Understanding the production and circulation of social media data: toward methodological principles and praxis

Susan Halford, Mark Weal, Ramine Tinati, Les Carr, Catherine Pope

Web Science Institute
University of Southampton

Draft: not for sharing or quotation.

1: Introduction

The phenomenal growth of online social media is a significant feature of the past decade. Social media are now firmly embedded across economies, cultures and political processes and in the lives of hundreds of millions of people, most famously a billion users signing on to Facebook in one day. It is not just the new forms of social practice associated with social media that is extraordinary - and their consequences, shaping social relationships, political regimes and business models, for instance - but that the very nature of these activities, as digital and online, also constitutes them as a remarkable new source of social data.

These data are generating widespread interest from business and government (Langlois et al 2015) but the response from researchers has been mixed. On one hand, social media offers new insights to the things that people say and do, in real time, and over time, at a scale and pace Weller et al 2013). Not least at a time when research funding is constrained (certainly in the US and UK) social media data appear as an unexpected gift, with rich potential to sustain social research. Indeed, recent years have seen significant take-up of these data across the humanities, social and computational sciences. On the other hand, there are concerns that social media data are problematic, that they are flawed by demographic biases and unknown provenance. The well-established principles of social research are grounded in clearly understood populations, carefully controlled sampling and well-known methods for collecting data. Social media data offer none of this. Accordingly, it is suggested, this may lead to poor research and unsustainable claims (Goldthorpe 2016; Hardaker 2016). At the same time, even amongst some enthusiasts – as our experience of working with social media data grows – there is rising awareness of the challenges in using these data for robust research (Weller et al 2014; Langlois et al 2105).

This paper seeks to trace a middle path in the space between *'giving in and getting out'* (Gehl 2015; 148). We are optimistic that social media data have something valuable to offer social research and also have concerns about the uses of social media data in this context. Our way forward is methodological. Working with conventional sources of data, professional standards demand that we make the details of our research design, methods of data collection and data management explicit. To date, this kind of transparency has not featured in research using social media data. In May 2016, we analysed the 115 papers with Twitter in the title or abstract published from 2013-2015, selected from the top 15 Social Science journals (57 papers) and 3 Computer Science social network journals (58 papers), ranked by H5 index. Of these, 90 papers contained Twitter data but few offered any methodological consideration of these data. Certainly this is challenging to do. The most popular social media platforms are privately owned and, make their data available, if at all, on their own terms and with differing levels of information. However, just because social media are a novel, and opaque, source of secondary data, is no less reason to consider these issues and their implications. To the contrary, there is all the more reason, if we are to allow social media data to be a credible and sustainable source for research.

In Section 2, we begin with our theoretical approach to data, drawing on Science and Technology Studies, long used to conceptualise data infrastructures (Bowker and Starr 1999; Bowker 2005) and influential in recent theorisations of the broader 'dispositifs' (Ruppert et al

2012) or 'assemblages' (Kitchin and Lauriault 2013) that produce new forms of digital data. Whilst social media data emerge from beyond the conventional practices of social science research, they are - despite the rhetoric sometimes deployed – anything but 'naturally occurring'. Our theoretical approach insists that data are constructed through the activities of heterogeneous actors, from data bases, interfaces and browsers to consumers, markets and legal regulations. Our specific focus in this paper is on how social media data are produced and made available to researchers: that is, the processes through which social media data are *produced and consumed*. In Section 3 we suggest that this has implications in three core methodological areas: the population, the sample and the methods that are used to create these data.

Our investigation is driven by a core philosophical research principle: we should understand the nature of our data, what we do *and do not* know about them, in order to be clear about the claims that we might make. This is not to insist on a gold standard for research data, whereby we must have full knowledge of data provenance, or to offer a finite description of data production and consumption for any particular social media service. Indeed, neither is possible partly because we cannot access all the information that would be necessary; and partly because social media services – and the data available from them – are highly dynamic. But it is to insist on methodological rigor. Section 4 suggests some key methodological principles for the use of social media data that might strengthen – and thereby protect – this new source of data for social research. Our conviction is that this will produce better academic research and will also develop our critical capacity to contribute to, and where necessary critique, the claims that are increasingly made from social media data by governments, the media and other commercial organizations.

