A framework for privacy preserving digital trace data collection through data donation

Laura Boeschoten
Dept. of Methodology and Statistics, Utrecht University

Jef Ausloos
Institute for Information Law, University of Amsterdam

Judith E. Möller
Amsterdam School of Communication Research (ASCoR), University of Amsterdam

Theo Araujo
Amsterdam School of Communication Research (ASCoR), University of Amsterdam

Daniel L. Oberski
Dept. of Methodology and Statistics, Utrecht University
Julius Center, University Medical Center Utrecht

Abstract

A potentially powerful method of social-scientific data collection and investigation has been created by an unexpected institution: the law. Article 15 of the EU's 2018 General Data Protection Regulation (GDPR) mandates that individuals have electronic access to a copy of their personal data, and all major digital platforms now comply with this law by providing users with "data download packages" (DDPs). Through voluntary donation of DDPs, all data collected by public and private entities during the course of citizens' digital life can be obtained and analyzed to answer social-scientific questions – with consent. Thus, consented DDPs open the way for vast new research opportunities. However, while this entirely new method of data collection will undoubtedly gain popularity in the coming years, it also comes with its

own questions of representativeness and measurement quality, which are often evaluated systematically by means of an error framework. Therefore, in this paper we provide a blueprint for digital trace data collection using DDPs, and devise a "total error framework" for such projects. Our error framework for digital trace data collection through data donation is intended to facilitate high quality social-scientific investigations using DDPs while critically reflecting its unique methodological challenges and sources of error. In addition, we provide a quality control checklist to guide researchers in leveraging the vast opportunities afforded by this new mode of investigation.

Keywords: Digital trace data, Data donation, Total Error Framework, Informed consent, Privacy

Introduction

Digital traces left by citizens during the natural course of modern life hold an enormous potential for social-scientific discoveries (King, 2011), because they can measure aspects of our social life that are difficult or impossible to measure by more traditional means (Pentland, 2010). For example, classic sociological theory describes citizens' interactions (Coleman, 1990), but large-scale data on such interactions are only now becoming available from digital platforms (e.g. Szell et al., 2010). Similarly, experiments show that news reports can produce different opinions depending on the consumer's political motivations (e.g. Bolsen et al., 2014), but only through digital media we can now observe the simultaneous dynamics of consumed (mis)information, motivation, and opinion. With increased datafication and digitalization of our societies, the study of digital traces gains even more relevance. As more and more of our social lives happens on platforms, the digital traces we leave behind on those platforms become an important object of study (Papacharissi, 2010). Further examples of digital traces' potential abound, and indeed, digital trace data collected through Application Programming Interfaces (APIs) and web scraping have been used in many applications, including network analysis from mobile phone data (Blondel et al., 2015); price indexing from online shop prices (de Haan & Hendriks, 2013); political opinion and electoral success prediction from Twitter data (Jungherr, 2015; Schoen et al., 2013); and personality profiling from Facebook "likes" ((Kosinski et al., 2013); see also (Settanni et al., 2018) for an overview of similar studies).

In recent times, however, the faucet of social science data from APIs and web scraping has been decisively turned off by the relevant tech companies

(Bruns, 2019; Freelon, 2018; Perriam et al., 2020). Through mutual agreement and negotiation between academia and industry, new efforts to make such data available to social scientists are underway, for example through "Social Science One" (King & Persily, 2019). These data are now becoming available in aggregated form under strict privacy protections (D'Orazio et al., 2015; Messing et al., 2020). While this new collaborative model is useful for social-scientific investigation of certain research questions, it does not fit all purposes mentioned above and has raised concerns about the role of platforms (for an overview, see Halavais (2019)). First, by their very definition, the imposed data protection regulations ensure these data cannot address questions of individual (user-level) dynamics or networks (Oberski & Kreuter, 2020). Second, APIs provide public data only; much of digital trace data's putative power, however, lies in private data that is too sensitive to share, such as location history, browsing history, or private messaging (Quan-Haase & Young, 2010). Third, the available data generally pertain to a nonrandom subset of the digital platform's user group (e.g. Facebook or Twitter) which is not representative of many populations of social-scientific interest (Mellon & Prosser, 2017; Pfeffer et al., 2018). Fourth, for both approaches, the researcher is entirely dependent on the private company that holds the data; sudden retractions of this collaborative spirit can, and have, occurred, posing a risk to the research process (Bruns, 2019). In addition. there is no possibility to independently verify that the data is complete and checked for measurement errors. Finally, even when a data processing company decides to share data for scientific purposes, the citizens who actually generated those data are generally impossible to contact for their consent, in some cases putting a firm legal basis for further data analysis in question (for example, following article 6, EU General Data Protection Regulation, or similar laws in other jurisdictions). Data donation can be defined as the act of an individual who actively consents to donate their personal data for research (Skatova & Goulding, 2019). By definition, the issue of consent is overcome in a situation of data donation. Many initiatives already allow to use data donation for research purposes. For example Andrews et al. (2015) used an app that records phone use, Reeves et al. (2019) used an app that collects screenshots, Haenschen (2020) used an app that retrieved posts and liked pages on Facebook during a period of 6 months. Participants of a study by Araujo et al. (2017) installed software that tracked internet use on PCs and android tablets. Alternatively, Menchen-Trevino (2016) developed a browser extention to extract browsing history data, which has been used for example by Wojcieszak et al. (2021) and Weeks et al. (2021). These initiatives all allow for the collection of individual level private data.
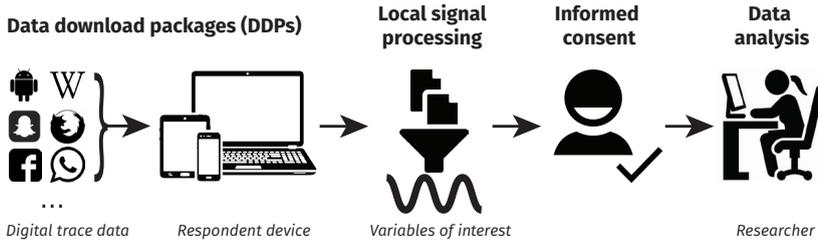
Figure 1. A workflow illustration how a respondent's data download packages (DDPs) can be leveraged for socialscientific research after local processing and informed consent.

Thereby, the issues of informed consent and (unknown) selectivity of the obtained sample can be overcome. However, these initiatives are generally limited in terms of platform from which data can be collected, in terms of how much data over time can be collected and are dependent on updates in software or operating systems.

In this paper, we present a new and alternative workflow for collecting and analyzing digital traces that overcomes all issues previously discussed, based on *data download packages* (DDPs). As of May 2018, any organisation subject to the GDPR, whether public or private, is legally required to provide all personal data to the data subject upon request, and in digital format (GDPR Article 15; Ausloos (2019)). Most major private data processing entities, comprising social media platforms as well as smartphone systems, search engines, photo storage, e-mail, banks, energy providers, and online shops comply with this right to data access by providing DDPs to the data subjects. Thanks to the GDPR, by far the majority of all digital traces left behind by people in the EU can be collected by means of DDPs. To illustrate, in the Netherlands 95.6% of the population had access to internet at home in 2020, 67.2% made use of social network sites, 86.1% send messages through digital platforms such as WhatsApp, and 83.8% used a form of mobile banking ("CBS StatLine," 2020). This means that DDPs containing digital traces on social network sites are available for 67.2% of the Dutch population, DDPs containing archives of digital (private) messsages are available for 86.1% of the Dutch population and bank transaction history is available for 83.8% of the Dutch population. DDPs can therefore be seen as a widely available sources to collect digital trace data and thereby a useful tool for researchers (Ausloos & Veale, 2020). Furthermore, to our knowledge, most large companies that operate internationally provide the same service to their users outside European Union.

Our workflow, proposed to collect digital trace data in a privacy preserving manner, consists of five steps (see Figure 1). First, data subjects are recruited as respondents using standard survey sampling techniques (Valliant et al., 2018) and the researcher determines which DDPs are relevant for the particular research question under investigation. Second, respondents request their DDPs with the various selected providers, storing these locally on their own device. Third, the stored DDPs can then be locally processed to extract and potentially transform the information from the DDP in such a way that is relevant for the particular research question under investigation. This step takes place locally at the device of the respondent by means of an extraction (and potentially also transformation) script that is tailored towards both the particular research question and DDP under investigation. Once this process is finished, the respondent can provide consent (step four) to send these derived variables to the researcher for analysis (step five). Thus, in the proposed framework no data is sent over the network until step five. To aid researchers in planning, executing, and evaluating studies that leverage the richness of DDPs, we discuss the steps involved our proposed workflow. Collecting digital trace data using the proposed workflow allows researchers to collect data that is individual, private, without requiring cooperation with the companies where the data is initially collected, by having control of the sample of respondents and with their consent. However, having control over both the obtained measurements and the sample, also means that considered decisions should be made regarding these issues by the researcher. In such cases, traditional survey research has benefited greatly from the "total survey error" framework (Groves & Lyberg, 2010); here we therefore present DDP data collection in a "total error" framework (Biemer, 2016) adapted specifically to this new mode of data collection (for a generic total error framework for "big data" see also Amaya et al. (2020)).

