CHAPTER TWO

Bad Judgment, Bad Ethics?

Validity in Computational Social Media Research

CORNELIUS PUSCHMANN

Data quality is a key concern in empirical social science. In quantitative research paradigms, data quality reflects the ability of a variable to allow valid inferences about social processes or entities (Trochim & Donnelly, 2006, p. 20). In this chapter, I discuss the role of data quality in relation to research ethics. I frame data quality as an ethical issue (in addition to being a methodological one) because a particular set of assumptions about what data is shapes both the methodological and ethical considerations of researchers. I draw on several cases that have been critically discussed by the scientific community in relation to their operationalization, including Google Flu Trends (Carneiro & Mylonakis, 2009) and the so-called Facebook emotional contagion study (Kramer, Guillory, & Hancock, 2014). I close by showing how the field is progressing in terms of both ethical and methodological considerations.

INTRODUCTION

How valid are the results of analyses that rely on the digital breadcrumbs that all of us leave behind when we use the internet? While initially this hardly seems to be a question related to ethics,[1] I argue that in computational research, data quality and operationalization are equally methodological and ethical issues that impact both academia and industry research. Billions of users log on to their preferred platforms on a daily basis, generating petabytes of what is sometimes called digital trace data for its ability to function as the record of interaction with a platform, as well as

a basis for inferences about social behavior more broadly (Cioffi-Revilla, 2010; Golder & Macy, 2014; Lazer et al., 2009; Ruths & Pfeffer, 2014; Strohmaier & Wagner, 2014). These data can be harnessed for a variety of purposes, from disaster prevention to credit scoring and is hoped to shed a new light on well-established social phenomena (Shah, Capella, & Neuman, 2015; Watts, 2015).

The Google Flu Trends (GFT) case discussed by Lazer, Kennedy, King, and Vespignani (2014) is a case in point because it highlights several of these entirely practical problems. GFT's predictions turned out to be inaccurate because of confounded variables or, as the authors wryly acknowledge, "the initial version of GFT was part flu detector, part winter detector" (p. 1203). Mahrt and Scharkow (2013) highlight this difficulty, phrased in slightly different terms, when they caution that "scholars should be careful to take data for what it truly represents (e.g., traces of behavior), but not infer too much about possible attitudes, emotions, or motivations of those whose behavior created the data" (p. 24), while Giglietto and Rossi argue somewhat more optimistically that "the idea of using user-generated content for sociological research may be considered an extension of traditional study based on the content analysis of data produced by mass media" (p. 34). Yet the operationalization in GFT that equates fluctuations in search queries with flu outbreaks was clearly inaccurate, because too many sources of error stand between the human expression via a search query and the ability for its sheer frequency to reliably and robustly predict a particular medical condition. In what follows, I will outline some considerations regarding the quality and qualities of social media data in relation to social research with a computational focus. In particular, I will highlight the unintended consequences of faulty operationalization.

WHERE DOES THE DATA COME FROM, AND WHAT IS DONE WITH IT?

Just as digital data are widely seen as the raw material that fuels social media research, its methods are the tools that transform this raw material into knowledge. The picture of handling physical objects, while evocative, comes with certain limitations. Data in this area of research are often secondary, meaning they are generated for a purpose other than research and later appropriated (or "cooked") for this end, often raising a range of complex ethical questions (Bowker, 2013; boyd & Crawford, 2012; Metcalf & Crawford, 2016; Zimmer, 2010). All data need interpretation, but appropriating content created for other purposes than research is inherently risky. Data from a survey or experiment may be detrimentally affected by biases, such as social desirability in responses, or by the artificiality of a laboratory setting, but experimental data, though cumbersome to produce, are also under much closer control by the researcher than communication or log

data that are collected as an afterthought, subjected to post-hoc analysis, and often interpreted at the aggregate level. Judging people by the digital traces that they leave behind is different from following a physical trail. Hypothesis-testing in particular is problematic when the articulation of questions takes place after collecting data, when an incentive exists to confirm a hypothesis, rather than to reject it.

Just as social media research draws on a wide range of data, from tweets and Facebook comments to network data and log files, the methods used by most quantitative computational researchers are collected from a variety of academic fields and from industry research, and assembled depending on the concrete aims involved in a project. Put together, these methods form an eclectic toolbox that leaves much room for interpretation and speculation. An underlying argument in what follows is that the data used in social media research are *signs* rather than *traces*, and that, accordingly, a semiotic perspective on their meaning as relatively flexible is instructive. Social media researchers accordingly are interpreters of signs and the methods at their disposal aim to enable powerful analyses based on large volumes of signs. However, there is something unusual about the understanding of signs in (quantitative) social media research, namely that their malleability is very much productively utilized in research, while the programmatic discourse tends to downplay it. This tension is exemplified by what Jungherr (2015) refers to as *the mirror-hypothesis* and *the mediation-hypothesis*. According to the mirror-hypothesis, digital trace data represent the social world. This analogical view is comparable to that implicit in experiments in disciplines such as economics or social psychology, with the important difference that in those contexts the laboratory setting is general more similar to that of a controlled trial in the natural sciences. In social media research, data are appropriated and re-conceptualized by scientists from their original context of use and purpose. The mediation-hypothesis, according to Jungherr, posits that media have an inherent logic through which they breed their own self-referential effects; effects which are not based on analogy with the physical social world:

