Jeremy Hunsinger
,
Lisbeth Klastrup
and
Matthew M. Allen
Second International Handbook of Internet Research
10.1007/978-94-024-1202-4_15-1

Blended Data: Critiquing and Complementing Social Media Datasets, Big and Small

sky croeser[1] and Tim Highfield[2]
(1)Curtin University, Perth, WA, Australia
(2)University of Amsterdam, Amsterdam, North Holland, Netherlands

sky croeser (Corresponding author)
Email: s.croeser@curtin.edu.au

Tim Highfield
Email: tim.highfield@gmail.com

Abstract

Internet research, and especially social media research, has benefited from concurrent factors, technological and analytical, that have enabled access to vast amounts of user data and content online. These trends have accompanied a prevalence of Big Data studies of online activity, as researchers gather datasets featuring millions of tweets, for instance – here, Big Data is a reference not solely to the size of datasets but to the wider practices and research cultures around large-scale and exhaustive (and often ongoing) capture of data from large groups, often (but not always) studied quantitatively (see Kitchin and Lauriaut 2014a; Crawford et al. 2014). However, the accessibility of "big social data" (Manovich 2012) for Internet studies research is not without its limitations and challenges, and while extensive datasets enable valuable research, combining them with small data can provide more rounded perspectives and encourage us to think more about what we are studying. Similarly, privileging the online-only or the quantitative analysis of social media activity may overlook or mask key practices and relevant participants not present within the datasets. We argue for a blended data model as a critique and complement for different social media datasets, drawing in part on our research into social movements and activists' use (and non-use) of online technologies. Together, these approaches may overcome and negotiate the respective limits and challenges of social media data, both big and small.

Keywords Social media - Big Data - Ethics - Methods - Social movements
Internet research, and especially social media research, has benefited from concurrent factors, technological and analytical, that have enabled access to vast amounts of user data and content online. Researchers can automate the collection of social media activity through application programming interfaces (APIs) on Twitter, Facebook, and Instagram and employ free tools and techniques for gathering and analyzing this data. These trends have accompanied a prevalence of Big Data studies of

online activity, as researchers gather datasets featuring millions of tweets, for instance – here, Big Data is a reference not solely to the size of datasets but to the wider practices and research cultures around large-scale and exhaustive (and often ongoing) capture of data from large groups, often (but not always) studied quantitatively (see Kitchin and Lauriaut *2014a*; Crawford et al. *2014*). (As an aside, one question of "Big Data" is whether it should be capitalized or presented in inverted commas (or neither); the various studies we cite here do not provide a consistent take on this. For this chapter, we have referred to Big Data, following boyd and Crawford's use of capitals to denote a set of practices (*2012*), with exceptions made for quotations and citations.) However, the accessibility of "big social data" (Manovich *2012*) for Internet studies research is not without its limitations and challenges, and while extensive datasets enable valuable research, combining them with small data can provide more rounded perspectives and encourage us to think more about what we are studying – as also raised by André Brock (*2015*) in arguing for, and developing, methodologies for "deep data" analysis. Similarly, privileging the online-only or the quantitative analysis of social media activity may overlook or mask key practices and relevant participants not present within the datasets. We argue, then, for a blended data model as a critique and complement for different social media datasets, drawing in part on our research into social movements and activists' use (and nonuse) of online technologies. Together, these approaches may overcome and negotiate the respective limits and challenges of social media data, both big and small. While "small data" may share characteristics with Big Data but offer less exhaustive populations as the object of study and represent infrequent or one-off data collection processes (Kitchin and Lauriaut *2014a*), our reference to "small" data here is not intended to portray such datasets as less complex or rewarding than Big Data: rather, small data may be treated as "deeper data," offering an opportunity to examine practices, uses (and nonuses), and other aspects of the data in detail – especially in combination with additional data sources and methods.

The critiques and approaches we outline here have been developed through our *Mapping Movements* research (Croeser and Highfield *2014*, *2015a*, *b*): this has encompassed fieldwork and social media analysis of social movements and events including Occupy Oakland, antifascist activism in Greece, and the 2013 World Social Forum in Tunisia. Our methodology brings together interviews and ethnographic approaches with digital methods, such as online issue mapping and social network analysis. In combining qualitative and quantitative approaches, for datasets which if not Big Data in themselves at least share characteristics with the large-scale study of social media activity, we reconcile the experiences of social movements as both physical and online manifestations. Using interviews and observations from the scene provides one perspective on a movement like Occupy, while its socially mediated online form may differ dramatically in terms of scope and participants – yet both elements are part of the same overall context. Our mixed methods utilize blended data as a means of addressing the limitations and gaps present in focusing on only one of these elements.

Our focus on social movements and their use of – and presentation through – online platforms raises questions about the value and limits of Big Data and about ethics in social media and social movement research. While some of these questions are perhaps more specific to controversial and sensitive contexts of activism and protest, they serve to demonstrate broader issues applicable to Internet research into diverse topics, including the everyday and banal online activity. It is also vital to remember that data which researchers may see as unimportant, or not particularly emotionally charged, may nevertheless be deeply personal to some social media users, precisely because of its quotidian nature. Lessons taken from the social movements' context, including around ethics and analytical and data biases, are relevant to other social media datasets to varying degrees. Social movement research may be considered as an edge case, and in this chapter we use this context as a means of identifying and addressing methodological and analytical issues, questions, and limitations

that are critical considerations for the field. As we developed this chapter, we found that many of our concerns mapped onto boyd and Crawford's (*2012*) provocations around the use of Big Data methods: here, we attempt to further expand on their work and discuss some of the responses and reflections which emerged out of the *Mapping Movements* project that may be relevant to research drawing on Big Data approaches. We also note that this remains a work in progress: as we develop our research, new issues arise, and activists and other social media also raise new critiques about how academics, journalists, and others use social media "data."

Why Blended Data?

Blended data approaches are a particular kind of mixed methods research, which is usually understood to draw on both quantitative and qualitative methodologies (Johnson et al. *2007*). It is also useful to bear in mind Hesse-Biber and Johnson's broader definition of mixed methods as "research and inquiry that includes 'multiple and mixed' research projects that facilitate and reside at the intersections of multiple methods, purposes, kinds of data, and levels of analysis (e.g., micro, meso, macro), as well as a range of academic disciplines, paradigms, axiologies, stakeholders, and cultures of research and practice" (*2013*, p. 103). This definition suggests research which goes beyond simply drawing together different methods: it also involves building a dialogue around the limitations and benefits of different approaches, delving more deeply into the underlying assumptions underpinning these approaches, and being prepared to unsettle them. In our case, Tim's Internet studies setting developed via communication studies and French studies, and sky comes from a background in political science and international relations. This means not only different methodological training but also different forms of analysis, concerns, and even ethical frameworks around the purpose and possibilities of research.

