Journal Title
XX(X):1–11
©The Author(s) 2021
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

More than Weight, Less than Driver's License Number: Comparative Sensitivity of Social Media Data and Their Acceptable Use in Research

Libby Hemphill[1,2], Angela Schöpke-Gonzalez[2], and Anmol Panda[2]

Abstract

NOTE: Submitted for peer review June 2022

Social media data offer a rich resource for researchers interested in public health, labor economics, politics, social behaviors, and other topics. However, scale and anonymity mean that researchers often cannot directly get permission from users to collect and analyze their social media data. This article applies the basic ethical principle of respect for persons to consider individuals' perceptions of acceptable uses of data. We draw from scholarship about individuals' perceptions of acceptable uses of other types of sensitive data, which we compare with individuals' perceptions of acceptable uses of social media data. Our survey of 1018 people shows that individuals think of their social media data as moderately sensitive and agree that it should be protected. Respondents are generally okay with researchers using their data in social research but prefer that researchers clearly articulate benefits and ask for explicit consent before conducting research. Our findings suggest that, much like other types of sensitive data, social media data researchers should carefully balance risks for individuals with benefits to society and avoid revealing personal details when possible.

Keywords

social media, information sensitivity, privacy, research data

Introduction

In early 2022, Twitter released its no-code API, making it possible for researchers to access Twitter data without needing to have programming expertise. Twitter data is now more accessible for a wider range of researchers to study topics like social capital (Steinfield et al. 2008), political agendas (Hemphill et al. 2020), labor economics (Antenucci et al. 2014), and public health (Ordun et al. 2013). This increased data availability makes ethical questions about social media data use for research, all the more pressing. For example, *should* researchers be collecting social media data? If so, when and how ought researchers to use it? Legally, social media users grant broad permissions for their data to be used when they agree to platforms' terms of service (TOS). However, users often do not actually read or understand TOS (Obar and Oeldorf-Hirsch 2020), may not think of their data as public (Fiesler and Proferes 2018), and may not realize that researchers are among that public (Bernstein et al. 2013; Marwick and Boyd 2010). Prior work finds that users actually prefer to grant explicit consent to use their data in research despite agreeing to TOS (Fiesler and Proferes 2018), and their attitudes toward acceptable data uses depend heavily on the research context and goals (Gilbert et al. 2021). How can social media researchers reconcile legal ability to use social media data with individuals' preferences about their social media data being used for research?

A basic principle in research called *respect for persons* can bring clarity to how researchers can think about ethical social media data use practices. Respect for persons (Office for Human Research Protections (OHRP)) requires that researchers orient their practices around individuals' perceptions of acceptable uses of data. Respect for persons is often addressed through informed consent processes that obtain explicit permission from individuals to use their data. Explicit consent serves to inform individuals of the opportunity to have data about them collected for research, and to express their preferences about their data being collected by accepting or declining participation.

However, explicit consent processes with social media users are often infeasible. The scale and anonymity of social media mean that researchers often cannot directly elicit individuals' perceptions about the acceptable uses of data they generate on social media. One exception is the Documenting the Now Project, which created "Social Humans" labels to attach explicit use permissions to content and analyses from social media (Documenting the Now). Social Humans labels aim to bridge the gap between the legal permissions users grant when agreeing to platforms'

[1] ICPSR, University of Michigan, USA
[2] School of Information, University of Michigan, USA

CRediT author statement: **Libby Hemphill**: Conceptualization, Methodology, Validation, Formal Analysis, Resources, Data Curation, Writing - Original Draft, Writing - Review and Editing Supervision, Project administration, Funding acquisition; **Angela Schöpke Gonzalez**: Validation, Writing - Original Draft, Writing - Review and Editing; **Anmol Panda**: Conceptualization, Methodology, Writing - Review and Editing

Corresponding author:
Libby Hemphill, University of Michigan, Ann Arbor, MI 48104, USA
Email: libbyh@umich.edu

TOS and content creators' wishes. However, this method has yet to be widely adopted, meaning even explicit use permissions like Social Humans labels cannot effectively guide researchers about how to ensure that they are respecting individuals' preferences about use of their data.

To understand how we can realize respect for persons in the particular context of social media data research, we look to scholarship on users' perceptions about acceptable uses of other types of sensitive yet widely available data. Prior work finds that social media data users think of their social media data similarly to how they think about widely-recognized sensitive data types (Milne et al. 2017; Markos et al. 2017) like voter files (Rubinstein 2014), cell phone records (Richards and King 2014), and large-scale surveys (Kenny et al. 2021). Comparing individuals' acceptable use perceptions for their social media data with these other types of sensitive data opens opportunities for us to learn from existing respect-for-persons best practices aside from explicit consent developed to support sensitive data. Our survey study thus addresses the following research questions:

- RQ1: How do participants perceptions of acceptable social media data use *compare to other types* of sensitive data about them?
- RQ2: How do participants' perceptions of acceptable social media data use relate to the *data analyst*, their *purpose* for using the data, and perceptions of *sensitivity*?

Answering RQ1 allows us to assess social media data's relative sensitivity, and to understand whether other sensitive data types are an appropriate context from which to seek wisdom about how researchers can approach social media data use. Answering RQ2 allows us to clarify sensitivity and other variables' relationships to acceptable use. Together, the answers to these questions provide insights for social media researchers about which best practices to follow when working with social media data and to ensure respect-for-persons.

Factors Affecting Perceptions of Acceptable Use

Prior work argues that individuals' willingness to share their personal data is a function of: data sensitivity, who will use the data (data analyst), what users hope to gain and who who else will benefit from using the data (data use purpose), which data they will use (data type), and personal characteristics of the person sharing the data (data sharer characteristics). The following subsections reviews each of these factors.

Data Sensitivity: Risk Perception

Data sensitivity describes how risky a person perceives sharing their data to be. In their work on consumer willingness to share data with marketers, Milne et al. (2017); Markos et al. (2017) identify four types of risk that may be particularly important for people's characterization of data as sensitive: monetary, psychological, physical, and social risk. A person's perceptions of how sensitive a type of data is – or how much risk they perceive they will incur by sharing that type of data – may inform how willing that person may be to share their data. However, while this correlation is implied by marketing literature, the existence and quality of this relationship have yet to be empirically evaluated. Further, whether a person characterizes data about themselves as sensitive is not a fixed characteristic, but rather can change according to the contexts in which data might be used and the data sharer's personal characteristics (Milne et al. 2017; Markos et al. 2017).

Data Analyst: Known Identities

Knowing who will use their data shapes data sharers' perceptions of acceptable use. People tend to find use of their social media data by *known* data analysts more acceptable than *unknown* data analysts. While social media users try to limit who sees and uses their data to only intended audiences, their data are still often seen by unintended or unknown audiences (Marwick and Boyd 2010). When people learn that these unintended audiences – among them researchers – use their data, they tend to find this use less acceptable than if their data is used by intended audiences (Williams et al. 2017).