2: What are Social Media Data?

As we enter the era of Big Data, Geoffrey Bowker's now emblematic statement that *'"raw data" is … an oxymoron'* (2005; 184) was never more apposite. As routine activities – from travel, shopping and energy consumption to web browsing and, of course, social communication – generate data of unprecedented volume, variety and velocity some dramatic claims have been made that these data constitute 'the new oil' (Humby 2014), a new natural resource that will, at last, reveal the mysteries of the social world (Anderson 2008, Watts 2011, Mayer-Schönberger 2013). Our starting point is that these data, like all data, are *'always already social'* (Bowker 2013; 168). Data do not exist 'in the wild' but are generated (Manovich 2001): they are produced rather than discovered, through a network of activities involving both human and non-human actors: social scientists and users, concepts and categories, survey tools, statistical measures, publication infrastructures, and so on (Hacking 2007). As Gitleman and Jackson (2013) argue, data are both framed – actively produced in specific contexts – and framing – themselves producing objects and subjects of knowledge.

This applies to any data (Scott 1998; Brine and Poovey 2013; Garvey 2013) but acquires particular significance in the current context, where researchers are increasingly drawing on new forms of digital data, not of our own making. In doing so, we make ourselves *'…reliant on platforms, methods, devices for data processing that have been developed in contexts and for purposes that are in many ways alien to those of social research'* (Marres and Weltervrede 2013; 13)

These data are generated beyond the orbit and control of researchers, used to producing their own data and/or working with carefully described secondary sources of data. Social media data are largely produced and owned by commercial companies, for whom the data are their only asset (Burgess and Bruns 2012): an asset that is carefully protected and if it is shared, usually with monetisation in mind. Reflecting on the implications of digital devices and their data for the social sciences Ruppert et al (2013) call for attention to

> *'… the lives and specificities of devices and data themselves, where and how they happen, who and what they are attached to and the relations they forge, how they get assembled, where they travel, their multiple arrangements and mobilizations and, of course their instabilities, durabilities and how they sometimes get disaggregated too'* (ibid; 31-32)

The production and circulation of social media data involves a heterogeneous network of actors. As with any network, there are many places that we could begin. Since we focus here on methodological questions about social media data, we start with 'pipeline' of data production and consumption. Conventionally used in Computer Science this metaphor describes the linear processes that shape the technical management of data (Patterson and Hennesey 1998). In what follows we corrupt this metaphor in two ways. First, we understand the processes shaping the evolution of data along the pipeline to be deeply social, political and economic, as well as technical. Second, we understand these processes as relational and dynamic, shaping each other iteratively and over time, rather than in fixed or necessarily linear ways.

-- Figure 1 about here --

Figure 1 provides an abstraction of actors in the 'pipeline' of social media data production and consumption: the subject who creates the content, posting to a social media platform, usually through client software on a phone, laptop, etc. that represents the data to the Application Programming Interface(s) (API), which enforces rules to determine what is passed through to the company's server software, and how, and the server software that organizes content into databases that store data in particular formats and structures. This is a thoroughly sociotechnical process, shaped by technical interfaces and protocols, data storage and software applications. And by popular culture, business models, organizational resources and so on. In turn, all this shapes if and how these data are shared with users – including researchers - back down the pipeline. The 'output' is *not* a simple reversal of the 'input' created by subjects, it is shaped by the methods that researchers use to access data, the economics and practicalities for the companies in sharing data, with whom and on what basis, both shaped by legal and ethical considerations.

3: Methodological Implications

In principle, the processes described in the data pipeline have profound methodological implications for use of social media data. In practice, we do not have access to all the information necessary to complete a full description of any given pipeline. Even if we did, the pace of change (in markets, technological infrastructures and social practices and so on) is rapid such that any detailed description is only ever a snapshot. However, this is no reason to abandon the search for a more robust methodological approach to social media data. In what follows we draw together the information that is available from social media companies with the published

experiences and experiments of a small group of researchers interested in this area, including ourselves, to generate a better understanding of the generic methodological issues that arise along the data pipeline. Given the caveats outlined above, our purpose here is to sensitise researchers to these issues, specifically as these affect methodological concerns in using social media data and the robustness of claims that might be made from social media data. We focus on three central elements of research methodology: population, sample and method. From this, we hope to inform stronger methodological practice in the field, improve the credibility of social media data and help to sustain its future in academic research.