Within our *Mapping Movements* research , "blended data" has not taken one specific form: while seeking to combine digital methods and fieldwork, our case studies have featured different foci in considering the physical and the digital. Our objects of research have varied, from Twitter to activist blogs and Web radio and different interviews and participant observation across the case studies thus far. There is no explicit formula to "blended data." Rather, there are multiple approaches researchers may follow to extend and develop their analysis of Internet-related practices and phenomena. Blended datasets might reflect online-only activity, bringing together data from different social media platforms (Driscoll and Thorson *2015*; Burgess and Matamoros Fernández *2016*) or mix documentary and archival research with social media, considering both the physically tangible and the tweeted, for instance. Here, the digital may have its analogues in, and demonstrate similar practices to, older media or physical phenomena, drawing parallels between social media and, variously, gazeteering, postering, pamphleteering, and diaries (Moe *2010*; Humphreys et al. *2013*). Drawing upon multiple data sources, of different types and contexts, can serve a complementary function.

The value of blended data within Internet research, and especially Big Data-driven studies, is in marking the limitations of Big Data while simultaneously addressing and complementing these aspects. boyd and Crawford (*2012*, p. 688) argue that:

Large data sets from Internet sources are often unreliable, prone to outages and losses, and these errors and gaps are magnified when multiple data sets are used together... A data set may have many millions of pieces of data, but this does not mean it is random or representative. To make statistical claims about a data set, we need to know where data is coming from; it is similarly important to know and account for the weaknesses in that data.

Addressing these challenges and limitations can be done *within* the data-gathering, but it can also be usefully done through blended data methods. These can serve to provide the context for online activity, to understand the motivations and decisions behind use and nonuse, and to include perspectives that are both important to the data yet not necessarily present within it. Blended data methods are one possible way of pursuing a "data-driven science" (in the broad sense of "science") – combining abduction, deduction, and induction (Kitchin *2014*). They can also encourage the cross-disciplinary discussions and collaborations which have been so fruitful in the development of Internet research as a field and which can lead to new paradigms, methods, and challenges.

From Big to Blended: Provocations and Responses

While Big Data as a concept has attracted extensive attention in Internet research and other fields, the scope of Big Data research can vary dramatically: what people are talking about when they talk about Big Data is not necessarily consistent. Indeed, we could ask how big the data has to be to qualify as Big Data. Mahrt and Scharkow (*2013*) note that Big Data has been used in part to refer to datasets that require specialist analytical processes that are too extensive for standard or commonplace software, for instance, and which within social media research may feature data points (such as tweets) numbering in the millions. However, the sheer volume of data is not the only factor that might contribute to a Big Data study – and indeed, size is not everything. Arguing that Big Data is perhaps a misnomer, boyd and Crawford (*2012*, p. 663) write that "Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets."

Accompanying the rise of Big Data has "emerged a system of knowledge that is already changing the objects of knowledge, while also having the power to inform how we understand human networks and community" (boyd and Crawford *2012*, p. 665). Rather than just being a descriptor for research involving large-scale and predominantly quantitative analysis of populations and their activities, Big Data now represents more than the objects of research. As Crawford et al. outline, Big Data can be variously considered as "theory, practice, archive, myth, and rhetorical move" (*2014*, p. 1665). We can conduct Big Data analyses of online activity (and other subjects) and follow principles of Big Data in how we frame and design our research – and indeed the wider Big Data context (not just within Internet studies) brings with it its own conventions, associations, hopes, and fears (see Puschmann and Burgess *2014*).

What is clear is that Big Data refers not just to the datasets themselves – the role of researchers in constructing Big Data needs to be recognized. Vis (*2013*) underlines that Big Data does not describe a set, pre-existing object: "Things are up for grabs so to speak, before the emerging ideas about Big Data become codified and institutionalised. There is therefore an urgent imperative to question the mechanisms and assumptions around Big Data." Similarly, the role of data providers in structuring what researchers can and cannot do, and what they are able to study, should not be ignored – this refers to what is obtainable via APIs and the terms of service introduced by individual platforms (and the politics of these platforms; see Gillespie *2010*; Langlois and Elmer *2013*) and also to what researchers are technically capable of doing. At the same time, though, and as will be outlined below, there's a risk that the privileging of automated processes for capturing data from APIs (which might then be studied from Big Data-related approaches) overlooks and negates other means for data to be collected, provided, or interrogated (Vis *2013*). While there is obvious value in long-term collection of data from populations of social media users, as a source for baseline analyses and providing context for smaller-scale analyses, the ways in which Internet research has responded to questions of Big Data and its common trends and approaches are not without their critiques or limits. Indeed, the data rush

noted by Mahrt and Scharkow (*2013*) has not been evenly distributed across social media platforms or practices.

In response to the rise of Big Data studies, especially within Internet studies, danah boyd and Kate Crawford set out key provocations for Big Data research, arguing that "Big Data creates a radical shift in how we think about research" (*2012*, p. 665). In particular, their critiques address the impact of Big Data on the definition of knowledge; that claims to objectivity and accuracy are misleading; that bigger data are not always better data; that when taken out of context, Big Data loses its meaning; that just because it is accessible does not make it ethical ; and that limited access to Big Data creates new digital divides. We use these critiques as foundation for our consideration of blended data, responding to the main challenges and questions raised by boyd and Crawford within an updated social media environment and the specific context of social movement research.

Big Data Changes the Definition of Knowledge

Big Data is not just about a quantitative shift in the volume of data analyzed: the first provocation that boyd and Crawford raise regarding Big Data relates to the ways in which it, as a system of practices, is "changing the objects of knowledge," creating "a radical shift in how we think about research" (*2012*, p. 665). Zizi Papacharissi, in her reply to boyd and Crawford's work, argues that we should draw on Sandra Harding's discussion of situated knowledges to understand Big Data as shaped by a greater social reality that it also reproduces (Papacharissi *2015*, p. 3). We might, then, query Big Data's relationship with neoliberal capitalism (Qiu *2015*, p. 4), the assumptions Big Data methods are based on, and the forms of analysis that they strengthen or undermine. Part of this involves remaining critical of "an arrogant undercurrent in many Big Data debates where other forms of analysis are too easily sidelined. Other methods for ascertaining why people do things, write things, or make things are lost in the sheer volume of numbers" (boyd and Crawford *2012*, p. 666). Using blended data methods – or even carefully considering whether to use blended methods – offers a partial balance to this: it invites us to think more critically about which methods, and which kinds of data, are most appropriate for exploring a particular set of questions, destabilizing claims that a particular set of methods are more "objective."

Hesse-Biber and Johnson (*2013*, 2014) argue that combining different methods – and debating the best ways in which to combine methods – results in a turbulence which "provides the space for innovation and productive dialogue across our methods and paradigmatic standpoints. It is in the gaps between points of view where we may go after new knowledge(s) and newly emergent practices and designs. Such a turbulent environment asks each of us to be reflexive on our own researcher standpoint and be open to dialogue across our paradigmatic and methods comfort zones." Blended methods can, and should, also lead to deeper challenges to how we think about research: "Can a researcher trained in qualitative methods with interpretative philosophy practice postpositivism? Can a postpositivist practice an interpretive method?" (Hesse-Biber and Johnson *2013*, p. 105). These are not simply abstract questions: they have direct consequences for the ways in which we carry out our work and for the ways in which we understand our positions as researchers.