Beyond known versus unknown data analysts, other data analyst identities can further affect sharers' perceptions of acceptable use. For example, Gilbert et al. (2021) asked respondents to rate the appropriateness of their personal Facebook data's use for research according to the discipline using it. They found that respondents were more concerned about studies in Computer Science, Gender Studies, and Psychology using their data than studies in Health Sciences. A related study about UK public health research shows that participants were much more willing to share their personal data for research by the UK's National Health Service than with a commercial company (Hill et al. 2013). Aside from explicitly intended audiences, people are most comfortable with their data being used by health researchers relative to other analysts. Overall, whether a data analyst is known or unknown and which discipline or professional domain they are affiliated with can affect individuals' perceptions of acceptable use.

Data Use Purpose: Benefit and Process

Literature suggests that individuals' perceptions of acceptable use also vary across data use purposes like health research or marketing. For example, in their study about UK public health research, Hill et al. (2013) found that when participants were told that mandating consent could lead to selection bias and adversely impact public health research, participants were more willing to share their data without explicit consent. For both health data (Howe et al. 2018; Tully et al. 2018; Hill et al. 2013) and social media data (Chen et al. 2021; Gilbert et al. 2021), how much participants believe that sharing their data will contribute to a purpose that will benefit society affects how acceptable they find the use of their data.

In addition to public benefits, people are more likely to find data use acceptable when it offers personal benefits like discounts or personalized service. Researchers studying public conversations about privacy controversies found that many discussants understand themselves and their data as a “product” that for-profit companies use for the purpose of making money, and in return, they receive some digital service – a personal benefit (Fiesler and Hallinan 2018). Discussants found this type of data use purpose acceptable. In another example, Dubois et al. (2020) study of journalists’ uses of social media data suggests that the more social media users want to feel heard, the more likely they are to find journalists’ use of their data acceptable. The personal benefits of receiving digital services and “feeling heard” mediated individuals’ perceptions of different data use purposes’ acceptability.

Beyond general use purpose (e.g., for public health research, for marketing, for public awareness, etc.), people’s understanding of exactly *how* their social media data will be used also affects their perceptions of its acceptable use. When participants know which analysis methods and data security measures researchers will use, they feel better about their data being used for research (Howe et al. 2018). Fiesler and Hallinan (2018) found that people who indicated understanding how their social media data would be used, including for research, were less concerned about its use relative to those who were *unaware* of how their data was later used. In general, when individuals are asked explicitly for their permission and understand what research will be conducted (i.e., data use purpose), they usually agree that their social media data can be used in research (Fiesler and Proferes 2018; Gilbert et al. 2021).

Data Type: Keys to Personal Identity and Social Networks

In addition to who will use their data and for what purposes, sharers care about *which* specific data will be used. Two studies examined US consumers’ willingness to share different types of information with marketers (Milne et al. 2017; Markos et al. 2017). Using a nationally representative survey, they found that respondents were just as unwilling to share their credit card number, financial account details, and driver’s license information as they were to share their social network profile, profile picture, and information about their friends or family. Respondents also considered their social media profile more sensitive – or risky to share – than basic demographics such as height, place of birth, and their occupation. Gilbert et al. (2021) asked respondents to rate the appropriateness of uses of their personal data for research by type of data. They found that respondents were most concerned about researchers using their photos and videos, data about sexual habits, data about preferences and behaviours, and posts about their friends or family members. These studies show that people weigh the riskiness and acceptability of sharing types of social media data differently.

Individual Data Sharer Characteristics

Research reports that perceptions of acceptable data use also vary based on a data sharer’s personal characteristics. For example, studies by Fiesler and Proferes (2018); Gilbert et al. (2021); Kass et al. (2003) find that how much people trust institutions – a type of data analyst – affects whether individuals are okay with those institutions using their social media data. For both health data (Howe et al. 2018; Tully et al. 2018; Hill et al. 2013) and social media data (Chen et al. 2021; Gilbert et al. 2021), how much people trust researchers in general also affects how acceptable they find their data’s use. Studies find mixed results concerning the effects on demographic characteristics on perceptions of acceptable use. For example, Fiesler and Proferes (2018) found that demographic characteristics have no statistically significant effect on survey respondents’ attitudes toward their Twitter data being used in various types of research. In contrast, Gilbert et al. (2021) showed that gender, age, education level, and frequency of social media use have significant effects on individuals’ attitudes toward their Facebook data being used in various type of research. Markos et al. (2017)’s comparative study of sensitivity perceptions in Brazil and the US found that perceptions do vary based on an individual’s country of residence (i.e., Brazil or the US) and age affect their willingness to share personal data. Milne et al. (2017) also reported significant effects for sex and education level on willingness to share data in their US-based survey. Finally, literature shows that pre-existing attitudes toward privacy (Fiesler and Proferes 2018) may affect perceptions of acceptable use. These data sharer characteristics can mediate the effects of the data analyst, data use purpose, and data type on individuals’ perceptions of acceptable use.

Summary of Factors Affecting Perceptions of Acceptable Use

Overall, social media users’ perceptions of of whether using their data is acceptable may be mediated by how sensitive they perceive their data to be, the data analyst, data use purpose, data type, and their personal characteristics. This web of factors shapes the challenges that researchers face in balancing respect for persons with research needs. Our study draws from existing scholarship’s insights on sensitive data and acceptable use to learn specifically about how people feel about their *social media data’s* use relative to other data types that people have indicated is sensitive to them like health records (Hill et al. 2013) and location (Martin and Nissenbaum 2020).

Methods

To understand participants’ perceptions about acceptable social media data use, we surveyed 1018 people through Qualtrics panels and Mechanical Turk. Qualtrics recruitment targeted US adults who posted publicly to social media at least once per week. We also used quotas for racial identity to ensure our sample was at least 10% African American individuals and at maximum was 80% white-only. We did not use targeting in recruiting MTurk participants.

We used statistical analyses to identify patterns in survey responses, specifically regarding (a) the sensitivity of individuals’ online identifiers relative to other personal identifiers, and (b) whether or not participants perceive a *particular data analyst* using a *specific type of data* for a *purpose* is acceptable or not.

Survey Population and Sample Size

We used a factorial ANOVA to estimate the appropriate size of our survey sample. The ANOVA indicated we need at least 306 respondents to detect significant differences in the effect of the interaction of analyst, data type, and purpose[*]. We contracted with Qualtrics to solicit responses from 586 survey panelists. Using a Qualtrics panel enabled us to set minimum quotas for our independent variables (e.g., non-white respondents) to ensure variability. We then used the same instrument to survey 432 crowdworkers through Amazon's Mechanical Turk. We recruited panels through both Qualtrics and Mechanical Turk to determine whether recruitment platform influenced findings. In our analyses, we include random effects for source to detect differences between the samples and their impacts on our observations (McKone and Lively 1993).