> Following the mirror- hypothesis, we should expect digital trace data to offer a true image of political reality. In contrast, the mediation-hypothesis leads us to expect the reflection of political reality, found in digital trace data, to be biased in accordance with the underlying data generating processes. (Jungherr, 2015, p. 63)

Imagine the notion of friends on Facebook for a moment. A naive interpretation would assume that Facebook friends faithfully represent "actual" friends. But it is common knowledge that this is not true and that Facebook friends are something very different from friends in the traditional sense. Not only this, but also does it become easier with the entrenchment of Facebook to rely on people's knowledge of *Facebook friends* as a distinct concept and to assume that others are familiar with the social conventions that form around Facebook friending. But the original

assumption of analogy between a real antecedent and a virtual shadow is perfectly rational before the background of the introduction of computer, the internet and the web. All of these are steeped in metaphors based on physical and familiar processes which are consequently applied to new and unfamiliar environments.

The quantitative methods used in social media research rely on turning information in various formats into numbers, and in applying statistical procedures to these numbers to express relationships among them. Information must be mathematized to be data that can be computationally analyzed and statistical procedures are applied both to inductively discover patterns in the data and to deductively test hypotheses. Far from there being "no theory" needed to interpret these numbers, the devices that record them are powerful mediating devices between social actors and between them and the researchers who study them (Gitelman, 2013; Manovich, 2012; Schroeder, 2014). Data analysis is usually preceded by a sequence of steps that include data acquisition, selection, conversion, restructuring, aggregation, and project-internal presentation. Furthermore, machine learning approaches allow both to discover clusters, mine association rules and construct decision trees from data, and to apply supervised learning where a manual annotation is reproduced by an algorithm. In some cases, the relevant statistical procedures are quite closely related, for example, when logistic regression is applied in supervised machine learning (another example is analysis of variance, or *ANOVA*). Yet, the disciplinary traditions even in the relatively narrow space of statistical analysis are clearly visible, with mathematically similar procedures playing distinctly different roles from one field to the next.

WHAT COUNTS AS DATA (AND TO WHOM)?

Data collection in the social sciences is traditionally an arduous enterprise, or, as Scott Golder and William Macy phrase it "social life is very hard to observe" (2014, p. 130). In addition to the risk of bias in its generation through factors such as social desirability, the sheer cost of data collection needs to be accounted for in every study. Golder and Macy point to longitudinal panel research such as the Framingham Heart Study (1948/2016), which is rare, costly, and often relies on relatively small samples, to underscore this precarious situation. Furthermore, as Murthy and Bowman note, "quantitative sociology has been traditionally driven by manageable, structured data sets" (2014, p. 2), in contrast to the massive volumes of largely unstructured data available online. Social scientists also differ considerably in what they accept as data. Individual fields, from anthropology to political science and sociology to economics, vary significantly in their data practices, often more than text book narratives suggest (cf. Borgman, 2015). Differences are not only patterned along disciplinary lines, but also reflect more granular

philosophical distinctions, from the area of research right down to national and local conventions, as well as personal preferences. In spite of these distinctions, to a social scientist data is generally something to be elicited, collected, or observed. It is brought into the world through a series of carefully planned and controlled actions, or culled from a larger body of information using specific sampling criteria. The function of data within the empirical paradigm is to represent the social world, and this is assured by its potential to be valid, reliable, and representative. Qualitative and quantitative research paradigms place very different demands on data and are subject to diverging assumptions and expectations, often by different (and sometimes warring) academic tribes. A broad consensus, however, is that data generation should be a visible part of the research cycle. Data is not natural, but profoundly man-made (Gitelman, 2013). It does not simply come into being by itself, but is either the result of a planned process of elicitation or of purposeful sampling. Such processes are often made to appear more straight-forward in the ideal environment of a text book or an introductory methods class than they turn out to be in actual research.

By contrast, data in computer science are usually considered to be any information in computable form. The ability to process information at scale is perhaps the oldest single research interest in the history of informatics. While the handling and storage of data is of key importance in this perspective, what it represents is not usually essential to the question of how it should be processed. Contrasting the understanding of data in social science and computer science reveals a combination of similarities and differences. A shared assumption is that data are an important resource for generating and transmitting knowledge, though opinions differ on what should be considered knowledge and what should not. Linked to this is the functional understanding of data as a representation of the social world in social science and the formal view of data as any machine-readable information in computer science. A further difference is sheer scale. Social scientists are familiar with data sets in which a few thousand observations are generally considered to be large, while computer scientists have long worked with data bases consisting of millions of records (cf. Schroeder, 2014). Datafication (van Dijck, 2014), or the tendency to create data to reflect more and more things digitally, extends the reach of computation to an ever-growing number of areas. Seen in a historical context, we can consider the high demands placed on the quality of data in the social sciences both as a function of its generation (often some form of dull physical or intellectual labor by the researcher) and its relative scarcity, while the quality of data in computer science is chiefly a formal concern in relation to its processing (previously costly and slow, increasingly cheap, fast, and easily extensible).