3.1 Population

In broad terms, all social research begins with a scoping of the 'population' to be researched, a definition of our empirical subject. Most commonly, social researchers think of this in terms of people, their characteristics, values and actions. Indeed, the appeal of social media – for most – is that they offer insights into the everyday lives of their 'users', the subjects that post content online, situated at the far right of the data pipeline. We know already that the users of social media are skewed sub-sets of a global population. In a world where over half the population does not have access to the internet, it could not be otherwise. Even among the 3bn internet users worldwide, whilst the number of social media users is impressive, it is no basis for making claims about the total population. Surveys tell us, for example, that Twitter users are middle class[1], more women than men use Pinterest[2] and that 70% of WhatsApp users are under 45[3]. These surveys are important because not all social media platforms reveal demographic information to researchers (Facebook does not), or even collect demographic details (anonymity is a key feature of YikYak; WhatsApp accounts are defined by a telephone number not an individual). Where demographic information is available, it is not necessarily evident if or how this is related to offline characteristics. Different social media companies take different positions on this: Facebook famously barring those it deems 'inauthentic', Instagram and Twitter taking little interest in this.

Location is a case in point. Mapping social media data is hugely popular (Leetaru et al., 2013, Doré et al., 2015, Rodríguez-Amat and Brantne 2016) accelerated by mainstreaming of location-based functionalities from specialist social network services (e.g. Foursquare) into many of the big social network platforms (Evans and Saker 2017). However, there has been little attention to the production of geolocated data, which is shaped along the pipeline. Users may add their location manually (e.g. to their profiles). This information may be more or less accurate. For example, there is evidence that the Iranian diaspora on Twitter selected Tehran as their location during the Iranian elections, in a show of solidarity whilst political activists in Iran chose to hide their location at the time (Gaffney 2010). Alternatively, users may enable their client software to add location to the metadata attached to each post. This is likely to be a more accurate method but, experiments suggest that fewer than 2% of Twitter users select this option and – consequently – that fewer than 3%[4] of tweets contain any geotagged metadata (Leetaru et al.,

[1] http://digital-stats.blogspot.co.uk/2012/07/demographics-of-uk-twitter-users.html Accessed 05/10/16
[2] http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/ Accessed 05/10/16
[3] http://www.statista.com/statistics/290447/age-distribution-of-us-whatsapp-users/ Accessed 05/10/16
[4] Suggesting that the 2% using this function are more active than average, producing 3% of content.

2013). Furthermore, the volume of geotagged tweets has its own geography, with 2.86% in Jakarta compared to 0.77% in Moscow and taking into account that some users are particularly active contributors to the overall Twitter stream, Leetaru et al (2013) conclude that just *'… one percent of all users accounted for 66 percent of georeferenced tweets'* (n.p). We cannot assume that this small number of geotagged accounts is a representative of the wider Twitter population. The decision to disclose – or not – is likely to shaped by a variety of individual and social factors, biasing the geotagged population in significant ways.

Finally, much social media research assumes that each account belongs to a sovereign human individual. However, on many platforms (for example Instagram, Tumblr, Twitter and WhatsApp) there is no limit to the number of accounts that anyone can register, and data may include corporate group and parody accounts. Corporate accounts are distinctively crafted to represent particular organizational interests, rather than personal views, activities and identities. Parody accounts (Highfield 2016), for example the Dark Lord (@Lord_Voldemort) and Elizabeth Windsor (@Queen_UK) on Twitter post thousands of tweets to their respective followers (currently 2.04m and 1.3m), with retweets reverberating across the Twitter network. Whilst some parody accounts may be easily spotted from their qualitative features – not necessarily examined in quantitative analytics – others may be less obvious. Related to this, automated accounts make up an increasing proportion of activity and content on social media. These computationally controlled 'bots' may be new accounts created for the purpose, or 'hijacked' dormant accounts providing cover for bot activity and can be readily bought to add friends and followers, push posts across social media networks (e.g. liking or reposting), or aggregate content. Estimates of bot activity vary, from 4.6% on Sina Weibo (Zhang and Lu 2016) to an official Twitter estimate of 8.5% and an unofficial estimate of 13.5% (Chu et al 2010). In addition it is estimated that as many as 38% of all Twitter accounts may be 'cyborgs' enhanced by computational agents (for example) pushing out queued posts at regular intervals or optimum times of the day (Chu et al., (2010). Bots may or may not be identifiable depending, in part, on their behaviours along the data pipeline. Whilst the simplest way for bots to operate is to engage directly with the API (feeding activity in directly, rather than through client software) these may be relatively easily detectable and so delete-able. Bots that simulate interaction with the user interface are less readily identifiable. Meanwhile, social media companies themselves may also use bots to drive up activity and enhance use of the platform.