Instrument Design and Variables

Our survey instrument measured relationships between acceptable use and relevant constructs like data sensitivity, and personal characteristics like trust in institutions. Table 1 summarizes the various measures we included in our instrument. The following subsections explain how we developed each measure.

Variable	Measure	Source
Acceptable Use	Binary (y/n)	Gilbert et al. (2021); Fiesler and Proferes (2018); Martin and Nissenbaum (2020)
Data Sensitivity	10-pt. scale	Milne et al. (2017)
Data Sharer Char.	7-pt Scale/SA	NORC (2021)
Trust in institutions		
Dig. privacy behav.	5-pt scale/SA	Steinbart et al. (2017)
Dig. privacy concern	5-pt sclae/SA	Steinbart et al. (2017)
Social media use	Binary (y/n)	Auxier and Anderson (2021)
Social media freq.	6-pt MCQ/SA	Auxier and Anderson (2021)
Income	8 levels/SA	ANES (2019)
Education	8 levels/SA	ANES (2019)
Gender	6-pt MCQ/MA	Smith et al. (2016)
Race	9-pt MCQ/MA	ANES (2019)
Sexual Orientation	5-pt MCQ/MA	ANES (2019)
Age	Manual Entry	ANES (2019)
State of Residence	Manual Entry	ANES (2019)

Legend:
MCQ/MA : Multiple Choice Questions / Multiple Answers
MCQ/SA : Multiple Choice Questions / Single Answer

Table 1. Variables included in our survey instrument

Dimension	Source
Data Analyst	
Academic researchers	Fiesler and Proferes (2018)
Journalists	Hill et al. (2013)
Social media companies	Dubois et al. (2020)
Data Type	
Social media content	Milne et al. (2017)
Cell phone location	Arie and Mesch (2016); Ratti et al. (2006)
Grocery store purchases	Ozgormus and Smith (2020)
Security camera	Schiff et al. (2009)
Voter file	Fraga and Holbein (2020)
Survey data	Hughes et al. (2021)
Data Use Purposes	Gilbert et al. (2021)
Social/behavioral research	Hui et al. (2009)
Research abt. natural world	Silva et al. (2018)
Interven. to change behav.	Kramer et al. (2014)

Table 2. Dimensions we varied in constructing our acceptable use scenarios

Acceptable Use We developed a measure of acceptable use perceptions motivated by existing work that uses scenarios. However, existing acceptable use scenarios did not explicitly vary three constructs that other literature proposes affect acceptable use perceptions ("data analyst", "data type", and "data use purpose"). Therefore, we developed our own set of scenarios to evaluate the effects of these three constructs (see Table 2 for a summary of how we operationalized these constructs). Specifically, we varied scenarios according to three *data analysts*: academic researchers, social media companies (Hill et al. 2013), and journalists (Dubois et al. 2020). We varied *data types* to include social media content (Milne et al. 2017), cell phone location (Arie and Mesch 2016; Ratti et al. 2006), voter file (Fraga and Holbein 2020), survey (Hughes et al. 2021; **?**), security camera(Schiff et al. 2009), and grocery store purchase data (Ozgormus and Smith 2020) because they are used widely in contemporary research. We included three *data use purposes*: social/behavioral research, research about the natural world, and interventions designed to change individual behavior (Gilbert et al. 2021). Scenarios describing these purposes include real-world research use such as vegetation phenology (Silva et al. 2018), emotional contagion (Kramer et al. 2014), and consumer spending (Hui et al. 2009). They also mirror vignettes used in prior work on attitudes toward Facebook and Twitter users' data (Gilbert et al. 2021; Fiesler and Proferes 2018) and individuals' location data (Martin and Nissenbaum 2020). We requested binary responses rather than scale-based responses to our questions about whether data use was acceptable in a scenario because scales are more difficult for respondents to interpret and answer (Singer et al. 2010; Couper et al. 2006).

Data Sensitivity In studying the relative sensitivity of respondents' data, we include items from Milne, et al.'s (2017) study of information sensitivity and willingness to provide it to various institutions for different purposes. Their study evaluated respondents' relationships to a specific kind of social media data: the *online screen name*. Evaluating respondents' relationships to their online screen name does not encompass all types of social media data. However, this particular social media data type, like a driver's license number, can be a key to accessing other data about a person. Online screen names thus offer a point of comparison with similarly sensitive data types. We mirror Milne et al. (2017)'s focus on the online screen name to facilitate comparisons to earlier work and other types of sensitive data. Sensitivity acts as a dependent variable in analyses responding to RQ1, and independent variable in analyses responding to RQ2.

[*] Specifics of the ANOVA are available in the Appendix

Independent Variables To understand whether perceptions of acceptable use vary by mediating factors, we included independent variables motivated by existing literature. Based on existing research, we expected participants' perspectives to vary based on personal characteristics like their trust in institutions generally (NORC 2021), their existing privacy practices and concerns (Steinbart et al. 2017), their demographics (ANES 2019; Smith et al. 2016), and their social media use (Auxier and Anderson 2021). Assessing these independent variables' effects also allows us to compare our findings with existing research on personal data sharing and sensitivity (Fiesler and Proferes 2018; Gilbert et al. 2021; Milne et al. 2017; Kass et al. 2003).

Analysis

We used generalized linear mixed models (GLMM) – specifically a cumulative link mixed model (CLMM) for ordinal outcome variables and a mixed effects logistic regression (MELR) for binary outcome variables – to analyze our survey's results. Through the ability to include random effects, GLMMs enabled us to understand whether individual variation among participants, in addition to personal characteristics like demographics and institutional trust that we measured, impacted participants' responses. We included a random effect for response platform (Qualtrics or MTurk) and for individual respondent. We estimated models for different combinations of independent variables, their interactions, and control variables. We include the models of best fit, determined by ANOVA, in the main text below.

Limitations

Our survey method likely underestimates individuals' agreement with various data uses because they may not be familiar with ways that data are used or the methods employed in analysis. For instance, Tully et al. (2018) and MORI (2016) found that users increased their willingness to share health data after they understood the potential for public benefit and data security measures.

We adopted the variable 'online screen name' from prior surveys. Had we included the specific data types we asked about in scenarios (e.g., social media posts) in the sensitivity questions, we may have gotten different results. We chose 'online screen name' because it enabled comparison to prior studies (Markos et al. 2017; Milne et al. 2017), and because it is analogous to other linkable personal identifiers such as driver's license number. Future work could examine the sensitivity of specific types of social media data.