While "big data," to computer scientists, can and often does include machine-generated information from remote sensors, such as telescopes, or from internet-enabled devices and the growing "internet of things," what is presently

often studied under the banner of social media research is digital discourse, taken from sites such as Wikipedia or from social media platforms such as Twitter, Facebook, or Reddit (Strohmaier & Wagner, 2014; Tufekci, 2014). Images, videos, and other user-generated media increasingly supplement this picture (Procter, Vis, & Voss, 2013), as do geolocation data, server log files, and search queries. Various long and short snippets of text (in practice e-mails, wall posts, tweets, comments, answers, messages) are enveloped by other written information in more structured form, such as user profiles, and surrounded by a combination of platform signals (friends, followers, faves, likes) and platform-generated meta-data (cookies, time stamps, client software ID strings) that are inadvertently recorded as the user interacts with the platform. This amalgam includes data that users themselves may not be aware of, even outside of any systematic analysis on the aggregate level. The analysis of these different data types requires a complex combination of skills, and this extends beyond mere handling to interpretation. Interpretation is particularly difficult, not only operationally (in terms of required skills) but conceptually (in terms of assumptions about what data represent). This applies to digital trace data such as a series of Facebook messages much more severely than it applies to, for example, a subject's behavior in a laboratory experiment. When the data was produced for a purpose other than research, with a particular audience in mind, and in a social or cultural context unfamiliar to the researcher, this opens the door to misinterpretation and "context collapse" (Marwick & boyd, 2010). Users address "imagined audiences" (Litt, 2012), rather than providing a convenient record of their emotions. Research should always be grounded in domain-specific knowledge, but this parameter is particularly easy to violate when large volumes of data are readily available and the data structurally fulfill properties that make them suitable for analysis with tools that are familiar to the researcher.

These challenges vary from one case to another, and much digital social research being conducted is not automatically subject to issues such as data privacy. Data produced by public institutions with citizens as their intended audience is unlikely to spark much criticism from institutional review boards (IRBs) or the media. A project that analyzes search query logs for popular topics in the US and Germany would ask different questions, use different methods, and have different ethical considerations, than one that investigates manifestations of depression through sentiment analysis of social media messages.

WHO HAS A STAKE IN DATA?

Thinking about how data is generated introduces another stakeholder to the picture, extending our view of the social data ecosystem. While vast troves of information have been digitized in recent years, and more and more traditional sources

of data, such as government statistics and public archives are continuously being made accessible online, this volume is dwarfed by what private individuals produce each day on internet platforms. This volume of information rises steadily as more people across the world gain access to cheap mobile devices and successful social media sites close the remaining gaps in their global coverage. Comparative media and communication technologies have taken considerably longer to proliferate to a level that the smart phone has achieved in barely half a decade. All this is not to say that the data from digital platforms offer a comprehensive picture of humanity, nor that they ever will. But the reach of traditional methods, such as surveys, is also severely limited, and their cost means that they must be employed much more selectively (Shah et al., 2015).

Digital traces left by users also underpin a personalization industry that has not only transformed advertising, but is also making inroads into the design of products and services previously unrelated to the internet. Knowing what people are doing and saying in digital media provides a competitive advantage whether it is in predicting sales or in tracking social and political movements. Social media platforms use their data, among other things, to continuously improve their products through intense experimental testing (Sandvig, Karahalios, & Langbort, 2014; Schroeder, 2014). Both these and future business opportunities considered, however, much of what they produce could be more valuable to scholarship than it is to improving products and services (Rudder, 2014). What Amazon knows about the literary preferences of people around the world goes far beyond what it needs to know in order to sell more books, and what Facebook activity reveals about a couple's relationship is at least as relevant to sociologists as it is to the company. While it is clear that global internet companies are ambitious and continuously adapt their business models to newfound innovations on the basis of the information that they have at their disposal, it also seems likely that they are producing more than they need, and that academia is increasingly cut off from their data and insights. Sharing data could result both in privacy headaches and in foregone revenue, which explains the hesitation of companies to engage in it more systematically, in addition to the costs associated with doing so, and the risk of raising ethical concerns (Bozdag, 2013; Puschmann & Bozdag, 2014). Twitter is a case in point for this hesitant approach. After sharing data comparably liberally in the early phase of the service, mostly to attract developers, and inadvertently instigating a veritable barrage of studies that use Twitter data, the company is now imposing increasingly stringent limitations on data access. It appears to regard data as one of its key assets, and sharing that asset too readily with anyone could be detrimental to the interests of shareholders at a time when they are not very forgiving. Before the background of recent media outrage over experiments conducted on social media platforms, it seems likely that collaborations between industry and academia will continue to raise complex legal and ethical questions

whose resolution is likely to take even longer than then proliferation of new methods (Puschmann & Burgess, 2013; Schroeder, 2014, p. 3).

Industry research can obviously not be entirely open, otherwise the above-mentioned social media stars risk losing their advantage to competitors. But perhaps it is possible to strike a balance between academic and economic interests. Apart from aiming to find patterns or mechanisms that can be considered even remotely universal, "predictive and analytic techniques can provide insight into, if not directly solve, significant social problems" (Shah et al., 2015, p. 9). Data from a wide range of contexts, from disaster relief and urban poverty to migration patterns and hate crimes, are relevant to research that can have a direct impact on combating social ills and improving government policy.

HOW DIVERSE AND REPRESENTATIVE IS DATA?