The aim of this paper is to introduce and discuss the idea of data donation for scientific research. As processing DDPs in such a way that high quality research can be performed is complex and challenging, a total error framework is introduced to guide researchers through this process. We first briefly discuss the right of access in the GDPR in the next section. We then present a research question that could hypothetically be addressed using Instagram DDPs collected from Dutch adolescents, such as collected in the "Adolescents, Well-being & Social Media" (AWeSome) study (Beyens et al., 2020). Note that in this paper, we only discuss this hypothetical research question for illustration purposes. Subsequently we present our total error framework for DDPs, and discuss the steps involved in answering such a research question in the context of this framework. Finally, we discuss

limitations of our approach, as well as future directions for methodological investigation. Appendix A provides a ready-to-use checklist as a guideline for researchers evaluating or conducting DDP studies.

The right of access in the GDPR

In recent years, jurisdictions around the world have enacted or are in the process of enacting new data protection legislation. Examples outside the EU include the 2017 Japanese Amended Act on the Protection of Personal Information (AAPI 2016), the 2020 Brazilian General Data Protection Law (LGDP 13.709/2018), the 2020 California Consumer Privacy Act (375/2018), the 2019 New York SHIELD act (S5575B/2019), and the proposed Personal Data Protection Bill (PDP Bill 2019) in India. Many of these laws have been designed specifically for their compatibility with the European Union's wide-reaching data protection legislation (Singh & Ruj, 2020; Suda, 2020), the General Data Protection Regulation (GDPR), which has applied across the EU and the UK since May of 2018. Together, these jurisdictions alone comprise about 2.2 billion people, over a quarter of the world's population.

The GDPR grants all natural persons ("data subjects"), whatever their nationality or residence, certain rights regarding their "personal data" with respect to "data controllers", such as tech companies, governments, mobile phone providers, etc. Although the GDPR is currently likely best known among data analysts for restricting what datacontrollers can do with personal data, the GDPR also grants data subjects the *right of access* (Article 15). This entails "the right to obtain from the controller confirmation as to whether or not personal data concerning him or her are being processed, and, where that is the case, **access to the personal data...**" (Article 15.1; emphasis added). Note that Article 15 also enables access to information regarding data recipients and sources, retention periods and data derived from your personal data. Article 15.3 further specifies the obligation for controllers to provide a copy of personal data, requiring them to do so "in a commonly used electronic form" whenever the data subject made their request by electronic means. The GDPR further grants the *right to data portability* in the closely related article 20, which states: "The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided".

In practice, most large "data controllers" currently comply with the right of access to one's personal data and the right to data portability by providing users with the option to retrieve an electronic "data download package" (DDP). For example, at the moment of writing, Google provides a "takeout" option[1], and Facebook[2], WhatsApp[3], Instagram[4], Uber[5], Apple[6], Netflix[7], and Microsoft[8] provide similar tools. Compliance with the right of data portability has sometimes been less straightforward for other data-controllers (Wong & Henderson, 2019). To our knowledge, with the exception of WeChat, none of the large global data controllers limit use of these tools to the European Union. Indeed, all other legislation mentioned above – including the California Consumer Privacy Act – grant some right of access, though often more limited than that found in the GDPR. Pursuant to GDPR article 20, data controllers cannot arbitrarily limit the data they provide in this package, or prevent their users from sharing its contents with third parties, such as social scientists. In principle, these third parties cannot be constrained by the original controller in how they process such packages, for example for scientific purposes, for as long as they comply with the GDPR themselves. The right of access is limited in that it cannot be invoked to infringe on the rights or freedoms of others, particularly on other natural persons' data protection rights, or on trade secrets; thus, the provided data should not a priori include personal data pertaining to other people (Wachter et al., 2017). For example, Facebook's data download packages do not include information on the user's "friends" (only the interactions these "friends" have with the data subject), nor does it provide details regarding Facebook's proprietary algorithms. In this sense, data included in DDPs are limited. Furthermore, in keeping with other rights granted by the GDPR, data subjects may also request deletion of their own data.

In spite of the limitations of the right of access, a wealth of information is contained in data download packages offered as its direct consequence. At the time of writing it appears likely that a large proportion of persons globally who use a smartphone or the internet will have some data in their DDPs. In the following section, we discuss how this fact can be leveraged for novel social-scientific research, as well as the pitfalls and errors that must be controlled along the way.

Using data-download packages (DDPs) for scientific research

To illustrate the considerations relevant when using DDPs for social-scientific research and thereby showing it potential, we will take the example of one

hypothetical research question that may be of interest to social scientists, and that we think could be answered using DDP collection. However, many other research questions can very well be answered by using DDP collection. For example research questions recently investigated using APIs and webscraping, such as the previously discussed network analysis from mobile phone data (Blondel et al., 2015), price indexing from online shop prices (de Haan & Hendriks, 2013), political opinion and electoral success prediction from Twitter data (Jungherr, 2015; Schoen et al., 2013), and personality profiling from Facebook "likes" (Kosinski et al. (2013) can be investigated while being more explicit regarding the generalizability of the findings. Alternatively, research questions typically investigated using surveys, such as energy consumption (Guerra-Santin & Itard, 2010), time spent (Elevelt et al., 2019) or budget research (Breedveld et al., 2002) can be executed without suffering from issues such as recall bias or bias due to social desirability.

Our example research question is inspired by the "Adolescents, Well-being & Social Media" (AWeSome) project (Beyens et al., 2020). In this study, Dutch adolescents participate in a panel study where they answer questions regardging their well-being and smartphone use, among other things (Beyens et al., 2020).

Here, we anticipate a larger follow-up study in which concepts related to well-being are further investigated. For example, adolescents' emotions are investigated *using* information obtained from their Instagram DDPs. For illustration purposes, we will work with a simple, descriptive, example research question:

> **Example RQ:** *How do emotions of Dutch adolescents differ when they are at home compared to when they are not?*

To answer this question, we must obtain (1) the consent and participation of a larger group of Dutch adolescents and their parents, and (2) measurements of the participants' emotions, as well as a measure of whether they are at home or not.

Here we will discuss the steps that would be required to obtain these data using DDPs. At each of these steps, errors can occur. In order to obtain useful answers to our research question, we must therefore take account of, and, where possible, control such errors. To enumerate the error sources associated with each step in a data collection, a highly convenient framework is the *total error* framework (Biemer, 2016; Japec et al., 2015). In a total error framework, each step of the data collection process is described, together
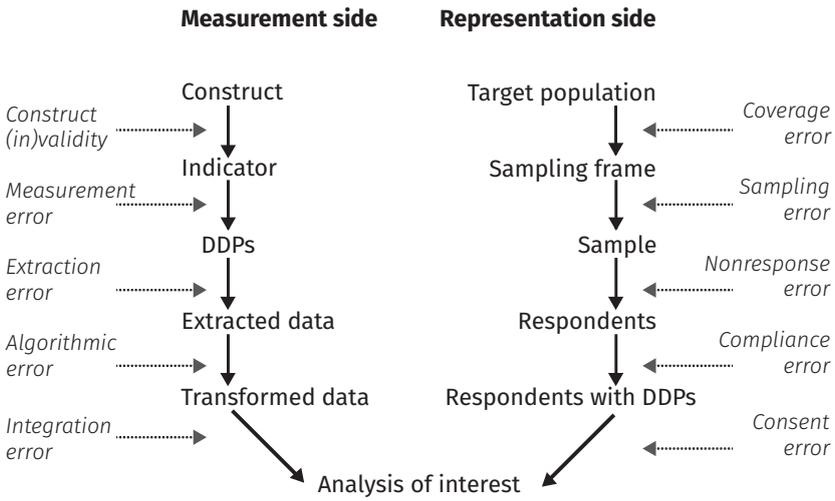
Measurement side   **Representation side**

Construct   Target population   *Construct (in)validity*   *Coverage error*

Indicator   Sampling frame   *Measurement error*   *Sampling error*

DDPs   Sample   *Extraction error*   *Nonresponse error*

Extracted data   Respondents   *Algorithmic error*   *Compliance error*

Transformed data   Respondents with DDPs   *Integration error*   *Consent error*

Analysis of interest

Figure 2. "Total error framework" for social-scientific data collection with DDPs. Each step in the data collection process is shown, together with the errors resulting from this step. Subsequent processing, modeling, and inference steps (Amaya et al., 2020) are omitted.
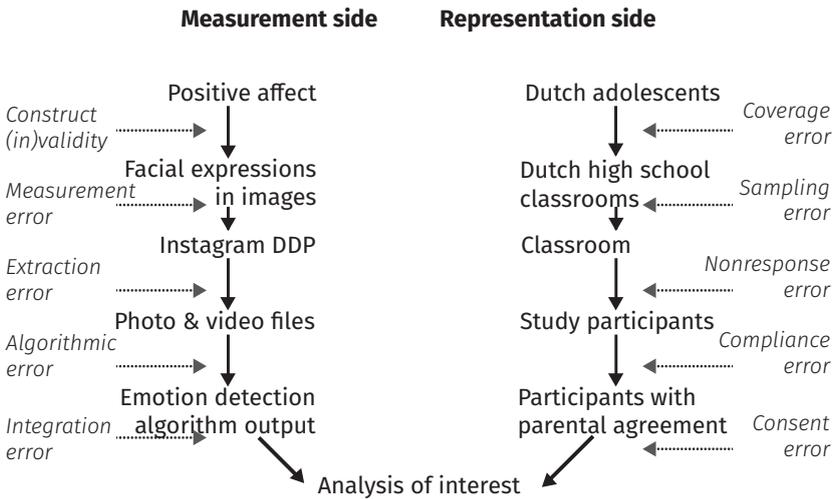
Measurement side   **Representation side**