Results

Table 3 offers summary descriptive statistics about our sample. We address each research question in its own section below. In each case, we have provided the regression models of best fit.

RQ1: How do participants perceptions of acceptable social media data use compare to other types of sensitive data about them?

We calculated CLMM using the ordinal package in R (Christensen 2019) to compare the relative sensitivity of one type of social media data (one's screen name) with other types of potentially sensitive data. Our results, presented in Table 4 and Figure 1, show that a screen name is more sensitive than demographic details such as race (OR = 0.36, $p < 0.001$), religion (OR = 0.50 $p < 0.001$), and weight (OR = .47 $p < 0.001$), but less sensitive than identifiers such as a driver's license number (OR = 4.67, $p < 0.001$) or data from one's medical history (OR = 4.46, $p < 0.001$). Figure 1 also shows that individuals exhibited more variation in the sensitivity of their online screen name than other types of data such as fingerprints or medical history that may also provide access to data about their behaviors.

Variables	Levels	N	%
Man	Yes	551	54%
	No	456	45%
Straight	Yes	844	83%
	No	161	16%
Race	Black	90	9%
	White	765	75%
	Other	140	14%
Age	18 - 24	345	34%
	35 - 64	398	39%
	65 and over	275	27%
Education	Less than college degree	342	34%
	College degree	486	48%
	Graduate degree	190	19%
Income	< $40K	394	39%
	$40K - 60K	271	27%
	> $60K	353	35%
	Max	**Mean**	**St. Dev.**
Trust	42	24.9	9.4
Digital privacy	12	10.5	1.9
Privacy behavior	21	15.4	3.3

Table 3. Overview of our survey sample's personal characteristics

Table 4 also indicates that respondents who were men and with higher trust in institutions were *less* likely to find their online screen name sensitive. Respondents from Qualtrics, who had higher digital privacy concerns, more privacy behaviors, were neither Black nor White, were straight, older, more educated, and had a higher income were *more* likely to find their online screen name sensitive.

RQ2: How do participants' perceptions of acceptable use relate to data type, data analyst, data use purpose, and sensitivity?

We calculated a mixed effects logistic regression using the glmer function from lme4 (Bates et al. 2015) to understand whether respondents indicated a particular combination of data analyst, data type, and purpose was acceptable. In this model, the dependent variable was whether respondents answered "yes" to a specific question, and the questions were the only fixed effect. We included random effects for respondent and source (Qualtrics or Mechanical Turk). The coefficient for source's random effects was nearly zero, indicating that between-subject differences based on source could be almost entirely explained by the other variables we assessed in our model. We also included demographic controls and scales for our questions about trust, digital

Variables of Interest	**OR**	**Control Variables**	**OR**
Data Type		**Recruitment Site**	
Drivers license no.	4.67^{***}	Qualtrics	1.42^{***}
Emotions	1.15	**Behavior Scales**	
Family friends	3.80^{***}	Trust	0.97^{***}
Fingerprint	5.25^{***}	Dig. priv. concern	1.28^{***}
Handwriting	1.71^{***}	Priv. behav.	1.09^{***}
Height	0.30^{***}	**Demographics**	
License plate no.	2.55^{***}	Man	0.85^{***}
Medical history	4.46^{***}	Race Other	1.38^{**}
Mental health	3.15^{***}	Race White	1.14
Mat. maiden name	2.35^{***}	Straight	1.45^{***}
Race	0.36^{***}	Age Bracket	1.17^{**}
Religion	0.50^{***}	Education Level	1.27^{***}
Vehicle registr. no.	3.10^{***}	Income level	1.13^{**}
Voice print	2.54^{***}		
Weight	0.47^{***}		

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 4. Predicting the sensitivity of various data types; controls include demographic variables; “online screen name” is the reference category. ’OR’ stands for odds ratio.

privacy concerns, and privacy behaviors. According to ANOVA analyses, the model of best fit did *not* include sensitivity as a predictor.

The results (see Table 5 indicate respondents found only one combination clearly unacceptable (i.e., significantly lower odds ratio):

> Is it ok for journalists to use posts you’ve deleted from social media in a story about natural disasters?

Overall, respondents said that it was *more* acceptable to use their social media data than other types of data (see Table 6).

However, respondents generally found academic researchers using social media data about them acceptable except for two scenarios (no significant difference between acceptable and not acceptable):

- images you’ve uploaded to social media to train facial recognition software?
- metadata from your photos in social media to create a public map of peony gardens in your area?

Respondents with higher scores on the digital privacy concern scale were less likely to agree that it their data’s use is acceptable (OR = 0.892). Those who had higher institutional trust scores (OR = 1.041) and privacy behavior scores (OR = 1.112) were more likely to indicate that using their data was acceptable.

Among our demographic controls, only straight (OR = 0.265) and older respondents (OR = 0.618) were less likely to answer that it was acceptable for their data to be used. Men were significantly more likely than women and other gender identities (OR = 1.980) to say it was acceptable for their data to be used. Individuals of races other than White and Black (OR = 2.116) were also more likely to find their data’s use acceptable. We observed no significant effects for income level.

We also fit a model where we collapsed our data type and purpose variables into categories. According to ANOVA analyses, we found that the best model included data type, use purpose, data analyst, behavioral scales, and demographics but *not* sensitivity as predictors of respondents finding their social media data’s use by researchers acceptable for social/behavioral research. In this model, presented in Table 6 and Figure 2, we use *academic researchers*, *social media data*, and *social research* as the reference categories for data analyst, data type, and purpose respectively. We included a random effect for respondent and another for source (Qualtrics or Mechanical Turk). The variance between sources was nearly zero, indicating that between-source differences could be almost entirely explained by the other variables we assessed in our model. Table 6 shows that respondents were more accepting of their data being used for social research than to generate interventions or research about the natural world.

We saw similar patterns among the control variables (trust, digital privacy, privacy behavior, and demographics) in both of our models. Individuals with concerns about digital privacy (e.g., they hesitate to provide information when its requested) were less likely to be accepting of their data being used. Individuals who generally trust institutions and governments accepted their data being used by these entities. Individuals who engaged in more privacy-protecting behaviors (e.g., removing cookies from their web browser, watching for ways to control what emails they receive) were more likely to agree that their data could be used. We find that information about how sensitive someone considers their online screen name does not provide statistically significant information about how acceptable they think a combination of data, user, and purpose are.