The involvement of powerful social media platforms also raises further issues, one of them being that a small number of platforms at present attract by far the most research, creating a skewed picture and risking a social data monoculture (Hargittai, 2015; Tufekci, 2014). Market concentration may well at some point in the future eliminate some of the concerns voiced by scholars. Communication scholars Merja Mahrt and Michael Scharkow (2013), for example, criticize the lack of cross-platform studies in internet research, arguing that "if researchers are interested in social network sites, multiplayer games, or online news in general, it is problematic to include only data from Facebook and Twitter, World of Warcraft and Everquest II, or a handful of newspaper and broadcast news sites" (p. 25). Zeynep Tufekci (2014) voices similar criticism when she speaks of "the model organism problem, in which a few platforms are frequently used to generate datasets without adequate consideration of their structural biases." The reference to newspaper and broadcast sites by Mahrt and Scharkow (2013) warrants emphasis because it suggests that social networking sites as a class are constituted by a large number of individual exemplars, just like individual newspapers or television broadcasters constitute "the news media." But the immense concentration of social media platforms suggests that this analogy is imperfect. Social networking sites as a class may matter less as a concept if in practice people mostly use Facebook. My point is not, by any means, that we should welcome concentration, but rather that our concept of diversity is built on a much less concentrated kind of media, where a diversity of sources differ in content, but hardly in form. Diverse sampling traditionally meant sampling across sources, but how plausible is sampling across different digital platforms? Of course commonalities are crucial, but it seems more honest to assume that the specifics of platforms shape their use, rather than aiming to generalize from one service to others on the grounds that their differences are

superficial. Stepping back from claims about generalizability is of course no small theoretical challenge and accepting both the terminology and intra-platform logic of sites such as Facebook and Twitter will no doubt be painful to social scientists.

Sampling is persistently noted in the literature as a thorny issue of social media research (Mahrt & Scharkow, 2013, p. 21). Observational studies that use online data frequently break apart the established cycle of data sampling, collection, and analysis, instead they are providing ex-post interpretations that dramatically overreach the data's validity. Sampling matters on two separate levels: to obtain an initial broad sample of everything that could be relevant to the research question, and for a second, narrower one, drawn to reduce the volume of data while retaining its representativeness. As Mahrt and Scharkow (2013, p. 28) point out, this second step may reduce big data to medium-sized data without a loss of quality, while the first step is the one that needs to be carefully tailored to the research question, and is oftentimes subject to convenience. Working with digital trace data highlights the similarities between traditional content analysis and computational social science. Media and communication research has long recognized that mediated behavior is not merely unmediated behavior that happens to be conveniently recorded in analyzable form. Mahrt and Scharkow (2013) point to a long tradition of studying and classifying messages in communication research and linguistics (p. 27). That this argument is not more widely heard has many reasons, scale being one. That discourse-analyzing computational methods such as sentiment analysis for the most part perform much less reliably than manual content analysis does is not widely acknowledged (González-Bailón & Paltoglou, 2015). Teasing out the exact relationship of sampling strategy and research objectives is crucial to evaluating how much data of what degree of diversity is needed both for adequate prediction and hypothesis-testing. While this is hardly a new issue, the tendency to use much more data than a given question requires is, and whereas in traditional research this involves the elicitation of a larger sample that is associated with more work for the researcher, this is not the case with observational data from digital media platforms. While sampling offline is subject to careful consideration not only to assure research quality, but also, one might suspect, because resources need to be strategically allocated in research projects by their principal investigators, this condition is relaxed considerably with "found" online data. As Carolin Gerlitz and Bernard Rieder (2013) observe: "The majority of sampling approaches on Twitter […] follow a non-probabilistic, non-representative route, delineating their samples based on features specific to the platform." In other words, most of the studies examined by Gerlitz and Rieder chose varieties of snowball sampling relying on keywords, seed users, or other aspects particular to Twitter. Bruns (2013) makes a similar argument when calling for "non-opportunistic data gathering," by which he means foregrounding data quality in favor of sheer quantity. Obviously the sample size does nothing to alleviate problems that follow from a strategy of convenience

sampling. David Lazer and colleagues (2014), after initially being very optimistic about the potential of computational social science, caution researchers not to succumb to "big data hubris," noting that "most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis" (p. 1203).

Speaking of *messages* or *discourse*, rather than *behavior*, seems not just to be terminological hairsplitting here, but my insistence to choose words carefully is underpinned by the observation that *behavior* often conjures up an image that is too simple and straightforward to be accurate. The operationalization in GFT that equates fluctuations in search queries with flu outbreaks is simply not a good one because too many sources of error stand between the human expression via a search query and the ability for its sheer frequency to reliably and robustly predict a particular medical condition. The signal sent by Google's users is an arbitrary one, more akin to a smoke sign than a trace.

WHAT DOES DATA SIGNIFY?