In our case, combining Big Data methods with participant observation and in-depth interviews raised questions about the relationship between "researchers" and "participants" that are rarely asked in Internet research which focuses purely on quantitative methods. For example, during fieldwork, it is common to try to make one's presence as a researcher visible, at the same time as attempting not to disrupt events. This raised questions about how to shift the same practice online: how would we make

it visible that we were collecting data around particular hashtags? What would we do if people using those hashtags challenged our data collection? These are ethical questions, and they are also questions which speak more deeply to the assumptions around how research works, what its purpose is, and what our roles are as researchers: considering these issues undermines assumptions around the subject/object dualism of much research and involves trying to build a different relationship between researchers and "participants" (or between researchers and "data"). Making data-gathering on spaces like Twitter visible opens researchers up to discussion and critique and hints at the possibility that they (or rather, *we*) may have to consider abandoning or significantly restructuring work on particular topics. Blended data methods also frequently require stepping away from one's position as a "researcher" and into participation, which often builds emotional lived experiences and connections (Reger *2001*). Attending meetings and protests, arguing with "participants" (or friends) late into the night about tactics, and sharing experiences of tension and threat give one a very different emotional relationship to your "data."

Similarly, even for researchers not directly involved in fieldwork, there is a different sense of emotional involvement and connection when you're not just collecting data on an event but rather collecting data with extended personal connections: this includes data specific to the fieldwork, such as a protest that your colleague (and, usually, friend) is attending, but also studies which are focused on a location with personal significance or writing about events, crises, and protests that other friends and colleagues have been affected by (apparent in such cases, for us personally, as the Boston Marathon bombing and London riots and natural disasters in Japan and Australia). In such examples, it is hard – or impossible – to become an objective voice studying relevant data; yet this lack of (emotional) distance from this setting may also allow for further (or different) questions and depth to the analysis, based on personal interest and familiarity, than might result if coming to the subject without prior contextual knowledge.

Claims to Objectivity and Accuracy Are Misleading

Big Data methods are often seen as more objective than qualitative methods. Kitchin and Lauriault (*2014b*, p. 4) argue that the pieces of information we have come to see as "data":

> are viewed as being benign, neutral, objective and non-ideological in essence, reflecting the world as it is subject to technical constraints ... the terms commonly used to detail how data are handled suggest benign technical processes: 'collected', 'entered', 'compiled', 'stored', 'processed' and 'mined.'

boyd and Crawford (*2012*, p. 667) note that this kind of analysis is often juxtaposed with a supposedly more subjective process of qualitative research: "there remains a mistaken belief that qualitative researchers are in the business of interpreting stories and quantitative researchers are in the business of producing facts." However, this division between "fact-based science" and "subjective humanities" is one that has been challenged on a variety of fronts, including by researchers within science and technology studies and technoscience studies.

"Data," including Big Data, is constructed through processes which are subjective and political. For example, Twitter's Data Grants program (Krikorian *2014*) simultaneously acknowledged and attempted to address the difficulty researchers face in using Twitter's public, historic data to tackle "big questions" and allowed Twitter to manage access to their data. (Twitter's Data Grants program ran once, in 2014: six projects were awarded grants, from 1300 submissions. At the time of revising this chapter in 2017, no repeat of the scheme had been announced.) "Data" might be the focus of research simply because it's available, which shapes the questions that we can ask about it (Vis *2013*),

and it is essential to recognize and address the limits of these datasets (especially through additional methods) rather than just accept them as "this is what (and all that) we can do" or to just do what is easiest and quickest: as Mahrt and Sharkow note on the availability bias of Big Data studies, "Rather than theoretically defining units of analysis and measurement strategies, researchers tend to use whatever data is available and then try to provide an ex-post justification or even theorization for its use" (*2013*, p. 25). Data is also co-produced through the tools available for data analysis, many of which are developed to meet commercial needs. And, of course, Big Data methods still require interpretation and framing: "We are on a mission to make a point about the data we made... we tell stories about the data and essentially they are the stories we wish to tell" (Vis *2013*). The comparative ease of access for gathering and analyzing data from Twitter has not just contributed to this platform's prominence within the existing literature, but the development of common methods and research designs has also led to extensive studies of similar topics. Elections and political communication on Twitter or the live-tweeting of television shows, sporting contests, and breaking news are well-represented by research. While this has clear benefits in affording comparative research between different types of event, audience, and national context, it is also important to consider the practices and discussions not featured here, including the everyday and the banal which might not, for instance, employ common hashtags or keywords to denote relevant content.

This includes practices that subvert or circumvent standard approaches and affordances and that use different mechanisms for sharing content: while quoting tweets via textual means is obviously visible in (automated) textual analysis of Twitter data, posting screenshots of tweets to present them in image form creates different methodological and analytical concerns. Given the importance of visual documentation of comments to controversial and sensitive subjects, recording evidence of tweets before they are deleted within contexts where, for instance, sexist remarks and rape threats are prevalent (Rentschler *2014*; Consalvo *2012*), it is important not to overlook these elements. Vis (*2013*) argues that, rather than being a "discarded data object" because of their unsuitability for existing analytical processes, new methodological innovation is required to include the visual – and the mixed media – aspect of social media activity within research (see also Highfield and Leaver *2016*). There are further practices which demonstrate active resistance and subversion of the norms and tropes of social media platforms and to avoid making explicit connections with specific users or topics (but where connections are implied through subtext), as outlined by Tufekci (*2014*), and using hashtags ironically or with deliberate misspellings or editorializing, to avoid the structural connection to wider, contested topics.

This is not to say that studying hashtags or networks of users is the wrong approach – there is a lot to be gained from this type of analysis, and of course we acknowledge that our own research has prominently featured these methods. Blended data offer one way to overcome various critiques and limits of Big Data – and online data-only – studies, to move beyond quick and easy analyses of social media data to explore context, motivation, and individual practices and behaviors within these settings. These methods have the potential to make resistant and subversive practices more visible. In our research, interviews, participant observation, and qualitative analysis of online practices highlighted a number of practices that might otherwise have remained hidden. These included the purposeful use of misdirection online as a response to ongoing surveillance and activists' exploration of platforms such as Pinterest which were beyond the purview of our quantitative data-gathering. Our methods also exposed us to prompts to consider our own positions, as activists took the opportunity to ask us (and sky in particular) about our politics and our relationships with movements and academia. Blended methods – and the interactions with other academics and "participants" that they frequently expose us to – offer the potential to highlight and revisit the subjective, constructed nature of Big Data research; they open the range of potential questions we might ask of a set of data and invite us to

consider more fully the ways in which we are making choices about the particular stories we are choosing to tell (or not to tell).