Discussion

Our results suggest that individuals are generally accepting of academic researchers using social media data and that online screen names are less sensitive than many other types of data, including individual identifiers such as driver’s license numbers and medical histories, but *more* sensitive than height, weight, race, and religion. Individuals indicated that academic researchers using their data was acceptable in more scenarios than journalists or social media companies doing so. One implication of these findings is that social media researchers can leverage the expertise and practices of researchers who use other types of sensitive or moderately sensitive data such as health records. However, one potential limitation of our study and Gilbert et al. (2021)’s is that our survey instruments clearly communicated the user, data, and purpose. In doing so, the instruments may have been specific enough that users found these scenarios acceptable; had we asked more generally about social media researchers using their data, they may not have been as accepting. We address findings about the comparative sensitivity of social media data and the relationships between sensitivity and acceptable data use below.

Sensitivity Comparison

Prior work on social media data use in research (e.g., Fiesler and Hallinan 2018; Gilbert et al. 2021) examines social media data on its own rather than in the context of private and/or sensitive personal information. Our survey

Data Type + Data Use Purpose	Researchers	Journalists	SMCs
Social media post content			
Antiracism bots	2.506***	1.514***	2.236***
Interventions to feel better	1.763***		1.480**
Language change	3.915***	1.894***	2.803***
Misinformation	3.886***	1.358*	2.698***
Predict elections	2.205***	1.451**	1.341*
Predict risk of harming others	2.421***		1.775***
Predict risk of harm to self	1.751***		1.488***
Exhibit about protest	1.917***	1.191	1.293*
Show relevant ads	2.141***		1.825***
Recognize emotions		1.015	
Deleted social media post content			
Information in natural disasters	1.304*	0.738*	1.099
Social media post timestamps			
COVID-19 spread	3.973***	1.861***	3.145***
Social media post location data			
COVID-19 spread	2.784***	1.499***	2.436***
Social media image content			
Train facial Recognition	1.198	0.885	1.038
Vegetation in national parks	3.745***	1.810***	2.251***
Social media image metadata			
Map of peonies	1.259	0.866	1.258
Security camera footage			
Train facial recognition	1.763***		1.232
Cell phone location			
Commuting patterns	1.449**	0.815	1.023
Grocery store purchases			
Send you coupons	3.386***		2.382***
Influence of other shoppers	2.316***	1.117	1.726***
Voter file			
Voter turnout	2.129***	1.029	1.156

Controls	OR	Controls	OR
Behavior Scales		**Demographics**	
Trust	1.041***	Straight	0.265***
Digital privacy concern	0.892**	Age brackets	0.618**
Privacy behaviors	1.112***	Man	1.980***
Demographics		Education level	1.356
Race other	2.116*	Income level	1.058
Num.Obs.	55 049	BIC	51 021.3

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 5. Is it ok for this data user to use this data type for this purpose? “No” is the reference category; ORs greater than 1 indicate respondents were more likely to say a use was “ok”. Missing cells mean that combination was not included on the survey.

instrument allowed us to compare individuals’ reports about the sensitivity of different types of data so that we can understand how social media data is similar to (or dissimilar from) other data often used in research. Looking at a specific type of social media data — one’s online screen name or the “key” to accessing one’s social media data, similar to one’s “offline” name – respondents suggest that their online screen name is more sensitive than demographic characteristics (e.g., one’s race) and less sensitive than other personally-identifiable data that can be linked to an individual and her behavior (e.g., one’s driver’s license number). Respondents also found online identifiers less sensitive than mental health and health records. There was also more variation in respondents’ perceptions of their online screen name’s sensitivity relative to other types of data like fingerprints or medical history that may also provide access to data about their behaviors. This variability may indicate uncertainty among individuals or true variation in our sample, and future work could attempt to verify this distribution and its causes.

Acceptable Use Comparison

To evaluate perceptions of social media data’s acceptable use relative to other widely used sensitive data types, we compared social media data with examples like cell phone data and voter files that carry varying re-identification risks. Compared to these other types of sensitive data, respondents were *most* comfortable with their social media data being used, and *least* comfortable with their cell phone location data being used. This finding resonates with prior work on location data arguing that individuals expect privacy even in public (Martin and Nissenbaum 2020), and this expectation extends to automatically collected data like location captured by cell phones and social media. Relative to social media data, respondents were also less likely to agree that it’s acceptable to use their voter file or security camera footage of them. While voter files are widely used in political science research (e.g. Hughes et al. 2021; Nyhan et al. 2017), researchers have made important efforts to protect data sharers’ privacy expectations including disclosure risk

Variable	Odds Ratio
Data Type	
Cell phone location	0.519***
Grocery store purchases	0.743**
Security camera	0.913
Voter file	0.705***
Data Use Purpose	
Intervention	0.643***
Natural research	0.622***
Data Analyst	
Journalists	0.490***
Social media companies (SMC)	0.730***
Behavior Scales	
Trust	1.044***
Digital privacy concern	0.880**
Privacy behaviors	1.130***
Demographics	
Man	1.945***
Race other	1.645
Race white	1.224
Straight	0.235***
Age brackets	0.646*
Education level	1.401
Income level	0.978
Interactions	**Odds Ratio**
Data Type × Data Analyst	
Cell phone location × journalists	1.161
Grocery store purchases × journalists	1.058
Security camera × journalists	0.924
Voter file × journalists	0.979
Cell phone location × SMC	0.917
Grocery store purchases × SMC	1.026
Security camera × SMC	0.828
Voter file × SMC	0.700*
Data Use Purpose × Data Type	
Intervention × grocery store purchases	2.270***
Data Use Purpose × Data Analyst	
Intervention × journalists	1.273**
Natural research × journalists	1.156
Intervention × SMC	1.204*
Natural research × SMC	1.052
Data Use Purpose × Data Type × Data Analyst	
Intervention × grocery store purchases × SMC	0.787
Num.Obs.	43 784
BIC	40 662.7

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 6. Summary of acceptable use where data type, date use purpose, and data analyst are aggregated. Academic researchers (data user), social media data (data type), and social research (purpose) are the reference categories.

mitigation techniques like aggregation to avoid revealing personally identifiable information.

Sensitivity and Acceptable Use

We found that sensitivity is not a good predictor of whether individuals thought it was acceptable for researchers to use social media data. This result is somewhat unexpected given prior literature that suggested sensitivity mediates acceptable use (Markos et al. 2017). One possible explanation is that 'online screen name' is not a useful example of social media data to ask about. It is possible, for example, that respondents may not understand how much data can be accessed when one's online screen name is known (e.g., one's tweet history). Another explanation is that for other types of data, sensitivity may predict acceptable use, but for social media data, acceptable use is a function of the data analyst, data type, and purpose of data use (Gilbert et al. 2021).