It pays off to examine the terminology used to describe what data are and where they come from both closely and critically. Van Dijck notes that "data and metadata culled from Google, Facebook, and Twitter are generally considered imprints or symptoms of people's actual behavior or moods" (2014, p. 199, my emphasis) and Mahrt and Scharkow (2013) speak of "traces […] automatically left" (p. 24). The term *traces* permeates across much of the literature, as does the analogy to the telescope (Watts, 2015). Golder and Macy (2014) prefer to speak of "digital footprints," while Strohmaier and Wagner (2014) provide the illustrative example of traces in a very physical sense, describing "the wear of floor tiles around museum exhibits as indicators of popular exhibits; the setting of car radio dials as indicators of favorite stations; the wear on library books and rub and fold marks in their pages" (p. 85). The vivid image that they provide makes a resounding point: Data traces are not always as readily interpretable as physical traces. If they were, the level of granularity that they would provide us with would be remarkable. But their potential for misinterpretation is, at least in the present stage, far greater. This need not deter us, but it is a powerful reminder that the interpretation of physical tracks on the ground is performed instantly and inadvertently by the brain, while the interpretation of a tweet's political relevance is a more complicated matter. When characterizing the generation of data, careful attention to detail is also warranted. Shah et al. (2015) refer to digital trace data as "naturally occurring," but put the adjective in quotations, as if wanting to express that such information is in many respects more "natural" than the data from surveys and traditional laboratory experiments, but also less natural than actual physical traces. Strohmaier and

Wagner (2014) discuss different terms, settling for "found data" also to express a situation where the data are generated without the researchers intervention (and chronologically before the researcher appears in the picture). They characterize such data as non-reactive and observational, in other words, as being collected without the possibility of the researcher influencing the research subject. The term is also somewhat suggestive of data as something that naturally occurs (or perhaps, that has been abandoned, and is conveniently discovered by the researcher by the side of the road), backgrounding the variation in the purpose of its creation (say, to communicate with a loved one), its storage (as part of a feature that a social media company hopes will bring more users to its site), and its analysis (by scientist hoping to publish a paper). Not much is natural about this form of eavesdropping on the conversations of others.

The fact that interaction in digital platforms is mediated should not be equated with the assumption that they are "not real." Golder and Macy (2014) ask rhetorically whether the online world is a parallel universe (p. 143) and go on to argue that it is not. They propose to "turn the tables […] rather than address the societal implications of the Internet, we survey studies that use online data to advance knowledge in the social sciences" (Golder & Macy, 2014, p. 130). Strohmaier and Wagner (2014) argue in the same direction, but add a qualification: "the World Wide Web represents not only an increasingly useful reflection of human social behavior, but every-day social interactions on the Web are increasingly mediated and shaped by algorithms and computational methods in general." Their comment suggests a double life of platforms (in this case the Web) as reflecting social life and at the same time influencing human behavior and enabling modes of expression that are intricately tied to the design of digital media services. The distinction between real and virtual, or online and offline, obscures the influence of these platforms on data creation. Online interactions are entirely real, but they are also subject to factors that do not exist in unmediated interactions, and that may change rapidly following the changing priorities of platform providers and their reflection in design. Herring (2004) identifies this kind of bias in the pre-social media Web when she argues that "computer-mediated discourse may be, but is not inevitably, shaped by the technological features of computer-mediated communication systems" (p. 338). In terms of their broad usage, their relevance to politics, the economy, and everyday life and the thoughts, emotions and relationships which they enable and support, digital platforms are entirely real. But all of these things take place on a cultural stage, to use the Goffmanian analogy, that is set by the companies running the services that we use – a set that changes with each scene, influencing the performance of the players in a variety of ways. David Lazer and colleagues (2014) seem most keenly aware of this complication, noting that a better understanding of the algorithms underlying Google, Twitter, and Facebook is crucial to both scholarship and civil society. The influence of platform providers on

data generation is interchangeably referred to as *platform politics* (Gillespie, 2010), *social media logic* (van Dijck & Poell, 2013), and *blue team dynamics* (Lazer et al., 2014). Lazer et al. note this continuous adjustment in service of the customer, pointing out that "in improving its service to customers, Google is also changing the data-generating process" (p. 1204). The problems caused by analytical feedback loops of the data generating process and subsequent data interpretation should be apparent. Data are only valid if the researcher's actions are not essential for its production. *Red team dynamics*, in Lazer et al.'s parlance, are those where "research subjects (in this case Web searchers) attempt to manipulate the data-generating process to meet their own goals, such as economic or political gain" (p. 1204). It should be conceivable that the distinction between social behavior and "manipulation" is quite a hazy one in many cases. All communication, ultimately, realizes a goal for the communicator, and often goals are determined strategically.

Imagining social media platforms as a stage that is set by the platform provider through the design of the site or app allows us to identify another complicating issue of digital trace data. Social media companies store data in structures that are reflected in the site design, or, turning this around, design the site in a particular way that has implications for what is stored, and how. Facebook likes and Twitter retweets are examples of such units of analysis that find their way into research. Likes and retweets at once serve a function for users and for Facebook and Twitter as companies (Gerlitz & Helmond, 2013). The functions and their utility for users individually is distinct from their function for the provider, particularly to the provider's advertisement-supported business model. But what is their function as indicators of social processes? Tufekci (2014) highlights this discrepancy when she argues that "users engage in practices that may be unintelligible to algorithms, such as subtweets (tweets referencing an unnamed but implicitly identifiable individual), quoting text via screen captures, and 'hate-linking' – linking to denounce rather than endorse" (p. 505). The issue in her examples applies less to algorithms, but to the expectations of platform providers towards users and their "correct" usage of the platform. These expectations may reflect business considerations, or simply result from failing to anticipate the subversive creativity of users. Often, both are at work. The debate around the introduction of a dislike button on Facebook serves to underscore this issue. Such a button does not exist because it would allow the expression of preferences which are not desirable to Facebook and could even result in legal challenges to the company. Some users have voiced their interest in such a button, and others have simply devised other ways of expressing what such a button would express. Subversive behavior is arguably of interest to social scientists, not just "proper" usage, particularly if the considerations at work are driven by what is desirable to companies. But in terms of design, if not expression, the considerations of the platform provider prevail. This simply underscores that Facebook is not a neutral observatory of social interaction. It is not just

subject to human biases in what people express, but its very design consciously aims to suppress behavior that could spell out legal or economic detriment to the provider. Among others, van Dijck (2014) notes this conflict between scientific and commercial interests when she speaks of "the paradoxical premise that social media platforms concomitantly measure, manipulate, and monetize online human behavior" (p. 200).