Positive affect   Dutch adolescents   *Construct (in)validity*   *Coverage error*

Facial expressions in images   Dutch high school classrooms   *Measurement error*   *Sampling error*

Instagram DDP   Classroom   *Extraction error*   *Nonresponse error*

Photo & video files   Study participants   *Algorithmic error*   *Compliance error*

Emotion detection algorithm output   Participants with parental agreement   *Integration error*   *Consent error*

Analysis of interest

Figure 3. DDP total error framework applied to the example. Measurement side shows measurement of emotion (positive affect) from Instagram photos and videos. Note that the errors made on the measurement side of "home" vs. "other" location status are not shown here, although they will be present and affect the analysis of interest as well.

with the errors that might arise from that step. The final "total" error in the analysis or statistics produced is then a combination of the sequence of preceding errors. The concept of "total error" arose from the survey methodology literature (Groves & Lyberg, 2010), where "total survey error" (TSE) is the standard framework for designing, evaluating, and optimizing data collection (Biemer, 2010; Biemer & Lyberg, 2003). Amaya et al. (2020) extended this framework to generic "big data" studies, Sen et al. (2019) extended this framework to digital trace data and Beinhauer et al. (2020) extended the framework to sensor data.

Here, we aim to aid future researchers in performing high-quality studies using DDPs by presenting a total error framework targeted specifically at DDP collection. Figure 2 presents a generic overview of our framework. In addition, Figure 3 applies the framework from Figure 2 to our example research question above. As shown in Figures 2 and 3, and following the standard TSE formulation, data collection consists of a "measurement side" and a "representation side". The measurement side deals with the extent to which the construct of theoretical interest is adequately measured by the procedure performed in the study. In a survey, this amounts to the extent to which answers to a survey question correspond to the construct of interest (e.g. well-being). With DDP collection, several additional steps are necessary, including definition of the construct, routine registration in the DDP, and extraction and transformation of the DDP into a variable to be analyzed. On the representation side, as with a standard survey, a population must defined, a sampling frame obtained, and respondents invited to participate. With DDP collection, additional steps are involved, which will lead to further respondent attrition.

The following describes the steps of the framework in more detail. Throughout, we refer to Figure 2 and our hypothetical example illustrated (in part) by Figure 3.

Measurement side

Construct

On the measurement side of the framework, the first step is to consider how the *constructs* (concepts) of interest can potentially be measured using *indicators* (proxies) found in DDPs. Following our example, it would appear

reasonable to presume that it is possible to determine whether a person is at home using location data, and indeed Elevelt et al. (2019) showed that this can be done relatively reliably. Similarly, the existence of the field of "affective computing" suggests it may be possible to determine a person's emotions from their facial expressions in photos and videos (Dibeklioğlu et al., 2015; Kaya, Gürpınar, et al., 2017; Li & Deng, 2020).

At this stage, errors can occur due to a mismatch between the chosen concept and the chosen indicator. For example, Instagram is often described as a "storytelling" device to assert the user's desired identity in contrast to the user's *true* identity (e.g. Martínez-García, 2017). In other words, Instagram photos and videos are likely to measure how adolescents wish to be seen by others – a construct that, as attested by popular culture, centuries of literature, and many readers' personal experience, may differ from their genuine emotional state.

Construct error is especially important since it enters at the very first step of measurement and has the potential to invalidate all downstream efforts unless controlled (Saris & Gallhofer, 2007). Methods of controlling construct error might include: careful elaboration of the theory underlying the research question, expert evaluation of the proposed indicator, and "triangulation" (Munafò & Davey Smith, 2018) – for example, comparison of research results between DDP and other types of measurement, or simultaneous DDP-survey measurement followed by multitraitmultimethod modeling (Oberski et al., 2017; Revilla et al., 2017). Because construct validity is such a crucial issue, we would suggest that simultaneous measurement using a combination of sources, including DDPs, is advisable; this idea is in line with similar advice given by Japec et al. (2015) and Konitzer et al. (2020). Our proposed workflow foresees in this need explicitly, by embedding the DDP collection step within a larger, more traditional, survey data collection effort.

Indicator

Once the researcher has identified valid *indicators* for the construct(s) of interest, the next step is to determine from which "data controller(s)" the *DDP(s)* is/are most useful to answer the research question. For our example research question, we are interested in whether adolescents feel different emotions when they are at home compared to when they are not. As individuals typically switch locations multiple times a day, a DDP that registers location only once a day would not be sufficient to make the distinction we are interested in. The location history listed in the Instagram DDP only logs a location when it is selected by the user while sharing media on the "timeline" or in the "stories" (Manikonda et

al., 2014) and would therefore not be sufficiently dense to appropriately distinguish between being home or not for every location the respondent visits throughout a day. Alternatively, Google Location History passively logs visited locations by combining internal phone GPS with connected WiFi devices and cell towers (Ruktanonchai et al., 2018) and is therefore much more appropriate for the research question under evaluation. In terms of measuring emotions via social platforms (Kramer et al., 2014), adolescents frequently use Instagram (Valkenburg et al., 2011), where emotions can be shared through both images and text (Bouko, 2020), which can be shared both publicly and privately.

At this stage, errors can occur when the measurements collected in the DDP diverge for some reason from what they intend to measure. For example, when satellites are temporarily out of order (Andrei et al., 2020), the measurements logged in Google Location History might diverge more from the user's true location.

Measurement error is particularly relevant because all measurements can be prone to error (Brakenhoff et al., 2018) and it can distort all relationships under evaluation (Biemer & Lyberg, 2003). A way to control for measurement error is by collecting multiple independent measurements of the construct of interest and investigate the variance of these measurements (Carroll et al., 2006) or their correlations (Bland & Altman, 1996). Furthermore, these independent measurements can be used to estimate the unobserved "true" variable (Biemer, 2011). In practice, this can be accounted for similarly as construct error, namely to supplement DDPs with survey measurements. In addition, measurement and construct error can be simultaneously estimated and accounted for using the previously discussed multitrait-multimethod modeling (Oberski et al., 2017). To investigate positive affect using images in Instagram DDPs, a way to account for measurement error here can be to measure facial expression from other sources, such as self reports, sharing of selfies through ESM or using another DDP. The information extracted from these different sources can then be used as indicators of the construct of interest by means of a latent variable model.

DDPs

Once a specific set of *DDPs* has been chosen to answer the research question of interest, the next step is to think more specifically which files of these DDPs are essential and how these relevant files are going to be *extracted* from the DDPs. For our example research question, we are interested in determining the emotional expressions of faces on images. The extraction step here would be to identify all images in the Instagram DDP.

Extraction error occurs when errors are made by the extraction algorithm, the image detection algorithm for example. A simple example of such an error is when the image detection algorithm only selects files with a .jpg extension, while the DDP of interest also contains images with a .png extension. Another example of this type of error is that images might be stored under different directories, and some are systematically missed. Moreover, controllers might also differentiate in their responses over time/location as well as file formats/structures(Ausloos et al., 2019).

To minimize the possibility of extraction error, researchers should extensively investigate the content of the DDPs of interest, and how the structures and data type might differ over different users. Although this step might sound straightforward, it can be quite a challenging and intensive procedure to thoroughly fathom a DDP structure. The script should not only be able to handle the particular structure, but should also be flexible enough to handle unexpected minor variations in the DDP structure. The approach that Boeschoten et al. (2021) used to develop a Python script that handles Instagram DDPs was that they started with the development of their script by inspecting the structure of the DDPs of a small number of researchers and acquintances. Next, with a group of researchers they all generated a new Instagram account which they intensively used in such a way that all features offered by Instagram were used, ensuring that the resulting DDPs contained de structures corresponding to all these features, and possible variations within. In addition, the corresponding script was developed in such a way that it could handle different and changing file formats. Although using this approach for data controllers such as social media sounds reasonable, it already becomes much more challenging when considering for example online shops or paid services, as it will then involve costs. However, for again another type of data controllers, such as for example banks, these efforts might not even be needed as they are required by law to provide the data in a highly structured format (which is known as 'SEPA' in the case of digital transactions ("The schemes rely on global open standards," 2021).

Extracted data
Once the relevant files have been *extracted*, an algorithm can be applied *transforming* the extracted files into data that can be used to answer the research question. In some cases, this step is very simple as data can be extracted from the files directly without further processing. Following the example research question, a face detection algorithm (Hjelmaas & Low, 2001; Hsu et al., 2002) followed by an emotional expression detection algorithm could be applied to the images in the Instagram DDP, for

example using pretrained models or by models further developed my means of transfer learning, such as by Kaya, Gürpinar, et al. (2017).