Concerning data analyst, type, and use purpose, our findings echo earlier results indicating users accept their social media data being used in research when told about who will use it and for what purpose (Fiesler and Proferes 2018; Gilbert et al. 2021). Our respondents were generally more accepting of researchers using their data than social media companies or journalists (see Figure **??**). We expect that this pattern holds because users better are able to imagine benefits from research. Given the increasing distrust of journalists in the United States (Fink 2019; Usher 2019), it's not surprising that our respondents did not welcome journalists using their data. In line with prior research about sensitive data use in health research (Kass et al. 2003) and marketing (Markos et al. 2017), our respondents were more comfortable sharing data with researchers looking to produce social benefit and understanding than with for-profit companies using their data for similar purposes.

Given these findings, rather than looking to similarly *sensitive* data for guidance on respect-for-persons practices with social media data, our research points us to data whose analysts have *clearly communicated their use purposes' benefits* and *cultivated trust* among prospective sharers. In fact, for both our survey and Gilbert et al. (2021), articulating data use scenarios explicitly may drive much of the acceptance individuals expressed.

Acceptable Use and Personal Characteristics

In our results, men were more likely to report that researchers could use their data without explicit permission, and that most uses of their data were acceptable. Related prior research on individuals' willingness to share data with marketers found that sensitivity was a function of perceived privacy controls and cultural context such as masculinity values and long-term orientation (Markos et al. 2017). The importance of masculinity values (e.g., " It is more important for men to have a professional career than for women.") in predicting sensitivity may explain why we observed differences between men and other gender identities. Women and members of gender identity minorities face greater risks when engaging in social media (Boulianne et al. 2021; Duggan 2017); those risks may lead them to be more conservative in their data sharing beliefs. As Mikal et al. (2016) point out, we must carefully consider who may opt-out of using social media publicly whenever we think of social media data.

We also found that older adults and straight respondents oppose the use of their data without explicit permission when asked about research generally. In these dimensions of identity, members of marginalized demographic groups may be more willing to share their data because they seek inclusion or because they see efforts to avoid surveillance as futile [†]. Opt-in consent formats introduce selection bias (Hill et al. 2013; Kho et al. 2009), and our results indicate that some groups will be over-represented in opt-in samples.

Implications for Researchers

Our results have two implications for realizing respect for persons in social media research. First, whenever possible, researchers should elicit informed consent from social media users to use their data in research. However, if researchers use data only from those individuals who have provided explicit permission and who are generally accepting of their data being used in research, their data will likely skew male, younger, more educated, and less straight. While bias in data cannot be eliminated and is not inherently bad, demographically-biased social media data limit their utility for population-level studies. Because individuals are generally more comfortable with population-level research than with individual-level research (Williams et al. 2017; Fiesler and Proferes 2018), this tension is especially important for researchers to consider. Given the burden of obtaining informed consent and the biases it introduces, when it is not possible to obtain, researchers should work to anonymize data as much as possible to reduce the risks of reidentification. Existing work such as Williams et al. (2017), the AOIR Ethics Guidelines (Franzke et al. 2020), and the STEP framework (Mannheimer and Hull 2018) provide useful tools for thinking through research processes, when and how to get consent, and how to mitigate risks to individuals.

Second, as Kass et al. (2003) suggest, educating the public about why research is important and why it requires their data is vital to ensuring individuals' comfort with their data being used. As Chen et al. (2021), Gilbert et al. (2021), and now our work show, how much data sharers trust researchers affects their perceptions of acceptable use. People who place more trust in institutions were more likely to accept their data being used. We can learn from health research that has been able to explain to individuals how and why their medical records are necessary for understanding diseases such as cancer. People now tend to find their data's use for health research more acceptable than for other uses (Gilbert et al. 2021; Hill et al. 2013).

Researchers who leverage social media data can engage in similar outreach and engagement efforts to understand individuals' hesitations and preferences. We do not, however, suggest that researchers try to cajole or coerce potential participants. Instead, social media researchers must work to ensure that their research *does* actually provide social benefits worthy of individual risks, and that they are consequently able to articulate the significance of their work so that individuals can decide whether that benefit is worth their risks. As Sloan et al. (2020) argue, the principles outlined in the Belmont Report apply even to social media research, and researchers have a responsibility to ensure that individuals understand their own data, how it could be used, and the risks associated with use.

In one example, Xafis (2015) demonstrated how they helped individuals understand how their data could be used and potentially competing interests between researchers and individuals represented in data. Their research showed that individuals could understand data linkage processes and the potential trade-offs quickly. Because it is not feasible to educate each potential participant that might share their social media data, the burden of education lies on researchers collectively. Researchers need to articulate the social benefits of their work so that individuals understand why disclosure and reputation risks are worth taking; if the benefits do not outweigh the risks, researchers need to be willing to abandon or avoid particular projects.

Conclusion

We began this article with an example of Twitter's new no-code API, highlighting how increased accessibility of Twitter data raises questions for social media researchers such as: *should* they use social media data in their research? If so, *when* and *how*? As with other large-scale data, it is not always possible to ask all prospective study participants to explicitly consent to use their social media data in research. However, through surveying individuals, our study offers a benchmark of individuals' perceptions of their social media data's sensitivity and acceptable use from which social media data researchers can cultivate general respect-for-persons practices. We show that people generally find their social media data moderately sensitive relative to other widely used data types. People generally find it acceptable for researchers to use their social media data but prefer that researchers clearly articulate the benefits of their work. Individuals are most concerned about *who, why*, and *which data* will be used. When these factors are clearly communicated, individuals are more likely to find their data's use for research acceptable.

Given our findings that individuals find their social media data moderately sensitive, our study invites social media researchers to learn from well-established best practices for using sensitive data and increasing public awareness about the benefits of research with social media data. Researchers must be clear with themselves and with the public about why social media data is necessary for their work and what benefits that work provides for society, especially for the individuals who carry risks by being included in the data. By using contemporary strategies to balance society media data's use risks to individuals with the benefits to society, social media researchers can more effectively realize respect for persons, ensuring greater public support for significant advances in research.

Acknowledgements

We are thankful for our colleagues' help and feedback on this project: Nazanin Andalibi, Ricardo Punzalan, and Kat Roemmich helped draft the survey questions. The Statistical Design Group directed by Brady West and James Wagner consulted on the structure and statistical design of our survey instrument. Lizhou (Leo) Fan provided comments on an earlier draft. Andrew Schrock provided editing services. Staff at Consulting for Statistics, Computing, and Analytics Research (CSCAR) at the University of Michigan advised on statistical analysis.

This material is based upon work supported by the National Science Foundation under Grant No. 1839868. Additional funding

[†] See Benjamin (2019) for a thorough discussion of the differential impacts of surveillance among racial groups, for instance.

was provided by the Propelling Original Data Science (PODS) grant program from the Michigan Institute for Data Science.