Bigger Data Are Not Always Better Data

Closely related to concerns about a false presentation of "objectivity," there is a temptation to think that more data will be better data: that an analysis of thousands of tweets, posts, videos, or tags will yield more meaningful and objective results than a smaller sample or qualitative material. Sampling issues are particularly of concern here: as boyd and Crawford (*2012*, p. 668) note, "Just because Big Data presents us with large quantities of data does not mean that methodological issues are no longer relevant. Understanding sample, for example, is more important now than ever." This is especially apparent when it comes to data drawn using automated processes through platform APIs or where visibility is partly determined by algorithms and/or choices which the researcher might not have been privy to (or conscious of making). Much Twitter research, for instance (our own included), uses the Streaming and Search APIs to locate relevant content based on hashtags, keywords, or other elements. However, these APIs are not providing *all* the relevant data: unless a researcher has access to Twitter's full data stream – something which is financially beyond many researchers without institutional funding – then their results are only a sample of Twitter activity, which they have not actively determined (for a full comparison of API results and rate-limiting, see Morstatter et al. *2013*; González-Bailón et al. *2014*). This becomes a particular methodological caveat for analysis of large-scale and trending topics, for the more popular a subject (in terms of volume of tweets), the more posts will be missed from automated data capture.

Similarly, the impact of platform policies and algorithms on user activity as well as research should not be overlooked. While features like Twitter's Trending Topics list or Facebook's News Feed are key elements of the user experience of these platforms, the mechanics behind determining what counts as trending or which content to display is not known to the user or to researchers – and the algorithms change, as do other architectural and aesthetic aspects (Bucher *2012*; Van Dijck and Poell *2013*; Ananny and Crawford *2016*; Duguay *forthcoming*). (Algorithmic decision-making and platform design has taken on increasing prominence following the 2016 US Presidential election, where the automated promotion of fake news on Facebook and questions of echo chambers, filter bubbles, manipulation of social media metrics and gaming algorithms, and algorithm-determined exposure to content and views took on popular and political dimensions (see Duguay *forthcoming*; for a full examination of media manipulation in the 2016 election, see Marwick and Caplan *2017*). Questions about algorithms, surfacing, and visibility have long-standing relevance to many platforms and have been invoked in various controversies and problematic behaviors (see, e.g., Massanari *2017*).)

Tarleton Gillespie (*2014*) has described "calculated publics," the result of user actions and decisions as determined by the algorithms of platforms: these processes may allow the identification of groups of users discussing a common topic, for example, yet as Gillespie notes "these algorithmically generated groups may overlap with, be an inexact approximation of, or have nothing whatsoever to do with the publics that the user sought out" (*2014*, p. 189). The publics presented by individuals and by algorithms are of course also shaped by practices, communities, and users whose presentation, identity, and activities are influenced by external factors beyond the online setting (and which are realized and challenged in various ways on social media), including race, gender, sexuality, class, and religion (among others); these are not necessarily obvious or apparent within large-scale, surface-level analyses. (What are also critical to note here, though, are the assumptions and biases within the algorithmic, machine learning, and platform development processes, which lead to situations where

digital media (intentionally or not) highlight structural inequalities around race or gender, for example (see Noble *forthcoming*).)

The impact of algorithms and black box processes on social media research sampling is twofold: first, what data is returned in searches and automated captures is often determined by factors beyond the researcher and second, algorithms influence what individual users are seeing and responding to on these platforms – yet this information may be lost to researchers looking at isolated posts, statuses, or tweets after the fact. It may not simply be a case of noting how social media data are representative (or not) of a given population but also of highlighting the multiple factors influencing what data are being studied. In the same way that Baym (*2013*) highlights benefits and shortcomings of various social media metrics – including the question of audiences and intentionality for user activity – we reiterate here the need to acknowledge how data has been sampled and to not treat all samples as shaped by the same factors.

A blended data model can do more to show us the gaps in our research: it can tell us which methods are "better," what we're missing, and where we might need to complement the existing data. As Mahrt and Scharkow suggest, despite issues around sampling with large-scale social media datasets, "the collection of Big Data can also serve as a first step in a study, which can be followed by analyses of sub-samples on a much smaller scale" (*2013*, p. 24). However, issues of sample selection do not disappear when using blended data. For example, social movement research often relies on a process of snowball sampling to gain contact with participants, which can lead to skewed demographics, and particularly to a focus on more accessible or visible participants (such as those who speak English or who are considered leaders within a movement). No single method is perfect, but blended data can both bring to light and partially address sampling biases.

Interviews and fieldwork carried out during *Mapping Movements* showed us gaps in our online data we might never have been aware of otherwise. In Oakland (Croeser and Highfield *2014*), participants emphasized the importance of organizers who were present on the ground but completely absent from Twitter (and therefore from our datasets). They also discussed strategic nonuse of Twitter and other platforms, as activists involved in Occupy tried to evade government surveillance. Similarly, our interviews in Greece (Croeser and Highfield *2015a*) showed that many activists modified their use of social media to avoid surveillance. They also eschewed the use of some online platforms because they were seen as out of alignment with activists' goals. We also note that the use of blended methods helped to address an additional gap in our research in the case of the Greek research: interviews not only brought to light important omissions in the data we were collecting they also helped to fill the linguistic and cultural gaps in our work, with participants translating ideas and concepts that we may otherwise have missed. For example, if we had only based our research on the online data available (and comprehensible) to us, we might not have understood the important distinction that many Greek anti-fascist activists make between the political left, autonomists, and anarchists. Finally, fieldwork during the 2013 World Social Forum provided a very different picture of the space of the forum than our online analysis did. While online analysis showed hashtagged commentary and contributions prominently featuring international organizations, movements, and independent media (including European Occupy sites and global justice groups), participant observation made it clear that the physical space was much more contested, with anarchist graffiti in particular spreading out across the days to reclaim the forum from more reformist perspectives. In each of these cases, limiting our research to Big Data methods would have left important gaps in our understanding.

While single-platform or hashtag-specific studies can be useful points of entry into studying the relationship between social movements and social media (or myriad other topics), since we completed our *Mapping Movements* fieldwork, there have been additional studies that combine multiple perspectives and sites of research to provide more rounded understandings of contemporary activism

and digital media. The "Beyond the hashtags" report by Freelon et al. (*2016*) drew on "40.8 million tweets, over 100,000 web links, and 40 interviews of BLM activists and allies" in its analysis of the Black Lives Matter movement. Here, while Black Lives Matter is a key example of a movement making extensive use of social media, to study social media content alone severely limits what is available for analysis, even when that is a dataset of more than 40 million tweets. Social media-only research may also restrict what can or cannot be said, whether due to the technical limitations of platforms or the publicness of the particular spaces chosen. As with our research, adding participant interviews to the study allowed Freelon et al. "to better understand [activists'] thoughts about how social media was and was not useful in their work" (p. 10). Similarly, Zeynep Tufekci's (*2017*) detailed examinations of protest and digital media draw on extended participant observation and interviews to offer insight into social movements and technology, from Egypt to Turkey to the USA.