Algorithmic error occurs when errors are made while generating transformed data from the extracted DDP files. When classifying emotions from faces, algorithmic error can be due to a face not being detected (as can be seen in Figure 4, a face incorrectly being detected, an incorrect emotional classification, or because the algorithmic uncertainty is lost once a classification is made. In other words, algorithmic error is the typical classification or prediction error in predicting social variables using found data, which is the focus of a large body of literature (Blondel et al., 2015; Elevelt et al., 2019; Jungherr, 2015; Kosinski et al., 2013; Settanni et al., 2018). In line with our example research question, research has also illustrated that algorithmic error can influence outcomes of computer vision algorithms (Buolamwini & Gebru, 2018). In the current work, we emphasize that, while this type of error is certainly important, it constitutes only one type of error within the total error framework. In other words, the "ground truth" employed by supervised modeling exercises is, within our framework, an error-prone and potentially partially invalid proxy of the concept of interest.

Algorithmic error in the current framework is essentially prediction error on an (error-prone) measure of some socially relevant variable. As such, it is among the most studied errors within the framework at the time of writing. As emphasized in every basic textbook on machine learning, a proper evaluation of the likely amount of error is key, and can be accomplished by separating training and test observations, whether this is using data splits or resampling techniques (Bengio et al., 2017; Bishop, 2006; Murphy, 2012). When applying pretrained models as an extraction method, the researcher should ideally evaluate whether the error incurred within the DDP dataset at hand is indeed similar to that within the test set of the original model. For example, the type of photographs taken by teenagers might be different from standard benchmark datasets on which image recognition models were trained. Obtaining an accurate estimate of the algorithmic error rate also makes it possible to handle downstream decisions more adequately, by using standard measurement error models. For example, when we know that a classification model has a 90% sensitivity and 75% specificity, a simple table of predicted counts from this model can be corrected by multiplying it by the inverse of a matrix with these rates on the diagonal (Beauxis-Aussalet & Hardman, 2017; Boeschoten et al., 2018). However, the difficulty of obtaining appropriate estimates of algorithmic errors should not be underestimated, particularly if pre-trained models are used for which training occurred using a different data-set.
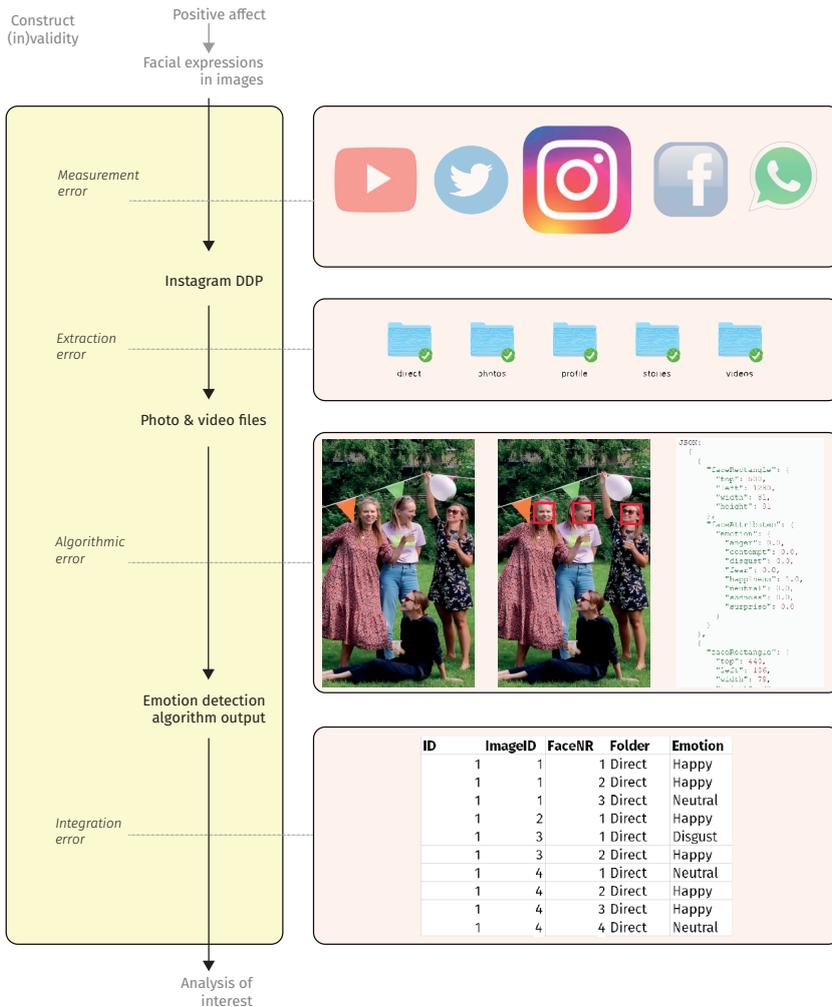
Figure 4. Visual illustration of the measurement side of the framework when using emotional expressions on images found in an Instagram DDP to measure positive affect.

Transformed data

After the *transformed data* files from all respondents are received and safely stored by the researcher, an integrated dataset can be generated containing data from all respondents, and linking the measurements received from possibly multiple DDPs to, for example, survey outcomes. As typically measurements at different time-points are collected through DDPs, attention should be paid to appropriately integrating the multiple datasets by linking on both person and time level (Harron et al., 2015; Zhang, 2012). For

our example research question, we should link the collected emotions to collected locations on time-level per person.

While linking the multiple sources on subject level and linking the subjects, **integration error** can occur (Kim & Tam, 2020), for example when time-stamps are not appropriately matched or when information collected from multiple sources is not appropriately linked on subject level (Doidge & Harron, 2019). Such errors can be prevented to an extent by creating software tests and other checks (Myers et al., 2004) at every stage of the linkage process that create reports which can be compared with sensible expectations. For example, the time period should not suddenly extend into unseen years, outliers should be detected, etc. In addition, the procedures used should be computationally reproducible, so that any errors can be detected in the future and easily corrected (Stodden et al., 2014; Stodden & Miguez, 2014).

See Figure 4 for a visual representation of how errors can affect outcomes on the measurement side of the framework. In addition, see the first part of Appendix A for guidance on how severe bias due to measurement errors can be prevented.

Representation side

Target population
On the representation side, researchers have in mind to what population their results should be generalized, a *target population*. For the example research question, the target population is Dutch adolescents. Furthermore, researchers investigate how a sample or participants can be selected from that target population, this is the *sampling frame*. If your target population is Dutch adolescents, it can be infeasible to randomly select a set of respondents out of that complete population directly. A practical approach can be to first select a sample of high schools and then select a number of adolescents here. Such a sampling scheme is known as clustered sampling (Bethlehem et al., 2011; Lohr, 2008).

The discrepancy between the target population and the sampling frame is denoted as **(under)coverage error**, as certain subgroups are not covered by the sampling frame. Coverage error can result in the problem that the obtained results cannot be generalized to the population of interest. For example, when Dutch high schools are used for the sampling frame, the subgroup of adolescents not going to high school have no probability of being included in the sample and obtained results can therefore not be

generalized to Dutch adolescents, but only to Dutch adolescents going to high school. A solution can be to use multiple sampling frames (Lohr, 2009).

Sampling frame

When the *sampling frame* has been determined, the *sample* can be selected using traditional sampling theory (Cochran, 2007), for example a simple random sample can be selected by randomly selecting a number of adolescents from the high school registers and invite them to participate in the research. Alternatively, using strata or clusters can be more convenient here, for example to first select a number of high schools and approach a sample of adolescents via these schools.

Failing to select a representative sample results in **sampling error**, failure to generalize results to the target population. Many large studies use model-based approaches (Chambers & Skinner, 2003), combining several stages of stratification and clustering to minimize sampling error (De Leeuw et al., 2008). Alternatively, adaptive designs can be used to minimize sampling error (Bethlehem et al., 2011) and to for example increase the sampling probability for certain subgroups if their response rate is relatively low. Stratification has also been listed by Japec et al. (2015) as an important contribution to the goal of generalizability in big data research.

Sample

Once the *sample* has been determined, its members can be invited to participate in the research. As with any type of research, part of the sampled subjects will not or only partly respond. This can be due to multiple reasons. First, the subject is not willing to participate at all. Second, the subject is willing to participate in the overall project, but is not willing to provide her DDP. Third, the subject is willing to participate, but does not use the platform from which the DDP is requested.