WHAT INTERPRETATIONS DOES DATA PERMIT?

Digital trace data are quite abstract. A physical trace seems for most purposes easier to interpret than a message, tweet, or like. A strategy to counter this problem is to utilize the powers of categorization and comparison. If there is something interesting to compare your data with, even simple descriptive statistics can be enlightening. By contrast, the application of complex methods to a single data set to produce a purely descriptive result can be frustrating because there is no reference point that would allow evaluating the results through a meaningful baseline. This point may seem trivial, but much online research is plagued by producing quantifications which are then left to stand on their own or by making comparisons which are, like sampling, the result of convenience rather than careful research design. Comparing men to women, American users to British ones, highly active individuals to sporadics, and regular weeks with unusual ones achieve this goal. The gender example is chosen deliberately because often comparisons are made without much of a clear motivation, but rather because categorical data are available by which groups can be conveniently compared. But once two groups are contrasted, this implicitly makes the claim that they exist, are sufficiently clear-cut, and play an important role in the analysis. Comparisons facilitate clarity, but what is compared by should be a conscious choice, rather than a matter of convenience. Categorizing people by certain criteria, whether it is gender, race, education, or income, is often useful in social science, but not to claim that they fit unequivocally into convenient ontological boxes, but because aggregation allows quantitative analysis, and quantitative analysis is essential to answering macro-level questions. Categorizing and comparing badly has unintended consequences which are a side effect of the distance that quantitative research creates between a researcher and her subject. That caveat is particularly important in computational social science.

I have previously established that the kind of data that computational social sciences are concerned with comes in many forms. However, for the purpose of most analyses, researchers encounter data in one of two formats in the phase that it is studied: It is usually either numeric or textual. Other formats, such as images and video, while both immensely popular and increasingly studied, place different demands on researchers in terms of skills, tools, and data processing

infrastructure. Numerical data in the aggregate take the shape of counted comments, clicks, friends, likes, or shares. A special significance, however, can be afforded to textual data. As Shah et al. argue (2015): "With much of the core social data now in textual form, changing in central ways how data are acquired and reduced, scholars will need to come to new agreements on what constitutes reliable and valid descriptions of the data; the categories used to organize those data; and the tools necessary to access, process, and structure those data" (p. 12). Textual data has a central role in social media research because so much of what people produce themselves, in contrast to information that is automatically collected about them, is text. Quantifying this in relation to images or video seems pointless, as nobody is likely to dispute the importance of these types of media, which once took dedicated equipment to produce, but can now be created with any common mobile device in excellent quality. Research methods for the analysis of digital images and video will take time to catch up for practical reasons and to make inroads into areas of social science where they are presently not widely used. Textual analysis has a long history in the social sciences and humanities, but it is worth noting that some distinctions made between different levels of analysis that are often conflated in computational approaches are of key importance to such accounts. For example, Herring (2004) distinguishes between four such levels of analysis on which data can be segmented: 1) structure, 2) meaning, 3) interaction, and 4) social behavior (p. 339). Structure covers aspects as orthography, the use of emoticons, or other properties on the level of words or sentences. Meaning relates to what words, speech acts or larger units of discourse express. Thirdly, interaction includes the properties of dyadic discourse such as turn-taking or topic development and other interactional dynamics. The fourth level indicates aspects that can be more abstractly labeled as forms of social behavior, such as expressions of play, conflict, and group membership. Herring's perspective is a linguistic one, therefor her differentiation of structural and socially functional aspects may not resonate with other social scientists (for example, differentiating levels one and two is didactically common in linguistics, but may not be very practical empirically). But it is worth pointing out that much of current research ignores intermediate levels of abstraction, going instead directly from words to social behavior. It is not yet broadly recognized that a word and its meaning is highly context-depended, and consequently a bad proxy for stable analytical units such as personality traits, social relationships, or public opinion. The reason for the popularity of words as a unit of analysis in computational textual research is to be found in the economics of research feasibility. Words are much more easy to extract than other units, and they are more widely accepted as a form of data than, for example, conversational turns. The approach taken in textual social media research of "operationalizing up" from words to more abstract categories is a lasting challenge to social media research.

WHO OWNS AND CONTROLS DATA?

