 2021).
- Chen SH (2018) *Big Data in Computational Social Science and Humanities*. Berlin, Germany: Springer.
- Crain M (2018) The limits of transparency: Data brokers and commodification. *New Media & Society* 20(1): 88–104.
- De Vaan M and Stuart T (2019) Does intra-household contagion cause an increase in prescription opioid use? *American Sociological Review* 84(4): 577–608.
- Diamond R, McQuade T and Qian F (2019) The effects of rent control expansion on tenants, landlords, and inequality: Evidence from San Francisco. *American Economic Review* 109(9): 3365–3394.
- Diekmann A, Jann B, Przepiorka W, et al. (2013) Reputation formation and the evolution of cooperation in anonymous online markets. *American Sociological Review* 79(1): 65–85.
- Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. London: St. Martin's Press.
- Federal Policy for the Protection of Human Subjects (1991) Common rule. §46 C.F.R. Available at: [www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html](http://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html) (accessed 5 June 2021).
- Federal Trade Commission (2014a) Data brokers: A call for transparency and accountability. Available at: [www.ftc.gov](http://www.ftc.gov) (accessed 5 June 2021).
- Federal Trade Commission (2014b) FTC charges data broker with facilitating the theft of millions of dollars from consumers' accounts. Available at: [www.ftc.gov](http://www.ftc.gov) (accessed 5 June 2021).
- Flick C (2016) Informed consent and the Facebook emotional manipulation study. *Research Ethics* 12(1): 14–28.
- Flores RD (2017) Do anti-immigrant laws shape public sentiment? A study of Arizona's SB 1070 using twitter data. *American Journal of Sociology* 123(2): 333–384.
- Golder SA and Macy MW (2011) Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333(6051): 1878–1881.
- Golder SA and Macy MW (2014) Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology* 40: 129–152.
- Greenberg A (2018) Marketing firm exactis leaked a personal info database with 340 million records. Available at: [www.wired.com](http://www.wired.com) (accessed 5 June 2021).
- Hallinan B, Brubaker J and Fiesler C (2019) Unexpected expectations: Public reaction to the Facebook emotional contagion study. *New Media & Society* 22(6): 1076–1094.
- Hofstra B, Corten R, Tubergen FV, et al. (2017) Sources of segregation in social networks: A novel approach using facebook. *American Sociological Review* 82(3): 625–656.
- Hurley M and Adebayo J (2016) Credit scoring in the era of Big data. *Yale Journal of Law and Technology* 18(1): 148–216.

- Infutor Data Solutions (2021) Our data is unmatched. Available at: [infutor.com](http://infutor.com) (accessed 5 June 2021).
- Lazer D and Radford J (2017) Data ex machina: Introduction to big data. *Annual Review of Sociology* 43: 19–39.
- Lin KH and Lundquist J (2013) Mate selection in cyberspace: The intersection of race, gender, and education. *American Journal of Sociology* 119(1): 183–215.
- Lotan G, Graeff E, Annany M, et al. (2011) The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions. *International Journal of Communication* 5: 1375–1405.
- Lucker J, Hogan SK and Bischoff T (2017) Predictably inaccurate: The prevalence and perils of bad big data. *Deloitte Review*: 9–25.
- McDonald AM and Cranor LF (2008) The cost of reading privacy policies. *I/S: A Journal of Law and Policy for the Information Society* 4(3): 543–568.
- Mayer R (2003) Payday loans and exploitation. *Public Affairs Quarterly* 17(3): 197–217.
- Metcalfe J and Crawford K (2016) Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society*: 1–14.
- National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1979) *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. Washington, DC: Department of Health, Education, and Welfare.
- National Research Council (2014) *Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences*. Washington, DC: National Academies Press.
- Newman AL (2015) What the ‘right to be forgotten’ means for privacy in a digital age. *Science* 347(6221): 507–508.
- Nissenbaum H (2004) Privacy as contextual integrity. *Washington Law Review* 79(1): 101–139.
- Phillips DC (2020) Measuring housing stability with consumer reference data. *Demography* 57(4): 1323–1344.
- Plesner U (2011) Studying sideways: Displacing the problem of power in research interviews with sociologists and journalists. *Qualitative Inquiry* 17(6): 471–482.
- Proferes N, Jones N, Gilbert S, et al. (2021) Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*: 1–14.
- Puig A (2019) *Equifax data breach settlement: What you should know*. Available at: [consumer.ftc.gov](http://consumer.ftc.gov) (accessed 5 June 2021).
- Rawls J (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Runte C and Kamps M (2021) GDPR enforcement tracker report - 2nd edition 2021. CMS legal. Available at: [www.cms.law](http://www.cms.law) (accessed 5 June 2021).
- Salganik MJ (2018) *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.
- Selbst A (2017) Disparate impact in big data policing. *Georgia Law Review* 52(1): 109–195.
- Stone HT (2003) The invisible vulnerable: The economically and educationally disadvantaged subjects of clinical research. *The Journal of Law, Medicine, and Ethics* 31(1): 149–153.
- US Security and Exchange Commission (2004) Form 10-K: Acxiom Corporation. Available at: [sec.gov](http://sec.gov) (accessed 5 June 2021).
- US Security and Exchange Commission (2017) Form 10-K: Acxiom Corporation. Available at: [sec.gov](http://sec.gov) (accessed 5 June 2021).
- Valentine S (2019) Impoverished algorithms: Misguided governments, flawed technologies, and social control. *Fordham Urban Law Journal* 46(2): 364–427.
- Vermont Secretary of State (2021) Data broker search. Available at: [bizfilings.vermont.gov/online/DatabrokerInquire/DataBrokerSearch](http://bizfilings.vermont.gov/online/DatabrokerInquire/DataBrokerSearch) (accessed 5 June 2021).
- Wang Q, Phillips NE, Small ML, et al. (2018) Urban mobility and neighborhood isolation in America’s 50 largest cities. *Proceedings of the National Academy of Sciences* 115(30): 7735–7740.
- Weiss N (2012) Combatting inaccuracies in criminal background checks by giving meaning to the fair credit reporting act. *Brooklyn Law Review* 78(1): 271–304.
- Young C, Varner C, Lurie IZ, et al. (2016) Millionaire migration and taxation of the elite: Evidence from administrative data. *American Sociological Review* 81(3): 421–446.
- Zimmer M (2018) Addressing conceptual gaps in big data research ethics: An application of contextual integrity. *Social Media + Society*: 1–11.