Regardless of the reason for not participating in the research, this will lead to **nonresponse error** an can lead to bias in results (Groves & Peytcheva, 2008). To minimize bias caused by respondents not willing to participate or only willing to partly participate, it is recommended to accompany data-download research with questionnaires. This provides the researcher with substantive information regarding the non-responders in terms of data-download packages. When viewed as a missing data problem, this means that once more information is known about the non-respondents, the likelihood increases that the Missingness is At Random (MAR), as variables are observed through which the missingness can be explained. This is in contrast to a situation when nothing is known about the nonrespondents, so that the

missingness cannot be explained (Missing Not At Random, MNAR) (Schafer & Graham, 2002). When the missingness can be explained, it can be accounted for by method such as multiple imputation or weighting (Boeschoten et al., 2017). To minimize the number of respondents that are willing to participate but to not use the platform under investigation, the researcher should also focus on how often the target population makes use of the platform under investigation when determining which platform to use for research. When considering the example research question, existing research showed that YouTube, WhatsApp, Instagram and Snapchat were used most frequently by adolescents in 2019 (van Eldik et al., 2019). Furthermore, Android had a market share of 86.1% in 2017 (Ahvanooey et al., 2020).

Respondents

If a respondent decides to participate in the research, she still needs to work through a process of multiple stages. The packages should be requested and downloaded. A piece of software should be installed and the packages should be opened and processed with this software, as can be seen in Figure 1. Next, the output is generated by the software and the respondent determines whether she is willing to share this output with the researcher and, if so, actually approve the sharing.

These steps are not straightforward. Therefore, clear guidelines, reminders and assistance are required to guide the respondents through this process (Shirima et al., 2007). Some attrition is likely to occur due to the fact that respondents are not willing or able to invest the time and effort in this procedure, resulting in **compliance error**. For example, when a respondent requests her Instagram DDP, it typically takes several hours to days for Instagram to prepare this DDP, so the respondent needs to reserve multiple moments throughout several days to successfully participate in this research, and the researchers should probably build in several reminders throughout this process to nudge the respondent into successfully completing the process. Furthermore, by processing and visualizing locally, the respondent has control over the data and is truly informed.

Respondents with DDPs

Once the respondent complied with all the steps required to complete the process, the transformed data is collected in a file. For our example research question, a respondent will for example review a csv file containing timestamps, classified emotions and supplementary information describing whether it was text or an image that was classified (as can be seen at the bottom of Figure 4.

This file should be reviewed by the respondent in order to give informed consent regarding sharing this information with the researchers. If the respondent decides to not or only partly share this file, this results in **consent error.** Consent error may be substantial, and could be related to topics of interest measured within the DDPs. Without any further information about the respondent, for example for a survey, this would lead to missingness "not at random" (MNAR; Schafer & Graham, 2002), which is difficult to account for. With information from surveys or other sources, it may be reasonable to assume the missingness is "at random" (MAR), especially when survey variables are strongly related to the study outcomes.

"Local signal processing" may alleviate the consent error considerably. First, local processing will allow researchers to avoid requesting sensitive information, perhaps making respondents more willing to share (Singer, 1993). For example, respondents could be more likely to give consent to share the datum of "looking unhappy" in a photograph than sharing all their private images. Second, the respondent can see that the only information that is requested is directly related to their interaction with the researcher: a scientific study. Most adolescents will intuit that, to study well-being, the researcher does not need to know their study habits, for instance. In other words, local signal processing is designed to comply with key data protection principles such as 'data minimization' and 'data protection by design' as well as more generally preserve the interaction's "contextual integrity" (Nissenbaum, 2004). Some studies have suggested that preserving contextual integrity can help improve consent (Hutton & Henderson, 2015).

When showing the respondent the extracted information for consent, the researcher should put effort in making the data easy and intuitive to understand. It can for example help to provide an explanation of what the extracted data exactly contain, or to make it visually more attractive in a figure.

Once the integrated data-set is finalized, it can be used to perform the final analyses to answer the research question of interest. For example, the researcher can investigate what type of emotions are more often detected while being at home and while being at other locations, and it can be investigated how these differences in emotional outings differ within and between persons.

See the second part of Appendix A for guidance on how severe bias due to representation errors can be prevented.

Discussion

Data-download packages (DDPs) allow us to study known phenomena in a novel manner, or even to study new social phenomena. Using DDPs for scientific research is attractive for multiple reasons. First, the existence of DDPs, and the right of the data subject to pass on information to social scientists, is guaranteed by EU law. Second, participants can easily investigate the data they share to give informed consent. Third, by starting off with a traditional random sample, the approach suggested in this paper allows researchers to generalize to populations of interest more easily than could be achieved with "found samples". This approach also allows for longitudinal data collection in parallel with the DDPs. More generally, fourth, DDPs do not only provide a very diverse set of available digital traces, but they can also easily be combined with other data, such as other DDPs, surveys, register data, and so forth. Finally, the approach suggested in this paper allows for experimental designs using digital trace outcomes, but under the same scrutiny as regular social-scientific experiments and with true informed consent that respects the contextual integrity of the researchparticipant interaction.

Of course, use of DDPs is also challenging. We have focused on summarizing some of the challenges to inference within our error framework, and hope this framework can serve as a guide to preventing errors where possible, and mitigating their effects otherwise. At the same time, the suggested approach also has several drawbacks that are unrelated to inference per se (Ausloos & Veale, 2020).

First, researchers should have good faith in not only respondents, but also in data controllers, as both have the opportunity to omit data during the process. DDPs may not be comprehensive and often do not include all information covered by the right of access in the GDPR (and similar legislation outside of the EU), and respondents can choose to remove parts of the DDP they are not willing to share with the researcher. As data not shared with the researcher can differ from shared data, this can bias results. A second challenge is that the world of DDPs changes rapidly. The structure and content changes continuously and individuals can be triggered to delete their own packages making them useless as research subjects. A third challenge is that, to safeguard participants' privacy and for scientists to comply with data protection requirements themselves, most research infrastructure should be set up in advance. For example, it should be clear which parts of which data-download packages are selected and an algorithm should be prepared to make transformations to a pre-defined format. A fourth disadvantage is

that available and free pre-trained algorithms are not always available for the specific research purposes, requiring the researcher to collect raw data and train an algorithm. Fifth, digital skills of participants are a major challenge. To address them an easy to use front end of the data collection tool is key. A sixth challenge is that data-download packages are not consistently formatted over different data controllers. For example, there are already many ways to provide timestamps (Dyreson & Snodgrass, 1993) so software should be adjusted to appropriately handle such differences. A seventh challenge is that a DDP itself is not formatted as a typical data set with respondents as rows and variables as columns. Instead, it typically comes as a zip file containing json files, images and videos, and processing should take place in order for it to be used for statistical analyses. A last challenge is that conducting research of this type should be carried out by a multidisciplinary team of social scientists, data scientists, computer scientists and data management experts.

Researchers can minimize the influence of issues such as the rapidly changing environment of DDPs and the inconsistency in DDPs by focussing their processes on structural characteristics such as for example usernames and timestamps. Issues such as setting up the infrastructure in advance and training algorithms without access to the complete data have been overcome before (Lovestone & Consortium, 2020), however ensuring that the usability of such infrastructures meets the level of digital skills of the participant remains an important challenge here. For challenges regarding data protection, informed consent, reproducibility and replicability, extensive research has been performed and guidelines have been developed on which we reflect in the following subsections.

Data protection and informed consent

Before a DDP of a respondent is shared, it is unknown what kind of information the package exactly contains. Social researchers will only be interested in the specific parts of the DDP that help to answer their research question, but a DDP possibly contains sensitive personal information. By using distributed local computation at the respondent's device to extract only the relevant information, it can be prevented that a researcher stores sensitive information. For example, an Instagram DDP can contain sensitive images. The researcher is not interested in the sensitive content, but in the emotional expressions of the faces on these images. Therefore, an emotional detection algorithm could be run locally and only the classifications of the emotional expressions are shared with the researcher.

During this privacy preserving transformation step, three aspects should be carefully considered. First, respondents store their DDPs locally on a device. After participating, respondents should be informed of this and should have the option to either preserve the packages under their own responsibility, or to permanently delete the packages from the device in use. Second, to maximize informed consent, respondents should be shown an example illustrating what information is extracted from the data-download package. In the case of transforming faces on pictures into classifications of emotional expressions, the example should show a picture as input and the classifications of emotional expressions per detected face at output, as can be seen in Figure 4. Such an example makes clear what information from the data-download package is shared with the researcher exactly. In addition, respondents should have access to output of the transformations applied to their own DDP to explicitly approve or reject sharing the transformations with the researcher. Existing research on successful informed consent can be consulted, see for example (Kreuter et al., 2016).

To ensure that sensitive information is not shared with the researcher and to ensure that the procedure of obtaining the transformed data occurs in a privacy preserving and ethical way, it is important that researchers consult ethical review boards of their universities in this process and obtain ethical approval for the research. Furthermore, researchers should consult data managers to develop a solid plan to receive the transformed data in a safe environment from the respondents and to generate an integrated database built with an architecture that can be accessed by the researchers, such as SURFsara in the Netherlands (Scheerman et al., 2020).

Reproducibility and replicability

Although reproducibility and replicability are essential for scientific research (Patil et al., 2016; Stodden & Miguez, 2014), these criteria are challenging to meet when using DDPs (Gayo-Avello, 2012). The field of digital trace data in general is a rapidly changing environment (Stier et al., 2019), and this holds for DDPs as well. When using local computation, reproducibility may only be feasible on the level of the transformed data received by the researchers, not on the raw DDPs, as they were never in the possession of the researcher in the first place.