References

ANES (2019) User guide and codebook. https://electionstudies.org/wp-content/uploads/2018/12/anes_timeseries_2016_userguidecodebook.pdf.

Antenucci D, Cafarella M, Levenstein M, Ré C and Shapiro MD (2014) Using social media to measure labor market flows.

Arie Y and Mesch GS (2016) Spatial distance and mobile business social network density. *Inf. Commun. Soc.* 19(11): 1572–1586.

Auxier B and Anderson M (2021) Social Media Use in 2021. URL https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

Bates D, Mächler M, Bolker B and Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1): 1–48. DOI:10.18637/jss.v067.i01.

Benjamin R (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Wiley.

Bernstein MS, Bakshy E, Burke M and Karrer B (2013) Quantifying the invisible audience in social networks. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13. New York, NY, USA: Association for Computing Machinery, pp. 21–30.

Boulianne S, Koc-Michalska K, Vedel T, Nadim M and Fladmoe A (2021) Silencing women? gender and online harassment. *Soc. Sci. Comput. Rev.* 39(2): 245–258.

Chen Y, Chen C and Li S (2021) Determining factors of participants' attitudes toward the ethics of social media data research. *Online Information Review* ahead-of-print(ahead-of-print).

Christensen RHB (2019) ordinal—regression models for ordinal data. R package version 2019.12-10. https://CRAN.R-project.org/package=ordinal.

Couper MP, Tourangeau R, Conrad FG and Singer E (2006) Evaluating the effectiveness of visual analog scales: A web experiment. *Soc. Sci. Comput. Rev.* 24(2): 227–245.

Documenting the Now (????) Social humans labels. https://www.docnow.io/social-humans/index.html. Accessed: 2021-12-14.

Dubois E, Gruzd A and Jacobson J (2020) Journalists' use of social media to infer public opinion: The citizens' perspective. *Soc. Sci. Comput. Rev.* 38(1): 57–74.

Duggan M (2017) Online harassment 2017. *Pew Research Center* .

Fiesler C and Hallinan B (2018) "we are the product": Public reactions to online data sharing and privacy controversies in the media. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, number Paper 53 in CHI '18. New York, NY, USA: Association for Computing Machinery, pp. 1–13.

Fiesler C and Proferes N (2018) "participant" perceptions of twitter research ethics. *Social Media + Society* 4(1): 1–14.

Fink K (2019) The biggest challenge facing journalism: A lack of trust. *Journalism* 20(1): 40–43.

Fraga B and Holbein J (2020) Measuring youth and college student voter turnout. *Electoral Studies* 65: 102086.

Franzke AS, Bechmann A, Zimmer M, Ess C and the Association of Internet Researchers (2020) Internet research: Ethical guidelines 3.0. Technical report.

Gilbert S, Vitak J and Shilton K (2021) Measuring americans' comfort with research uses of their social media data. *Social Media + Society* 7(3): 1–13.

Hemphill L, Russell A and Schöpke-Gonzalez AM (2020) What drives U.S. congressional members' policy attention on twitter? *Policy & Internet* .

Hill EM, Turner EL, Martin RM and Donovan JL (2013) "let's get the best quality research we can": public awareness and acceptance of consent to use existing data in health research: a systematic review and qualitative study. *BMC Med. Res. Methodol.* 13: 72.

Howe N, Giles E, Newbury-Birch D and McColl E (2018) Systematic review of participants' attitudes towards data sharing: a thematic synthesis. *J. Health Serv. Res. Policy* 23(2): 123–133.

Hughes AG, McCabe SD, Hobbs WR, Remy E, Shah S and Lazer DMJ (2021) Using administrative records and survey data to construct samples of tweeters and tweets. *Public Opin. Q.* 85(S1): 323–346.

Hui SK, Bradlow ET and Fader PS (2009) Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior. *J. Consum. Res.* 36(3): 478–493.

Kass NE, Natowicz MR, Hull SC, Faden RR, Plantinga L, Gostin LO and Slutsman J (2003) The use of medical records in research: What do patients want? *J. Law Med. Ethics* 31(3): 429–433.

Kenny CT, Kuriwaki S, McCartan C, Rosenman ETR, Simko T and Imai K (2021) The use of differential privacy for census data and its impact on redistricting: The case of the 2020 u.s. census. *Science Advances* 7(41): eabk3283. DOI: 10.1126/sciadv.abk3283. URL https://www.science.org/doi/abs/10.1126/sciadv.abk3283.

Kho ME, Duffett M, Willison DJ, Cook DJ and Brouwers MC (2009) Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ* 338: b866.

Kramer ADI, Guillory JE and Hancock JT (2014) Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl. Acad. Sci. U. S. A.* 111(24): 8788–8790.

Mannheimer S and Hull EA (2018) Sharing selves: Developing an ethical framework for curating social media data. *International Journal of Digital Curation* 12(2): 196–209.

Markos E, Milne GR and Peltier JW (2017) Information sensitivity and willingness to provide continua: A comparative privacy study of the united states and brazil. *Journal of Public Policy & Marketing* 36(1): 79–96.

Martin KE and Nissenbaum H (2020) What is it about location? *Berkeley Technol. Law J.* 35(1).

Marwick AE and Boyd D (2010) I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13(1): 114–133.

McKone MJ and Lively CM (1993) Statistical analysis of experiments conducted at multiple sites. *Oikos* 67(1): 184–186.

Mikal J, Hurst S and Conway M (2016) Ethical issues in using twitter for population-level depression monitoring: a qualitative study. *BMC Med. Ethics* 17: 22.

Milne GR, Pettinico G, Hajjat FM and Markos E (2017) Information sensitivity typology: Mapping the degree and type of risk consumers perceive in personal data sharing. *J. Consum. Aff.* 51(1): 133–161.

MORI I (2016) The One-Way mirror: Public attitudes to commercial access to health data. Technical report, Ipsos MORI.

NORC (2021) Documentation questionnaire. https://gss.norc.org/get-documentation/questionnaires. Accessed: 2022-2-24.

Nyhan B, Skovron C and Titiunik R (2017) Differential registration bias in voter file data: A sensitivity analysis approach. *Am. J. Pol. Sci.* 61(3): 744–760.

Obar JA and Oeldorf-Hirsch A (2020) The biggest lie on the internet: ignoring the privacy policies and terms of service policies of social networking services. *Inf. Commun. Soc.* 23(1): 128–147.

Office for Human Research Protections (OHRP) (????) Read the belmont report. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html#xbasic. Accessed: 2021-12-5.

Ordun C, Blake JW, Rosidi N, Grigoryan V, Reffett C, Aslam S, Gentilcore A, Cyran M, Shelton M and Klenk J (2013) Open source health intelligence (OSHINT) for foodborne illness event characterization. *Online J. Public Health Inform.* 5(1).