Finally, questions that go beyond the collection, analysis and interpretation of data also need to be addressed. Who owns digital platform data is a point of ongoing debate among legal scholars. Initially, in many contexts, the answer is "no one", at least not in the sense of legal ownership. Data, apart from a few exceptions, do not constitute intellectual property. While the suggestion has been made that users have a natural right to the data that they produce and the meta-data that surround it, such data are generally not considered to constitute property. Laws protecting the privacy of users apply to social media platforms, but the fact that most information is disclosed willingly in such platforms and that providers are usually granted the right to analyze the data and experiment with the site's features when users sign the terms of service means that companies are under relatively few constraints to make use of the data. Attempts to regulate data about people outside of the frameworks of ownership and privacy protection, such as the "Right to be forgotten" implemented in the European Union and imposed on search results that concern individuals have met with very mixed responses. Attempts in this and similar directions underscore that data from digital platforms and what is done with it is increasingly seen a human rights issue that transcends national regulation, though political solutions to these problems seem far off.

On the other hand, cases where social media data have been used in large-scale research projects have attracted considerable media attention, particularly when the results have been published in major scientific journals. The Facebook emotional contagion experiment (Kramer et al., 2014) is one such example. Legally, researchers at Facebook had done nothing wrong, despite widespread criticism of the ethics of the study. And while a perceived lack of scrutiny by the institutional review board (IRB) that cleared the research was criticized by some commentators, others did not find the research to breach ethics guidelines. There was, however, a consensus regarding the need to develop better standards and adapt ethical codes to new forms of research. Research in social media research underscores that simply having access to data is much less important than effectively being able to query it. This requires the right tools for infrastructure and analysis, as well as the competence to interpret results. In an environment where data are ubiquitous, their mere existence seems less of an issue than their use and the outcomes of these uses. An ethical use of data in social media research must therefore be more concerned with research results and their potential to clash with the interests of users than with the mere legality of data access. As Shah et al. (2015) argue: "The acquisition and archiving of complex data systems – let alone their manipulation – often involve collecting personally identifiable information […] this forces some reflection on issues of data privacy and de-identification, especially in an era of increased tracking of expression and action" (p. 8).

SUMMARY

This chapter has examined how data are collected, processed, and interpreted in computational social media research, and how a lack of concept validity frequently dogs ambitious research in this area of study. The flavor of social science that this emerging field embraces is strongly concerned with making scientific inferences on human behavior, yet it has been shown that observational findings based on social media data can frequently not be reproduced (Liang & Fu, 2015). From predicting elections to forecasting consumption, what people do is a key interest of the field. Even when asking very theoretical questions about human sociality, such questions need to be quantifiable in order to fit into the computational paradigm. This is not to say that social media research rejects qualitative insight. Qualitative and quantitative research can be integrated into new approaches and very often this yields the best results (Bastos & Mercea, 2015). Some warn of a crisis of empirical social research if new methods and sources of data are left to computer science and eschewed by social scientists (Savage & Burrows, 2007). But the role afforded in social media research to computation, and therefore some form of quantification, comes with certain limitations. Data are used to answer a set of questions or patterns within data are identified and related to particular behavior. Social media research thrives on data, particularly of the observational kind. These data are generally not produced with research in mind, but accumulate in online platforms as a by-product of a user's actions, sometimes without their knowledge. They are produced for particular purposes and with particular addresses in mind. The researcher ideally has a form of privileged access to these data, putting her under both the methodological and ethical obligation to produce a sound analysis.

NOTE

1. But see Giglietto and Rossi (2012) who also make this connection.

REFERENCES

Bastos, M. T., & Mercea, D. (2015). Serial activists: Political Twitter beyond influentials and the twittertariat. *New Media & Society*. Retrieved from http://doi.org/10.1177/1461444815584764

Borgman, C. L. (2015). *Scholarship in the digital age*. Cambridge: MIT Press. Retrieved from http://doi.org/10.1017/CBO9781107415324.004

Bowker, G. C. (2013). Data flakes: An afterword to "raw data" is an oxymoron. In L. Gitelman (Ed.), *"Raw data" is an oxymoron* (pp. 167–171). Cambridge, MA: MIT Press.

boyd, d., & Crawford, K. (2012). Critical questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society, 15*(5), 662–679. Retrieved from http://doi.org/10.1080/1369118X.2012.678878

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and Information Technology, 15*(3), 209–227. Retrieved from http://doi.org/10.1007/s10676-013-9321-6

Bruns, A. (2013). Faster than the speed of print: Reconciling "big data" social media analysis and academic scholarship. *First Monday*, *18*(10). https://doi.org/10.5210/fm.v18i10.4879

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases, 49*(10), 1557–1564. Retrieved from http://doi.org/10.1086/630200

Cioffi-Revilla, C. (2010). Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(3), 259–271. Retrieved from http://doi.org/10.1002/wics.95

Framingham Heart Study. (1948/2016). Retrieved from https://www.framinghamheartstudy.org/

Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society, 15*(8), 1348–1365. Retrieved from http://doi.org/10.1177/1461444812472322

Gerlitz, C., & Rieder, B. (2013). Mining one percent of Twitter: Collections, baselines, sampling. *M/C Journal, 16*(2), 1–15. Retrieved from http://www.journal.media-culture.org.au/index.php/mcjournal/article/viewArticle/620

Giglietto, F., & Rossi, L. (2012). Ethics and interdisciplinarity in computational social science. *Methodological Innovations Online, 7*(1), 25–36. Retrieved from http://doi.org/10.4256/mio.2012.003

Gillespie, T. (2010). The politics of "platforms." *New Media & Society, 12*(3), 347–364. Retrieved from http://doi.org/10.1177/1461444809342738

Gitelman, L. (Ed.). (2013). *"Raw data" is an oxymoron*. Cambridge, MA: MIT Press.