To support replicability, tools and analysis code should depend on structures specific for particular data controllers as little as possible, and should be easily updatable and extendable as structures of DDPs from

specific data controllers will inevitably change. To help achieve this goal, the highest standards of software engineering for architecture design, testing, documentation, version control and support should be applied and software engineers should be involved during all stages of the process (Myers et al., 2004). In addition, FAIR principles (Wilkinson et al., 2016) should be used for data archiving, documentation and long-term storage. As these go beyond the expertise of most social scientists, Research Data Management Offices should be involved or at least consulted, see for example ("Utrecht University Research Data Management Support," n.d.). Frameworks such as differential privacy (Dwork, 2008) are relevant to guarantee reuse.

Conclusion

If researchers interested in using DDPs for scientific research follow the proposed workflow, improvements can be made regarding generalizability of findings. This holds for the example research question discussed, but also for example for the research questions discussed in the introduction such as the network analysis from mobile phone data (Blondel et al., 2015), price indexing from online shops (de Haan & Hendriks, 2013), political opinion and electoral success prediction from Twitter data (Jungherr, 2015; Schoen et al., 2013), and personality profiling from Facebook "likes" (Kosinski et al., 2013). Furthermore, research questions typically investigated using surveys can be executed without suffering from issues such as recall bias or bias due to social desirability, such as the examples discussed in the introduction regarding such as energy consumption (Guerra-Santin & Itard, 2010) time spent (Elevelt et al., 2019) or budget research (Breedveld et al., 2002).

While collecting data with the proposed framework – i.e., relying on DDPs and the steps required to meaningfully make sense of their data – may be seen as effortful, this method brings a set of important opportunities for academic research, in two different fronts. First, using DDPs allows for the collection of individual-level data, with informed consent, and at a level of granularity simply not available in existing APIs from social media platforms, or via scraping. If the research interest is on the analysis of publicly available content at aggregate levels, then relying on APIs may be sufficient for platforms that may be open to this type of research (e.g., Twitter), and scraping may be a possibility – although open to legal contests– for those that do not. However, if the research interest is at the individual level (i.e., activities of a set of individuals), then using DDPs may be a promising

avenue to pursue. Working with informed consent and collecting multiple data points per individual, DDPs provide not only public but also private content created by individuals and, importantly, go beyond simply what one posts or who one follows on social media. DDPs encompass a diverse set of digital trace data – including user activity and/or profiling done by a platform about a user – often at a level of granularity and detail that is simply not available via APIs. Second, using DDPs complements – rather than substitutes – the information that can be gathered via self-reports. This brings two sets of advantages. On the one hand, given the ubiquitous, fragmented and always-on nature of the current media environment, respondents have difficulty in providing accurate estimates of their media exposure or usage of digital platforms (Araujo et al., 2017; Parry et al., 2021). Using data derived from DDPs instead of self-reports in these cases may help researchers ease the cognitive burden experienced by respondents during surveys and provide more accurate estimates of digital media use for researchers. On the other hand, certain respondent characteristics can also be inferred from DDPs that often suffer from social desirability issues when asked directly in a survey setting, such as political interest, religiosity or environmental behavior.

In addition to our strong recommendation to combine various data sources, the possibility always remains to collect digital trace data in a manner different from our proposed workflow. For example the previously discussed tracker apps and plugins remain useful and even more suitable for certain purposes, for example if the research focusses more on what participants see on social media instead of what they do themselves (Reeves et al., 2021). However, we think the response burden in terms of technological knowledge is not that different for these alternative approaches. At last, researchers can also consider to collect complete DDPs instead of only extracting features. This can for example be relevant if the data will be used for training purposes or for more indepth analyses (e.g., content analyses that may require access to a wide variety of textual data). However, even in such situations it remains recommended to preserve the privacy of participants and to for example run a de-identification algorithm (Boeschoten et al., 2021) locally before receiving the DDPs.

To summarize, it is clear that our proposal is no silver bullet for solving all problems associated with modern social science. In spite of these challenges, however, we believe that leveraging the advantages of DDP collection can become an important tool in the social scientist's arsenal.

Appendix A Checklist for social scientific research using Data Download Packages

Measurement side
(determine per construct separately)

Construct.
– The construct of interest is clearly defined.
– The construct of interest matches the scope of the research.

Indicator(s).
– All aspects of the construct can be sufficiently represented through observable indicators (proxies).
– The indicators can be measured by data controllers.

DDPs.
– Data controllers are selected in which the indicators of interest are measured.
– The denseness of the measured indicators matches the research purpose.
– The credibility of the data controller is positively evaluated.
– The number of different data controllers is minimized to reduce response burden.

Extracted data.
– Presence of the indicator is evaluated for all file formats present in the DDP.
– Relevant files are extracted using validated scripts with known accuracy rates.

Transformed data.
– A transformation method is selected that extracts the outcome values for each indicator.
– The transformation method is trained on a sample similar to the data collected by means of DDPs.
– The transformation method has a known accuracy rate estimated on a comparable data-set.
– The transformation method does not systematically include, exclude or misclassifies specific (identifiable) cases.
– The outcome values sufficiently represent all indicators identified.

Analysis of interest.
– The shared data is linked on person level, such that different sets of transformed data are represented by different columns in one data-set.
– Individual respondents can be clearly identified, for example by means of an anonymized identification number.
– The variables are clearly identified for each respondent.

Representation side

Target population.
– A target population is identified that matches the research purpose.
– All identifiable subgroups can in theory be included in the study.

Sampling frame.
– All identifiable subgroups of the target population are present in the sampling frame.
– Evaluate whether the available sampling frame matches the research purpose.

Sample.
– All subgroups in the sampling frame have a probability to be included in the sample.
– All subgroups in the sampling frame have an equal or known probability to be included in the sample.

Respondents.
– The communication towards the sample is clear and simple.
– Communication is possible in the respondent's language.
– The procedure is explained in a step-by-step manner for informed consent at the start of the procedure.

Respondent's DDPs.
– The software's usability has been validated on an independent validation sample.
– The software is available for different types of devices and different versions of operating systems.
– 24 hour assistance is available during the data collection period.

Analysis of interest.
–   The respondents can see the final data-set containing the transformed data before it is shared with the researcher for informed consent.

Appendix B Definitions

–   Personal data: Information relating to an identified or identifiable natural person (van der Sloot, 2020)
–   Data subject: The person that the personal data refer to.
–   Data processing entity / Data controller: The person or organization responsible for processing personal data. In this paper we refer to the online platforms providing data download packages as data controllers. However, note that as a researcher collecting DDPs, you are a data controller as well (van der Sloot, 2020).
–   Data controller: The person or organization responsible for processing personal data. The controller decides which data will be processed, how and why (van der Sloot, 2020).
–   Data download package (DDP): Because of the right of data access, data subjects are always allowed to retrieve their personal data from data controllers. Here, data controllers are obliged to comply with such a request and because of the right of data portability, provide the requested data in a machine readable format. To comply with these rules, social media platforms typically provide data subjects with a .zip file containing the personal data requested (van der Sloot, 2020).
–   Consent: When data subject provide researchers their DDPs, consent should be provided. This means that the data subject confirms that the data provided given freely; that the data subject is informed regarding what data are shared exactly and how the data will be processed by the researcher. Consent can be provided via a written, electronic or oral statement (van der Sloot, 2020).
–   Target population: The population to be investigated, and about which conclusions are to be drawn (Bethlehem et al., 2011).
–   Sampling frame: A list, map, or other specification of units in the target population from which a sample of data subjects may be selected (De Leeuw et al., 2008).
–   Sample: The set of data subjects within the sampling frame selected for participation in the research in practice.
–   Respondents: Data subjects within the sample who complied with participation in the research.

–   Responses: Data collected from the data subjects who complied with participation in the research.
–   Construct: A conceptual variable that is known to exist but cannot be directly observed (Privitera, 2018).
–   Indicator: Variables (constructed by means of measurement instruments) that aim to measure either the construct of interest or are closely related to the construct of interest.
–   Transformation method: Algorithm that is used to transform the data obtained from the DDPs into features and classifications that can be used for further research.
–   Transformed data: The features or classification extracted using transformation method which can be used for further research.
–   Data integration: The theory and techniques used for data linkage and micro integration. Here, data linkage techniques vary from record linkage to statistical matching. Micro integration techniques vary from harmonization of measures in concept to actual adjustments of data (Zhang, 2012).

Notes

1.   https://support.google.com/accounts/answer/3024190
2.   https://www.facebook.com/help/1701730696756992
3.   https://faq.whatsapp.com/general/account-and-profile/how-to-request-your-account-information/
4.   https://help.instagram.com/181231772500920?helpref
5.   https://help.uber.com/riders/article/request-a-copy-of-your-uber-data?nodeId=2c86900d-8408-4bac-b92a-956d793acd11
6.   https://privacy.apple.com/
7.   https://www.netflix.com/account/getmyinfo
8.   https://support.microsoft.com/en-us/help/4468251/microsoft-account-view-your-data-on-the-privacy-dashboard

References

Ahvanooey, M. T., Li, Q., Rabbani, M., & Rajput, A. R. (2020). A survey on smartphones security: Software vulnerabilities, malware, and attacks. *arXiv preprint arXiv:2001.09406*.