Taken Out of Context, Big Data Loses Its Meaning

The caveats that accompany rich, big social datasets include acknowledging the loss of context and recognizing the artifice of the researcher's context imposed upon the research object. Considering a dataset organized around a specific hashtag, for example, brings together diverse comments and users which otherwise might have nothing in common and no awareness of others featured in the same space. Context is determined by the presence of the chosen hashtag, yet the reasons and meanings behind the hashtag's inclusion in a given tweet may be lost. For political subjects, this can mean losing important contextual information about an individual's own views on a given issue, which might not be apparent in a single tweet but outlined in their surrounding comments and supporting information. Identifying sarcasm, irony, and other ways of presenting views which might hide or obscure an individual's own views – when read in isolation – are further challenges here.

The many ways in which language can be used and adapted are particularly important to communication by different groups. Sharma, for instance, studied "Blacktags" – hashtags associated with Black Twitter that are "expressive of everyday racialized issues and concerns" (2013, p. 51) and make use of slang, humor, and other linguistic elements in their presentation; as part of their discussion, Sharma interrogates how such tags may become popular and widely shared (including receiving attention, positive or negative, from individuals and groups outside of those initially using the hashtag), as a result of connections between users and platform algorithms. The context of Black Twitter also raises important considerations about race, visibility, and researcher/analytical biases (among other questions), as noted by Brock (*2015*): while the marker of "Black Twitter" is used to describe social media practices and communities of, especially, Black Americans, the use of this descriptor is also been criticized for potentially ascribing a singular identity or practice to a diverse (and already marginalized) population (see also Brock *2012*; McElroy *2015*; Ramsey *2015*). Furthermore, as Freelon et al. note within the context of Black Lives Matter, "'Black Twitter' is a widely-discussed cultural phenomenon that overlaps with BLM but remains distinct from it" (p. 8). They also underline that "Black Lives Matter" itself can also represent, variously, the wider movement, the specific organization, and the social media coverage and campaign of #BlackLivesMatter; these have crossover but are not entirely synonymous – yet the nuance here could be lost when taken out of context. In these cases, and many others, then, there is a risk of Big Data analysis imposing homogeneity on diverse practices, overlooking individuals' intentions and contexts for their remarks. Decontextualizing social media data can also remove external and personal motivations behind comments, masking the political, social, cultural, temporal, and other factors which may lead an individual to tweet or post a particular remark.

Furthermore, there are diverse practices and interests apparent within datasets and discussions organized around a particular subject: single hashtag studies, for instance, use a specific marker to identify relevant content, but this does not mean that all comments featuring the hashtag are covering the same topic. In our analysis of Twitter around the Occupy Oakland movement, the #oo hashtag was used for multiple purposes and co-occurred with other hashtags denoting, variously, specific protests, other branches of the Occupy movement, meetings, and local authorities and events. Not all users tweeting about #oo used these other hashtags, though: the risk of homogeneity is apparent here, as for some users Occupy Oakland was actually a peripheral topic – since they were not in Oakland or discussing that branch specifically – while others were directly discussing it. Similarly, Burgess et al. (*2015*) describe the hashtag as a "hybrid forum," where diverse coverage and practices are apparent within an ostensibly overarching, common context. This diversity, though, is revealed and interrogated more fully through a blended methods approach than purely quantitative processes and backed up with further blended data by combining social media analysis with, for instance, interviews and surveys of participants. Of course, not all Big Data analysis is quantitative – and not all quantitative research features Big Data – and we are not saying that focusing on Big Data is wrong: by definition, this research offers huge potential because of the richness and scope of its datasets, providing extensive resources to research. Similarly, while the wealth of Twitter research provides rich detail into particular contexts and practices on this platform, Twitter is not a representative of "social media" or "online communication" overall. There are many options available here and no one solution or template for analysis.

A secondary form of blended data is suggested here, then: studying different platforms and apps, beyond the major, general spaces like Twitter and Facebook, and especially a focus on platform use in concert. Users employ different platforms for different purposes, communicating with networks of friends and followers that may overlap considerably but are not entirely the same, and information and content spread between platforms. To investigate this extended coverage and activity, Driscoll and Thorson (*2015*) outline approaches to identifying and analyzing topical content across platforms, based on links made to common content and thematic overlap in tweets and posts on Twitter, Facebook, and more, among other potential means of exploring social media activity beyond the single platform context. The issue mapping of sociocultural controversies by Burgess and Matamoros Fernández (*2016*), too, extends existing methodologies to undertake cross-platform tracking of topics and identify cultural dynamics at play within, but also across, popular platforms. By following particular media objects, actors, or themes across digital media, these approaches allow for greater awareness of the different cultures and practices on particular platforms and the involvement of the platforms themselves in shaping activity and discussion (see also Matamoros Fernández *2017*).

The risks of loss of meaning and context in big social datasets also extend to the various groups and communities – and isolated individuals – contributing around a given topic or event. This is not just apparent online, of course: within the *Mapping Movements* context and the study of activists and social justice, "social movements are inherently multifaceted, fluid, and messy" (Croeser *2015*, p. 77), drawing together diverse themes, groups, and perspectives. Implying singular topical relevance or adherence within a big dataset ignores this messiness. Blended data provides further context to the analysis: it does not provide *all* the contexts but at least addresses some of the imbalances apparent here.

Just because it Is Accessible Does Not Make it Ethical

Gathering social media data comes with a number of ethical challenges. Ethical issues with gathering social media data are often addressed in the context of highly politicized and risky topics, such as participation in social movements. Big Data methodology can alleviate some of the concerns associated with using such data by, for example, de-identifying content, although, as we note elsewhere at greater length (Croeser and Highfield *2015b*), ethical challenges remain in these cases. However, even collecting data on more quotidian topics can be ethically problematic. As Mahrt and Scharkow note, "basically all Big Data research is based on the assumption that users implicitly consent to the collection and analysis of their data by posting them online" (*2013*, p. 26), which is not an assumption we should take for granted. Rather than simply accepting the ways in which social media companies are resetting the boundaries of privacy to push the "radical transparency" that their business model relies on (Raynes-Goldie *2012*; Hoffman *2014*), researchers need to think carefully about what it means to turn people's posts about their lives, or their everyday and mundane conversations, into "data." Vis (*2013*) notes the links between the "synoptic view" of users that Big Data researchers, social media companies, and those developing data-collecting tools adopt in the production of "data." Similarly, Anita Chan argues in her response to boyd and Crawford's work that we need to be particularly alert to the dangers of collaborations between academics and corporations offering access to private data pools and to the ways in which this might encourage a view of users "less as physical, embodied 'subjects' with rights and obligations for protections, than as data opportunities for experimental extractions" (Chan *2015*, pp. 3–4). It is useful to take this as a prompt to explore some of the deeper ethical issues involved in the use of Big Data methods on social media. One of the first concerns related to the ethics of online research concerns the kinds of data it is ethical to access. As Dorothy Kim (*2014*) argues, academics (and in particular white academics) have often assumed that anything shared on Twitter – and other online spaces – is public, ignoring the "the contextual and historical background [necessary] to understand the long history of minorities and especially black men and women being used as data and experimental bodies for research and scientific experiments." Activists have criticized the appropriation of data, content, and analysis shared on social media. For example, Sydette Harry (*2014*) talks about the urgent need for a better understanding of the ways in which marginalized groups, and Black women in particular, have been subject to data-gathering which becomes a weapon and which they cannot direct – or even participate in – themselves. In a case that highlights some of the failures of academia to properly address these issues, Lauren Chief Elk criticized the unattributed use of data developed by the Save Wįyąbi project in a funded student project and the lack of an appropriate response by the California College of the Arts (Steinhauer *2014*; Chief Elk *2014*). Even as we write this, we are aware that there is a danger that in drawing on work by activists discussing the appropriation of their analysis, we are repeating the same dynamic. We have attempted to address this by foregrounding and properly attributing the activists and academics who have led the conversation around the appropriation of their data while avoiding informal communications that seem intended for a more limited audience.