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Annual Review of Sociology, 40*, 129–152. Retrieved from http://doi.org/10.1146/annurev-soc-071913-043145

González-Bailón, S., & Paltoglou, G. (2015). Signals of public opinion in online communication: A comparison of methods and data sources. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 95–107. Retrieved from http://doi.org/10.1177/0002716215569192

Hargittai, E. (2015). Is bigger always better? Potential biases of big data derived from social network sites. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 63–76. Retrieved from http://doi.org/10.1177/0002716215570866

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online behavior. In S. Barab, R. Kling, & J. H. Gray (Eds.), *Designing for Virtual Communities in Service of Learning* (pp. 338–376). New York, NY: Cambridge University Press.

Jungherr, A. (2015). *Analyzing political communication with digital trace data*. Heidelberg: Springer International Publishing. Retrieved from http://doi.org/10.1007/978-3-319-20319-5

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

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

Lazer, D., Pentland, A., Adamic, L. A., Aral, S., Barabasi, A.-L., Brewer, D., … Van Alstyne, M. (2009). Computational social science. *Science, 323*(5915), 721–723. Retrieved from http://doi.org/10.1126/science.1167742

Liang, H., & Fu, K. (2015). Testing propositions derived from Twitter studies: Generalization and replication in computational social science. *PLoS One, 10*(8), e0134270. Retrieved from http://doi.org/10.1371/journal.pone.0134270

Litt, E. (2012). Knock, knock. Who's there? The imagined audience. *Journal of Broadcasting & Electronic Media, 56*(3), 330–345. Retrieved from http://doi.org/10.1080/08838151.2012.705195

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media, 57*(1), 20–33. Retrieved from http://doi.org/10.1080/08838151.2012.761700

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (pp. 460–475). Minneapolis, MN: University of Minnesota Press.

Marwick, A., & boyd, d. (2010). I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society, 13*(1), 114–133. Retrieved from http://doi.org/10.1177/1461444810365313

Metcalf, J., & Crawford, K. (2016, June). Where are human subjects in big data research? The emerging ethics divide. *Big Data and Society*, 1–34. Retrieved from http://doi.org/10.1177/2053951716650211

Murthy, D., & Bowman, S. A. (2014). Big Data solutions on a small scale: Evaluating accessible high-performance computing for social research. *Big Data & Society, 1*(2), 1–12. Retrieved from http://doi.org/10.1177/2053951714559105

Procter, R., Vis, F., & Voss, A. (2013). Reading the riots on Twitter: Methodological innovation for the analysis of big data. *International Journal of Social Research Methodology, 16*(3), 197–214. Retrieved from http://doi.org/10.1080/13645579.2013.774172

Puschmann, C., & Bozdag, E. (2014). Staking out the unclear ethical terrain of online social experiments. *Internet Policy Review, 3*(4), 1–15. Retrieved from http://doi.org/10.14763/2014.4.338

Puschmann, C., & Burgess, J. (2013). The politics of Twitter data. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and society* (pp. 43–54). New York, NY: Peter Lang.

Rudder, C. (2014). *Dataclysm: Who we are (When we think no one's looking)*. New York, NY: Crown.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science, 346*(6213), 1063–1064. Retrieved from http://doi.org/10.1126/science.346.6213.1063

Sandvig, C., Karahalios, K. G., & Langbort, C. (2014, July 24). *Christian Sandvig, Karrie G. Karahalios, and Cedric Langbort look inside the Facebook News Feed*. Retrieved from http://blogs.law.harvard.edu/mediaberkman/2014/07/24/christian-sandvig-karrie-g-karahalios-and-cedric-langbort-look-inside-the-facebook-news-feed-audio/

Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology, 41*(5), 885–899. Retrieved from http://doi.org/10.1177/0038038507080443

Schroeder, R. (2014). Big Data and the brave new world of social media research. *Big Data & Society, 1*(2), 1–11. https://doi.org/10.1177/2053951714563194

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science: Possibilities and perils. *The ANNALS of the American Academy of Political and Social Science, 659*(1), 6–13. Retrieved from http://doi.org/10.1177/0002716215572084

Strohmaier, M., & Wagner, C. (2014). Computational social science for the World Wide Web. *IEEE Intelligent Systems, 29*(5), 84–88. Retrieved from http://doi.org/10.1109/MIS.2014.80

Trochim, W., & Donnelly, J. P. (2006). *Research methods knowledge base* (3rd ed.). Mason, OH: Atomic Dog. Retrieved from https://doi.org/10.1515/9783110858372.1

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (pp. 505–514).

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & Society, 12*(2), 197–208.

Van Dijck, J., & Poell, T. (2013). Understanding social media logic. *Media and Communication, 1*(1), 2. Retrieved from http://doi.org/10.17645/mac.v1i1.70

Watts, D. J. (2015). Common sense and sociological explanations. *American Journal of Sociology, 120*(2), 313–351.

Zimmer, M. (2010). "But the data is already public": On the ethics of research in Facebook. *Ethics and Information Technology, 12*(4), 313–325. Retrieved from http://doi.org/10.1007/s10676-010-9227-5