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119. https://doi.org/10.1093/jssam/smz056

Andrei, C.-O., Johansson, J., Koivula, H., & Poutanen, M. (2020). Signal performance analysis of the latest quartet of galileo satellites during the first operational year. *2020 International Conference on Localization and GNSS (ICL-GNSS)*, 1–6.

Andrews, S., Ellis, D. A., Shaw, H., & Piwek, L. (2015). Beyond self-report: Tools to compare estimated and real-world smartphone use. *PloS one*, *10*(10), e0139004.

Araujo, T., Wonneberger, A., Neijens, P., & de Vreese, C. (2017). How much time do you spend online? understanding and improving the accuracy of selfreported measures of internet use. *Communication Methods and Measures*, *11*(3), 173–190.

Ausloos, J., Veale, M., Mahieu, R., et al. (2019). Getting data subject rights right: A submission to the european data protection board from international data rights academics, to inform regulatory guidance. *Journal of Intellectual Property, Information Technology and Electronic Commerce Law*, *10*.

Ausloos, J. (2019). *GDPR Transparency as a Research Method* (SSRN Scholarly Paper No. ID 3465680). Social Science Research Network. Rochester, NY. https://doi.org/10.2139/ssrn.3465680

Ausloos, J., & Veale, M. (2020). Researching with data rights. *Technology and Regulation*, 136–157.

Beauxis-Aussalet, E., & Hardman, L. (2017). Extended methods to handle classification biases. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 765–774.

Beinhauer, L., Snijkers, G., & Bakker, J. (2020). Towards a total error framework for sensor and survey data. *BigSurv20*.

Bengio, Y., Goodfellow, I., & Courville, A. (2017). *Deep learning*. MIT press.

Bethlehem, J., Cobben, F., & Schouten, B. (2011). *Handbook of nonresponse in household surveys* (Vol. 568). John Wiley & Sons.

Beyens, I., Pouwels, J. L., van Driel, I. I., Keijsers, L., & Valkenburg, P. M. (2020). The effect of social media on well-being differs from adolescent to adolescent. *Scientific Reports*, *10*(1), 10763. https://doi.org/10.1038/s41598-020-67727-7

Biemer, P. P. (2010). Total Survey Error: Design, Implementation, and Evaluation. *Public Opinion Quarterly*, *74*(5), 817–848. https://doi.org/10.1093/poq/nfq058

Biemer, P. P. (2011). *Latent class analysis of survey error* (Vol. 571). John Wiley & Sons.

Biemer, P. P. (2016). Errors and inference. In I. Foster, R. Ghani, R. S. Jarmin, F. Kreuter, & J. Lane (Eds.), *Big data and social science: A practical guide to methods and tools* (pp. 266–297). CRC press.

Biemer, P. P., & Lyberg, L. (2003). *Introduction to survey quality*. Wiley.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Bland, J. M., & Altman, D. G. (1996). Measurement error and correlation coefficients. *BMJ: British Medical Journal*, *313*(7048), 41.

Blondel, V. D., Decuyper, A., & Krings, G. (2015). A survey of results on mobile phone datasets analysis [Number: 1 Publisher: SpringerOpen]. *EPJ Data Science*, *4*(1), 1–55. https://doi.org/10.1140/epjds/s13688-015-0046-0

Boeschoten, L., Oberski, D. L., De Waal, T., & Vermunt, J. K. (2018). Updating latent class imputations with external auxiliary variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 750–761.

Boeschoten, L., Vink, G., & Hox, J. J. (2017). How to obtain valid inference under unit nonresponse? *Journal of Official Statistics*, *33*(4), 963–978.

Boeschoten, L., Voorvaart, R., Kaandorp, C., Goorbergh, R. v. d., & De Vos, M. (2021). Automatic deidentification of data download packages. *arXiv preprint arXiv:2105.02175*.

Bolsen, T., Druckman, J. N., & Cook, F. L. (2014). The Influence of Partisan Motivated Reasoning on Public Opinion. *Political Behavior*, *36*(2), 235–262. https://doi.org/10.1007/s11109-013-9238-0

Bouko, C. (2020). Emotions through texts and images: A multimodal analysis of reactions to the brexit vote on flickr. *Pragmatics*, *30*(2), 222–246.

Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G., Groenwold, R. H., & van Smeden, M. (2018). Measurement error is often neglected in medical literature: A systematic review. *Journal of clinical epidemiology*, *98*, 89–97.

Breedveld, K., Van Den Broek, A., & Huysmans, F. (2002). Background to the methods used in the time budget survey (tbo). *Social and Cultural Planning Office of the Netherlands. Disponible en internet via: http://www. scp. nl/onderzoek/tbo/english/achtergronden/history.pdf.*

Bruns, A. (2019). After the 'apicalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, 77–91.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. CRC press.

Cbs statline. (2020). https://opendata.cbs.nl/#/CBS/nl/dataset/84888NED/table?ts=1619697629253 Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data*. John Wiley & Sons.

Cochran, W. G. (2007). *Sampling techniques*. John Wiley & Sons.

Coleman, J. S. (1990). *Foundations of social theory*. Belknap Press of Harvard Univ. Press.

De Leeuw, E. D., Hox, J. J., & Dillman, D. A. (2008). *International handbook of survey methodology*. Taylor & Francis Group/Lawrence Erlbaum Associates.

de Haan, J., & Hendriks, R. (2013). Online data, fixed effects and the construction of high-frequency price indexes. *Economic Measurement Group Workshop*, 28–29.

Dibeklioğlu, H., Salah, A. A., & Gevers, T. (2015). Recognition of Genuine Smiles. *IEEE Transactions on Multimedia*, *17*(3), 279–294. https://doi.org/10.1109/TMM.2015.2394777

Doidge, J. C., & Harron, K. L. (2019). Reflections on modern methods: Linkage error bias. *International Journal of Epidemiology*, *48*(6), 2050–2060.

D'Orazio, V., Honaker, J., & King, G. (2015). Differential Privacy for Social Science Inference. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2676160

Dwork, C. (2008). Differential privacy: A survey of results. *International conference on theory and applications of models of computation*, 1–19.

Dyreson, C. E., & Snodgrass, R. T. (1993). Timestamp semantics and representation. *Information Systems*, *18*(3), 143–166.

Elevelt, A., Bernasco, W., Lugtig, P., Ruiter, S., & Toepoel, V. (2019). Where You at? Using GPS Locations in an Electronic Time Use Diary Study to Derive Functional Locations. *Social Science Computer Review*. https://doi.org/10.1177/0894439319877872

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

Gayo-Avello, D. (2012). No, you cannot predict elections with Twitter. *IEEE Internet Computing*, *16*(6), 91– 94.

Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. *Public Opinion Quarterly*, *74*(5), 849–879. https://doi.org/10.1093/poq/nfq065

Groves, R. M., & Peytcheva, E. (2008). The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public opinion quarterly*, *72*(2), 167–189.

Guerra-Santin, O., & Itard, L. (2010). Occupants' behaviour: Determinants and effects on residential heating consumption. *Building Research & Information*, *38*(3), 318–338.

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

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society*, *22*(11), 1567–1581.

Harron, K., Goldstein, H., & Dibben, C. (2015). *Methodological developments in data linkage*. John Wiley & Sons.

Hjelmaas, E., & Low, B. K. (2001). Face detection: A survey. *Computer vision and image understanding*, *83*(3), 236–274.

Hsu, R.-L., Abdel-Mottaleb, M., & Jain, A. K. (2002). Face detection in color images. *IEEE transactions on pattern analysis and machine intelligence*, *24*(5), 696–706.

Hutton, L., & Henderson, T. (2015). "i didn't sign up for this!": Informed consent in social network research. *ICWSM*.

Japec, L., Kreuter, F., Berg, M., Biemer, P. P., Decker, P., Lampe, C., Lane, J., O'Neil, C., & Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, *79*(4), 839–880.

Jungherr, A. (2015). *Analyzing Political Communication with Digital Trace Data: The Role of Twitter Messages in Social Science Research*. Springer International Publishing. https://doi.org/10.1007/978-3-319-20319-5

Kaya, H., Gürpinar, F., & Salah, A. A. (2017). Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, *65*, 66–75.

Kaya, H., Gürpinar, F., & Salah, A. A. (2017). Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, *65*, 66–75. https://doi.org/10.1016/j.imavis.2017.01.012

Kim, J.-k., & Tam, S. M. (2020). Data integration by combining big data and survey sample data for finite population inference. *arXiv preprint arXiv:2003.12156*.