Kim and Kim's (*2014*) #TwitterEthics Manifesto, developed through a broader conversation on Twitter, sets the minimum for Twitter research at "credit, citation, attribution" but say that additionally academics "should ask each individual user on Twitter for consent. They should explain the context and the usage of their tweets." This is clearly impractical for most Big Data research on Twitter and other platforms. There are, however, some ways in which researchers can, at the least, make their data-gathering more visible, including by tweeting with the relevant hashtag (or in some other way reaching out to the community they're gathering data from). This not only gives users the opportunity to avoid using a particular hashtag or taking part in the community which is the subject of data-gathering, it also opens the possibility of dialogue and of critique. This may be uncomfortable at times for researchers: this kind of discomfort can be productive in rethinking the ways in which

academia is structured and the relationship between researchers and those whose data they draw on. In addition, some of Kim and Kim's other recommendations around building dialogue, recognizing the role of activists (as well as, or sometimes rather than) academics as experts, and seeking to build more radical research models may be applicable (Kim and Kim *2014*). Many of these mesh well with research approaches that attempt to challenge existing power structures within academic research, including social movement research which positions activists as experts (Chesters *2012*); critical, indigenous, and anti-oppressive approaches to research (Brown and Strega *2005*); and intersectional approaches (Collins *1989*; Cho et al. *2013*; hooks *2000*). Integrating these approaches into Big Data methodology can be further strengthened by drawing on blended methods.

Blended methods offer opportunities to further deepen our engagement with the ethical challenges of Big Data research. They are a potential means for addressing questions of ethics more generally within Internet research, including platform ethics (what is allowed on and by individual platforms, from Twitter to Facebook to Instagram to Tumblr) and the blurring of personal, private, and public across multiple levels of "publicness" (see Highfield and Leaver *2015*). Such considerations should take into account the information volunteered by social media users both deliberately and inadvertently – the presentation not just of automatically generated geolocation and other metadata, for instance, but also the tagging and visibility of other people in posts and photos or the other potentially revealing elements depicted in images. As part of this, the use and nonuse of social media, and subversive and unorthodox practices, can be identified and examined through blended data approaches. Within our *Mapping Movements* research, interviews have in part examined why activists might not use particular platforms, including for reasons of surveillance – information not necessarily available from a Big Data-only analysis. Awareness of these processes, which are clearly attempts to protect certain data from particular audiences, needs to be taken into account when considering which data researchers gather and what we publish.

Reflections on the ethics of Big Data research, and on the ways in which different research methods might lead to more engagement with participants (including the possibility that participants will raise critiques of researchers' methods or analysis), must extend beyond institutional ethics processes. As boyd and Crawford (*2012*, p. 672) note, many ethics boards are not fully aware of the details of Big Data methodology or of the potential ramifications of Big Data. Internet research, and academia more generally, should also learn from important conversations being held among social media users about the ways in which privilege and surveillance structure the (in)visibility of different contents across platforms. This requires actively seeking out, listening to, and attributing the analysis of marginalized groups, including trans people, people of color, disabled people (We recognize this is a contentious term, and while we draw on disabled people's such as Stella Young's (*2012*) understanding of their disability as socially constructed, we recognize that some people may be uncomfortable with or reject this model.), and women.

Limited Access to Big Data Creates New Digital Divides

Questions of access to data and tools, and understanding of Big Data analysis, are not purely methodological: they are also political. As Kitchin and Lauriault note, "Databases and repositories are expressions of knowledge/power, shaping what questions can be asked, how they are asked, how they are answered, how the answers are deployed, and who can ask them" (*2014b*, pp. 4–5). The recent turn toward Big Data has implications not just for the kinds of research that we do but also for power structures within academia. Big Data methods – and the current enthusiasm for these methods – among other effects, exclude or discourage some groups from participation in research processes,

open spaces for others who might previously not have been involved in Internet studies, and shift funding distribution.

The politics of social media platforms (see also Gillespie *2010*) contribute to such a divide among researchers: access to large-scale data from Twitter, Instagram, or Facebook, for instance, is limited either by what is and is not permitted through the platforms' APIs or by researchers' funding to use data on sellers like Gnip. For many individual researchers, including postgraduate students, fractional and sessional staff, and adjuncts, such a financial outlay may well be unrealistic, especially for the longer-term study of users and practices beyond snapshots. Furthermore, the terms of use for Twitter which explicitly prevent the sharing of datasets mean that not only are researchers without formal affiliations and collaborations unable to use resources that might be gathered by colleagues but that replicating and reviewing datasets and analyses are also unlikely. As Vis (*2013*) notes, this impacts upon the validity of research: issues "arise in relation to the often opaque and unclear ways in which researchers themselves make and collect data for research purposes" and without a clear way for studies to reproduce datasets and replicate analyses.

Many researchers are aware of these issues; the interviews carried out by Weller and Kinder-Kurlanda (*2015*) highlight various concerns and inconsistencies around doing social media research, including questions of sharing data and independent testing of analyses. Similarly, Zimmer and Proferes (*2014*) have noted recurring questions around ethical research in this field and the inconsistent approaches to addressing such concerns within different studies and institutional contexts (from the collection of public or semipublic data to the presentation of information within published research). While these divides may appear neutral – all research is affected by platform policies, for instance – they are exacerbated by the structural inequalities within and outside of academia: this includes the varying support and resources available to permanent and fractional staff and the impact of precarity noted above but also reflects other gaps and divisions that may be widened by the privileging of particular types of research (and the potential to more quickly publish Big Data studies than those relying on interviews, fieldwork, surveys, focus groups, or other sources and methods beyond the online-only and the automated).

Given the continued underrepresentation of women, people of color, and other marginalized groups in science, technology, engineering, and math, it's also worthwhile considering the ways in which an uncritical valorization of Big Data approaches might widen gaps within academia. boyd and Crawford (*2012*, p. 674) note that women are underrepresented in computational research, which affects the kinds of questions posed and topics studied. There are echoes here, with our own experiences: sky began her tertiary education studying science and engineering and was put off by a multitude of barriers, including orientation events which were unsafe, creepy male tutors, and a lack of signals to counter imposter syndrome (these were also balanced by the ways in which being a white, middle-class cis woman offered forms of privilege).