Ozgormus E and Smith AE (2020) A data-driven approach to grocery store block layout. *Comput. Ind. Eng.* 139: 105562.

Ratti C, Frenchman D, Pulselli RM and Williams S (2006) Mobile landscapes: Using location data from cell phones for urban analysis. *Environ. Plann. B Plann. Des.* 33(5): 727–748.

Richards NM and King JH (2014) Big data ethics. *Wake Forest L. Rev.* 49: 393.

Rubinstein IS (2014) Voter privacy in the age of big data. *Wis. L. Rev.* : 861.

Schiff J, Meingast M, Mulligan DK, Sastry S and Goldberg K (2009) Respectful cameras: Detecting visual markers in real-time to address privacy concerns. In: *Protecting privacy in video surveillance*. Springer, pp. 65–89.

Silva SJ, Barbieri LK and Thomer AK (2018) Observing vegetation phenology through social media. *PLoS One* 13(5): e0197325.

Singer E, Couper MP, Raghunathan TE, Antonucci TC, Burmeister M and Van Hoewyk J (2010) The effect of question framing and response options on the relationship between racial attitudes and beliefs about genes as causes of behavior. *Public Opin. Q.* 74(3): 460–476.

Sloan L, Jessop C, Al Baghal T and Williams M (2020) Linking survey and twitter data: Informed consent, disclosure, security, and archiving. *J. Empir. Res. Hum. Res. Ethics* 15(1-2): 63–76.

Smith TW, Marsden P, Hout M and Kim J (2016) General Social Surveys, 1972-2014 [machine-readable data file] .

Steinbart P, Keith M and Babb J (2017) Measuring privacy concern and the right to be forgotten. In: *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)*. Hawaii International Conference on System Sciences, pp. 4967–4976.

Steinfield C, Ellison NB and Lampe C (2008) Social capital, self-esteem, and use of online social network sites: A longitudinal analysis. *J. Appl. Dev. Psychol.* 29(6): 434–445.

Tully MP, Bozentko K, Clement S, Hunn A, Hassan L, Norris R, Oswald M and Peek N (2018) Investigating the extent to which patients should control access to patient records for research: A deliberative process using citizens' juries. *J. Med. Internet Res.* 20(3): e112.

Usher N (2019) Putting "place" in the center of journalism research: A way forward to understand challenges to trust and knowledge in news. *Journal. Commun. Monogr.* 21(2): 84–146.

Williams ML, Burnap P and Sloan L (2017) Towards an ethical framework for publishing twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology* 51(6): 1149–1168.

Xafis V (2015) The acceptability of conducting data linkage research without obtaining consent: lay people's views and justifications. *BMC Med. Ethics* 16(1): 79.

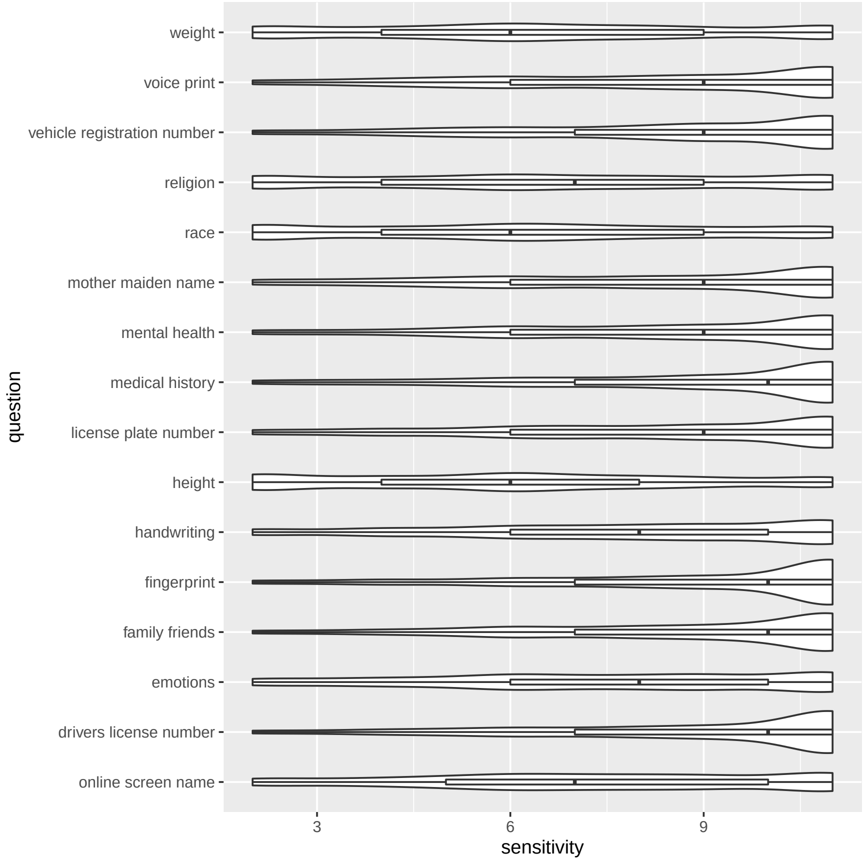

Figure 1. Violin plots showing the distribution, median, and quartiles for sensitivity of various types of data where 10 = 'very sensitive'

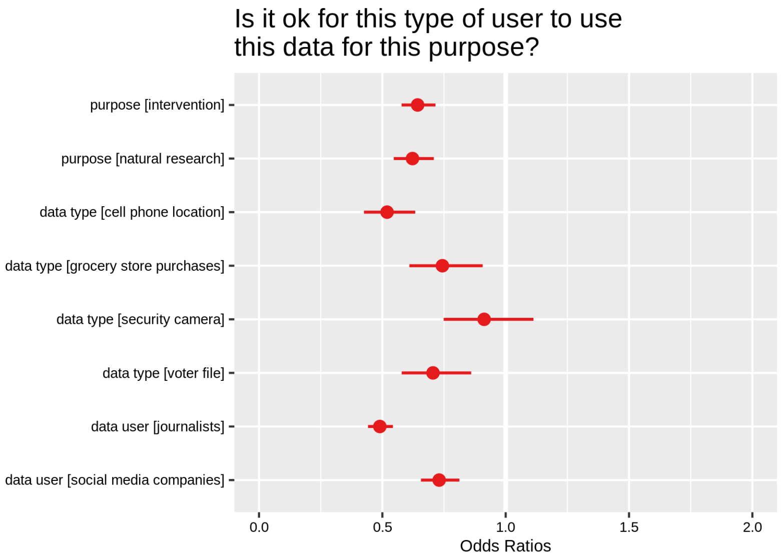

Figure 2. Odds ratio plot where DV = "is ok" on each data analyst, data type, purpose combination