King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. *Science*, *331*(6018), 719–721. https://doi.org/10.1126/science.1197872

King, G., & Persily, N. (2019). A New Model for Industry–Academic Partnerships. *PS: Political Science & Politics*, 1–7. https://doi.org/10.1017/S1049096519001021

Konitzer, T., Allen, J., Eckman, S., Howland, B., Mobius, M. M., Rothschild, D. M., & Watts, D. (2020). Measuring News Consumption With Behavioral Versus Survey Data. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3548690

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

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

Kreuter, F., Sakshaug, J. W., & Tourangeau, R. (2016). The framing of the record linkage consent question. *International Journal of Public Opinion Research*, *28*(1), 142–152.

Li, S., & Deng, W. (2020). Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing*, 1–1. https://doi.org/10.1109/TAFFC.2020.2981446

Lohr, S. L. (2008). Coverage and sampling. *International handbook of survey methodology*, 97–112.

Lohr, S. L. (2009). Multiple-frame surveys. *Handbook of statistics* (pp. 71–88). Elsevier.

Lovestone, S., & Consortium, E. (2020). The european medical information framework: A novel ecosystem for sharing healthcare data across europe. *Learning Health Systems*, *4*(2), e10214.

Manikonda, L., Hu, Y., & Kambhampati, S. (2014). Analyzing user activities, demographics, social network structure and user-generated content on instagram. *arXiv preprint arXiv:1410.8099*.

Martínez-García, Á. (2017). *La imagen en la era digital*. Egregius. Retrieved August 12, 2020, from https://idus.us.es/handle/11441/91571

Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, *4*(3), 2053168017720000. https://doi.org/10.1177/2053168017720008

Menchen-Trevino, E. (2016). Web historian: Enabling multimethod and independent research with real-world web browsing history data. *IConference 2016 Proceedings*.

Messing, S., DeGregorio, C., Hillenbrand, B., King, G., Mahanti, S., Mukerjee, Z., Nayak, C., Persily, N., State, B., & Wilkins, A. (2020). Facebook PrivacyProtected Full URLs Data Set [type: dataset]. https://doi.org/10.7910/DVN/TDOAPG

Munafò, M. R., & Davey Smith, G. (2018). Robust research needs many lines of evidence. *Nature*, *553*(7689), 399–401. https://doi.org/10.1038/d41586018-01023-3

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Myers, G. J., Badgett, T., Thomas, T. M., & Sandler, C. (2004). *The art of software testing* (Vol. 2). Wiley Online Library.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. L. Rev.*, *79*, 119.

Oberski, D. L., Kirchner, A., Eckman, S., & Kreuter, F. (2017). Evaluating the quality of survey and administrative data with generalized multitraitmultimethod models. *Journal of the American Statistical Association*, *112*(520), 1477–1489.

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

Papacharissi, Z. (2010). *A networked self: Identity, community, and culture on social network sites*. Routledge.

Parry, D. A., Davidson, B. I., Sewall, C. J., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour*, 1–13. https://doi.org/10.1038/s41562-021-01117-5

Patil, P., Peng, R. D., & Leek, J. T. (2016). A statistical definition for reproducibility and replicability. *BioRxiv*, 066803.

Pentland, A. (2010). *Honest signals: How they shape our world*. MIT press.

Perriam, J., Birkbak, A., & Freeman, A. (2020). Digital methods in a post-api environment. *International Journal of Social Research Methodology*, *23*(3), 277–290.

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

Privitera, G. J. (2018). *Research methods for the behavioral sciences*. Sage Publications.

Quan-Haase, A., & Young, A. L. (2010). Uses and gratifications of social media: A comparison of facebook and instant messaging. *Bulletin of science, technology & society*, *30*(5), 350–361.

Reeves, B., Ram, N., Robinson, T. N., Cummings, J. J., Giles, C. L., Pan, J., Chiatti, A., Cho, M., Roehrick, K., Yang, X., et al. (2019). Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them. *Human–Computer Interaction*, 1–52.

Reeves, B., Ram, N., Robinson, T. N., Cummings, J. J., Giles, C. L., Pan, J., Chiatti, A., Cho, M., Roehrick, K., Yang, X., et al. (2021). Screenomics: A framework to capture and analyze personal life experiences and the ways that technology shapes them. *Human–Computer Interaction*, *36*(2), 150–201.

Revilla, M., Ochoa, C., & Loewe, G. (2017). Using Passive Data From a Meter to Complement Survey Data in Order to Study Online Behavior. *Social Science Computer Review*, *35*(4), 521–536. https://doi.org/10.1177/0894439316638457

Ruktanonchai, N. W., Ruktanonchai, C. W., Floyd, J. R., & Tatem, A. J. (2018). Using google location history data to quantify fine-scale human mobility. *International Journal of Health Geographics*, *17*(1), 28.

Saris, W. E., & Gallhofer, I. N. (2007). *Design, evaluation, and analysis of question-naires for survey research* [OCLC: ocm80358503]. Wiley-Interscience.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological methods*, *7*(2), 147.

Scheerman, M., Voort, L., & Zarrabi, N. Secure processing of sensitive data on shared hpc systems. In: 2020. https://www.compbiomed-conference.org/wp-content/uploads/2020/03/Fri_11_15_Narges_Zarrabi-slides-compbiomed-website_smaller.pdf

The schemes rely on global open standards. (2021). https://www.europeanpaymentscouncil.eu/what-we-do/sepa-payment-scheme-management/schemes-rely-global-open-standards

Schoen, H., Gayo-Avello, D., Takis Metaxas, P., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media (D. Gayo-Avello Panagiotis Takis Metax, Ed.). *Internet Research*, *23*(5), 528–543. https://doi.org/10.1108/IntR-06-2013-0115

Sen, I., Floeck, F., Weller, K., Weiss, B., & Wagner, C. (2019). A total error framework for digital traces of humans. *arXiv preprint arXiv:1907.08228*.

Settanni, M., Azucar, D., & Marengo, D. (2018). Predicting Individual Characteristics from Digital Traces on Social Media: A Meta-Analysis. *Cyberpsychology, Behavior, and Social Networking*, *21*(4), 217–228. https://doi.org/10.1089/cyber.2017.0384

Shirima, K., Mukasa, O., Schellenberg, J. A., Manzi, F., John, D., Mushi, A., Mrisho, M., Tanner, M., Mshinda, H., & Schellenberg, D. (2007). The use of personal digital assistants for data entry at the point of collection in a large household survey in southern tanzania. *Emerging themes in epidemiology*, *4*(1), 1–8.

Singer, E. (1993). Informed consent and survey response: A summary of the empirical literature. *Journal of Official Statistics*, *9*(2), 361.

Singh, R. G., & Ruj, S. (2020). A Technical Look At The Indian Personal Data Protection Bill [arXiv:2005.13812]. *arXiv:2005.13812* [*cs*]. Retrieved August 12, 2020, from http://arxiv.org/abs/2005.13812

Skatova, A., & Goulding, J. (2019). Psychology of personal data donation. *PloS one*, *14*(11), e0224240.

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field.

Stodden, V., Leisch, F., & Peng, R. D. (2014). *Implementing reproducible research*. CRC Press.

Stodden, V., & Miguez, S. (2014). Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *Journal of Open Research Software*, *2*(1).

Suda, Y. (2020). Japan's Personal Information Protection Policy Under Pressure. *Asian Survey*, *60*(3), 510–533. https://doi.org/10.1525/as.2020.60.3.510

Szell, M., Lambiotte, R., & Thurner, S. (2010). Multirelational organization of large-scale social networks in an online world [Publisher: National Academy of Sciences Section: Social Sciences]. *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1004008107

Utrecht university research data management support [Accessed: 2020-07-18]. (n.d.).

Valkenburg, P. M., Sumter, S. R., & Peter, J. (2011). Gender differences in online and offline self-disclosure in pre-adolescence and adolescence. *British Journal of Developmental Psychology*, *29*(2), 253–269.

Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical Tools for Designing and Weighting Survey Samples*. Springer International Publishing : Imprint: Springer. Retrieved August 5, 2020, from https://link.springer.com/10.1007/978-3-319-93632-1

van der Sloot, B. (2020). *The general data protection regulation in plain language*. Amsterdam University Press.

van Eldik, A., Kneer, J., & Jansz, J. (2019). Urban & online: Social media use among adolescents and sense of belonging to a super-diverse city. *Media and Communication*, *7*(2), 242–253.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* (*Harvard JOLT*), *31*(2), 841–888. Retrieved August 7, 2020, from https://heinonline.org/HOL/P?h=hein.journals/hjlt31&i=860

Weeks, B. E., Menchen-Trevino, E., Calabrese, C., Casas, A., & Wojcieszak, M. (2021). Partisan media, untrustworthy news sites, and political misperceptions. *New Media & Society*, 14614448211033300.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., et al. (2016). The fair guiding principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1–9.

Wojcieszak, M., Menchen-Trevino, E., Goncalves, J. F., & Weeks, B. (2021). Avenues to news and diverse news exposure online: Comparing direct navigation, social

media, news aggregators, search queries, and article hyperlinks. *The International Journal of Press/Politics*, 19401612211009160.

Wong, J., & Henderson, T. (2019). The right to data portability in practice: Exploring the implications of the technologically neutral gdpr. *International Data Privacy Law*, *9*(3), 173–191.

Zhang, L.-C. (2012). Topics of statistical theory for registerbased statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63.