If funding and attention shifts to Big Data approaches – and consequently away from researchers using other methodologies (a possibility that Sawyer noted on the horizon in *2008* and which Chan (*2015*, p. 2) notes has precedents in other areas) – which perspectives might this (further) marginalize? What might it mean for those who already find it hard to find a voice within academia, for those who cannot easily access the necessary tools or training for Big Data research, or for those who find their critiques of Big Data analysis undermined by assumptions that they are too "subjective"? The possibility that an enthusiasm for Big Data methodology might increase existing inequalities within (and beyond) academia does not necessitate abandoning these tools, but it does require deep critical reflection.

Conclusion

Our response to boyd and Crawford (*2012*), and to questions of Big Data in general, is intended to highlight the limits to Big Data analyses while recognizing the opportunities that they afford. We have outlined a way that we have attempted to combine various approaches to social media analysis through blended data, as a means to negate and overcome some of these concerns. However, this is not the only possible solution: our approach focuses on social movements in addressing the limits and pitfalls of social media research, including ethical considerations, focusing on social movements. Other contexts might warrant other approaches, such as the deeper data analyses suggested by Brock (*2015*), yet a blended model in general is one that may prove beneficial to researchers and their subjects.

In suggesting this model, though, we are not arguing that big/small data are the problem, that all Big Data research is reductive or misguided, or that looking solely at small data will suddenly remedy all problems. What is needed is *good* data, in whatever form, supported by methodologies and analyses that are robust, flexible, and complementary. The datasets featured in studies, however large, are not the research or its findings: researchers have a responsibility to ask *good* questions of their data, of what it depicts and what it does not, and of the contexts, motivations, and cultures apparent within them.

Using blended data does not, of course, completely address the many provocations raised around Big Data. Our research is open to many of the same concerns that boyd and Crawford highlight, including a positioning of some knowledge as more objective, gaps and elisions in our analysis, and an uneasy ethical relationship with "participants" and "data." In many senses, these problems are inherent in humanities research, particularly given the structural inequalities which underlie academia (and its broader social and political context). The questions and concerns raised and responded to here around Big Data analysis, and around Internet studies more generally, are not just questions of these contexts; in responding to the provocations by boyd and Crawford, then, and promoting blended and deeper analyses that take into account other perspectives, practices, and contexts, we are also attempting to improve research cultures and capabilities. There is no one right way of doing this, and not all approaches will be relevant or appropriate for all settings. However, being aware of these concerns and consciously attempting to address them, collectively and collaboratively promoting different methods, is a step toward doing research that is inclusive, accessible, open, and supportive, rather than what might just provide quick results or fit a currently trending paradigm.

References

Ananny M, Crawford K (2016) Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability. New Media Soc. https://doi.org/10.1177/1461444816676645

Baym NK (2013) Data not seen: the uses and shortcomings of social media metrics. First Monday 18(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4873/3752

boyd d, Crawford K (2012) Critical questions for Big Data. Info Comm Soc 15(5):662–679. https://doi.org/10.1080/1369118X.2012.678878

Brock A (2012) From the blackhand side: Twitter as a cultural conversation. J Broadcast Electron Media 56(4):529–549. https://doi.org/10.1080/08838151.2012.732147
CrossRef

Brock A (2015) Deeper data: a response to Boyd and Crawford. *Media*. Culture Society 37(7):1084–1088

Brown LA, Strega S (2005) Research as resistance critical, indigenous and anti-oppressive approaches. Canadian Scholars' Press, Toronto. Retrieved from http://site.ebrary.com/id/10191692

Bucher T (2012) Want to be on the top? Algorithmic power and the threat of invisibility on Facebook. New Media Soc 14(7):1164–1180
CrossRef

Burgess J, Galloway A, Sauter T (2015) Hashtag as hybrid forum: the case of #agchatoz. In: Rambukkana N (ed) Hashtag publics. Peter Lang, New York, NY, pp 61–76

Burgess J, Matamoros Fernández A (2016) Mapping sociocultural controversies across digital media platforms: one week of #gamergate on Twitter, YouTube, and Tumblr. Comm Res Pract 2(1):79–96
CrossRef

Chan A (2015) Big data interfaces and the problem of inclusion. Media Cult Soc 0163443715594106. https://doi.org/10.1177/0163443715594106

Chesters G (2012) Social movements and the ethics of knowledge production. Soc Mov Stud 11(2):145–160. https://doi.org/10.1080/14742837.2012.664894

Chief Elk L (2014) Teach-In Summer Fundraiser. YouCaring. Retrieved 3 June 2015, from http://www.youcaring.com/nonprofits/teach-in-summer-fundraiser/212206

Cho S, Crenshaw KW, McCall L (2013) Toward a field of intersectionality studies: theory, applications, and praxis. Signs 38(4):785–810. https://doi.org/10.1086/669608

Collins PH (1989) The social construction of black feminist thought. Signs 14(4):745–773

CrossRef

Consalvo M (2012) Confronting toxic gamer culture: a challenge for feminist game studies scholars. J Gender New Media Technol 1. Retrieved from http://adanewmedia.org/2012/11/issue1-consalvo/

Crawford K, Miltner K, Gray ML (2014) Critiquing Big Data: politics, ethics, epistemology. Int J Comm 8:1663–1672

Croeser S (2015) Global justice and the politics of information: the struggle over knowledge. Routledge, Hoboken, NJ

Croeser S, Highfield T (2014) Occupy Oakland and #oo: uses of Twitter within the occupy movement. First Monday 19(3). https://doi.org/10.5210/fm.v19i3.4827

Croeser S, Highfield T (2015a) Harbouring dissent: Greek independent and social media and the antifascist movement. Fibreculture 26:136–157

Croeser S, Highfield T (2015b) Mapping movements - social movement research and Big Data: critiques and alternatives. In: Langlois G, Redden J, Elmer G (eds) Compromised data: from social media to Big Data. Bloomsbury, pp 173–201

Driscoll K, Thorson K (2015) Searching and clustering methodologies: connecting political communication content across platforms. Ann Am Acad Pol Soc Sci 659(1):134–148. https://doi.org/10.1177/0002716215570570
CrossRef

Duguay S (forthcoming) Social media's breaking news: the logic of automation in Facebook trending topics and twitter moments

Freelon D, McIlwain CD, Clark MD (2016) Beyond the hashtags: #Ferguson, #Blacklivesmatter, and the online struggle for offline justice. Centre for Social Media & Social Impact, Washington, DC. http://cmsimpact.org/wp-content/uploads/2016/03/beyond_the_hashtags_2016.pdf

Gillespie T (2010) The politics of "platforms". New Media Soc 12(3):347–364
CrossRef

Gillespie T (2014) The relevance of algorithms. In: Gillespie T, Boczkowski PJ, Foot KA (eds) Media technologies: essays on communication, materiality, and society. The MIT Press, Cambridge, MA, pp 167–194

González-Bailón S, Wang N, Rivero A, Borge-Holthoefer J, Moreno Y (2014) Assessing the bias in samples of large online networks. Soc Networks 38:16–27. https://doi.org/10.1016/j.socnet.2014.01.004
CrossRef

Harry S (2014, October 6) Everyone Watches, Nobody Sees: How Black Women Disrupt Surveillance Theory. Model View Culture. Retrieved from https://modelviewculture.com/pieces/everyone-watches-nobody-sees-how-black-women-disrupt-surveillance-theory

Hesse-Biber S, Johnson RB (2013) Coming at things differently: future directions of possible engagement with mixed methods research. J Mixed Methods Res 7(2):103–109. https://doi.org/10.1177/1558689813483987

Highfield T, Leaver T (2015) A methodology for mapping Instagram hashtags. First Monday (1):20

Highfield T, Leaver T (2016) Instagrammatics and digital methods: studying visual social media, from selfies and GIFs to memes and emoji. Comm Res Pract 2(1):47–62
CrossRef

Hoffman AL (2014, June 30) Reckoning with a decade of breaking things. Model View Culture. Retrieved from https://modelviewculture.com/pieces/reckoning-with-a-decade-of-breaking-things

hooks b (2000) Feminist theory: from margin to Center. Pluto Press, London

Humphreys L, Gill P, Krishnamurthy B, Newbury E (2013) Historicizing new media: a content analysis of twitter. J Commun 63(3):413–431
CrossRef

Johnson RB, Onwuegbuzie AJ, Turner LA (2007) Toward a definition of mixed methods research. Journal of Mixed Methods Research 1(2):112–133. https://doi.org/10.1177/1558689806298224

Kim D (2014, October 7) Social media and academic surveillance: the ethics of digital bodies. Model View Culture. Retrieved from http://modelviewculture.com/pieces/social-media-and-academic-surveillance-the-ethics-of-digital-bodies

Kitchin R (2014) Big Data, new epistemologies and paradigm shifts. Big Data Soc 1(1):2053951714528481. https://doi.org/10.1177/2053951714528481

Kitchin R, Lauriault TP (2014a) Small Data, Data Infrastructures and Big Data (SSRN Scholarly Paper No. ID 2376148). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2376148

Kitchin R, Lauriault TP (2014b) Towards critical data studies: charting and unpacking data assemblages and their work (SSRN Scholarly Paper No. ID 2474112). Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2474112

Kim D, Kim E (2014, April 7) The #TwitterEthics manifesto. Model View Culture. Retrieved from https://modelviewculture.com/pieces/the-twitterethics-manifesto

Krikorian R (2014) Introducing Twitter Data Grants. Retrieved 4 Aug 2015, from https://blog.twitter.com/2014/introducing-twitter-data-grants

Langlois G, Elmer G (2013) The research politics of social media platforms. Culture machine (14)

Mahrt M, Scharkow M (2013) The value of Big Data in digital media research. J Broadcast Electron Media 57(1):20–33. https://doi.org/10.1080/08838151.2012.761700
CrossRef

Manovich L (2012) Trending: the promises and the challenges of Big Social Data. In: Gold MK (ed) Debates in the digital humanities. University of Minnesota Press, Minneapolis, pp 460–475
CrossRef

Marwick A, Caplan R (2017) Media manipulation and disinformation online. Data Soc, New York City. https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf

Massanari A (2017) #Gamergate and the Fappening: how Reddit's algorithm, governance, and culture support toxic technocultures. New Media Soc 19(3):329–346
CrossRef

Matamoros Fernández A (2017) Platformed racism: the mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. Info Comm Soc 20(6):930–946
CrossRef

McElroy K (2015) Gold medals, black twitter, and not-so-good hair: framing the gabby Douglas controversy. ISOJ 1(1). Retrieved from https://isojjournal.wordpress.com/2015/04/15/gold-medals-black-twitter-and-not-so-good-hair-framing-the-gabby-douglas-controversy/

Moe H (2010) Everyone a pamphleteer? Reconsidering comparisons of mediated public participation in the print age and the digital era. Media Cult Soc 32(4):691–700
CrossRef

Morstatter F, Pfeffer J, Liu H, Carley KM (2013) Is the sample good enough? Comparing data from Twitter's Streaming API with Twitter's Firehose. In: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, pp 400–408

Noble SU (forthcoming) Algorithms of oppression: how search engines reinforce racism. NYU Press, New York City

Papacharissi Z (2015) The unbearable lightness of information and the impossible gravitas of knowledge: Big Data and the makings of a digital orality. Media Cult Soc 0163443715594103. https://doi.org/10.1177/0163443715594103

Puschmann C, Burgess J (2014) Metaphors of Big Data. Int J Comm 8:1690–1709

Qiu JL (2015) Reflections on Big Data: "just because it is accessible does not make it ethical". Media Cult Soc 0163443715594104. https://doi.org/10.1177/0163443715594104

Ramsey DX (2015, April 10) The truth about Black Twitter. The Atlantic. http://www.theatlantic.com/technology/archive/2015/04/the-truth-about-black-twitter/390120/

Raynes-Goldie K (2012) Privacy in the age of facebook : discourse, architecture, consequences. Curtin Universitye. Retrieved from http://espace.library.curtin.edu.au/R?func=dbin-jump-full&local_base=gen01-era02&object_id=187731

Reger J (2001) Emotions, objectivity and voice: an analysis of a "failed" participant observation. Women's Stud Int Forum 24(5):605–616. https://doi.org/10.1016/S0277-5395(01)00190-X

Rentschler CA (2014) Rape culture and the feminist politics of social media. Girlhood Studies 7(1):65–82. https://doi.org/10.3167/ghs.2014.070106
CrossRef

Sawyer S (2008) Data wealth, data poverty, science and cyberinfrastructure. Prometheus 26(4):355–371. https://doi.org/10.1080/08109020802459348

Steinhauer J (2014, July 28) Native activist charges art students with plagiarism. Hyperallergic. Retrieved 3 June 2015, from http://hyperallergic.com/139769/native-activist-charges-art-students-with-plagiarism/

Tufekci Z (2014) Big Questions for social media Big Data: Representativeness, validity and other methodological pitfalls. In: ICWSM '14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media. Ann Arbor

Tufekci Z (2017) Twitter and tear gas: the power and fragility of networked protest. Yale University Press, New Haven, US

van Dijck J, Poell T (2013) Understanding social media logic. Media Comm 1(1):2–14
CrossRef

Vis F (2013) A critical reflection on Big Data: considering APIs, researchers and tools as data makers. First Monday 18(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4878

Weller K, Kinder-Kurlanda KE (2015) Uncovering the challenges in collection, sharing and documentation: The hidden data of social media research? Ninth International AAAI Conference on Web and Social Media

Young S (2012) I identify as a disabled person. Retrieved 27 Aug 2015, from http://www.mamamia.com.au/news/stella-young-i-identify-as-a-disabled-person/

Zimmer M, Proferes N (2014) A topology of Twitter research: disciplines, methods, and ethics. Aslib J Manag 66(3):250–261