In short, we cannot and should not assume that the social media population is representative of the wider population. More than this, however, the constitution of social media population along the data pipeline shapes the nature of the data generated. Our brief survey of research using Twitter data shows that such issues are rarely considered. We found that only 9% of papers discussed the demographics of their data sample, whilst only 4% mentioned corporate accounts and 3% mentioned the presence of automated accounts. Whilst 25% were interested in geographically defined samples, there was little consideration of the mechanisms through which location is provided, their reliability or social selectivity.

3.2 Sample

Despite the often made claim that big data provides total populations, ending our reliance on samples, this is rarely the case for social media data (Highfield et al 2013). Whilst in principle, all activity is captured whole data sets are rarely shared fully with researchers. Some are given, or buy, full data sets but the vast majority are dependent on smaller samples of data, drawn either from their own web scraping or the use of a public Application Programming Interface (API), directly or indirectly through the services of a third party data broker. Figure 2 elaborates the basic pipeline representation, with particular reference to the 'output' back down the data pipeline. The mode of consumption chosen shapes the nature of the data that are derived and may, in turn, be significant for the research.

-- Figure 2 about here --

Web scraping uses automated agents (computational programmes) to process web pages and extract specific pieces of information e.g. the content of social media posts. This is advantageous where social media services provide no access to their data or if the researcher wishes is seeking for a different kind of data to the officially provided streams (although this may contravene company terms and conditions). Web scraped data has some distinctive characteristics. Most notably, the automated agents are searching pages listed by a browser – Google for instance – which has its own algorithmic processes for returning pages, not least shaped by data that can be accessed about the characteristics of the account making the request (the so called 'filter bubble' (Pariser 2012)). Any content received through web scraping is, thus, already sampled and, furthermore, web-scraped data can only include whatever information is available online to the browser, which may be different to data sourced directly from the companies. Although additional information may be inferred through web scraping – for instance, linking to other data on the web pages searched – this will require particular assumptions and inferences to be built into the computational processes.

Alternatively, data can be harvested directly from many social media companies. Whilst full data sets may be commercially available, sometimes public APIs offer data for free. How this sample is structured may have significant methodological implications. Let us take Twitter as an example. Launched in 2006, Twitter was initially open about sharing data, particularly encouraging developers to build applications to promote use of the platform. In turn researchers benefitted from this enhanced data access. However, as the company moves *'… from many possibilities to a narrower commercial entity'* (Burgess and Bruns 2015; 97) access has been progressively restricted and third party data brokers (who added their own functionalities to data streams) have largely been subsumed within the company (GNIP was bought by Twitter in 2014) or had their data access restricted (Datasift lost access to the full live Twitter data stream and to historical data in 2015). Whilst Twitter pursues a model of commercialisation for its most valuable data streams – the full 'firehose' of all tweets – access to other data streams is still available through the public API or, rather, two different APIs: the Streaming (live) API and Search API, with *'[e]ach offer[ing] a different set of methods for interacting with the system and each constrains the user in different ways'* (Driscoll and Walker 2014; 1748).

At present, the Streaming API provides real time data, in two ways: (i) a 1% sample of all tweets, 'pushed' through the API on a continuous basis. We do not know how this sample is generated but the company states it is random (perhaps a time-stamped sample[5]) and this is validated in experimental work (Morestatter, 2013; Wang et al 2015). This sample may be very useful for looking at 'what is happening on Twitter' but less so if the research aims to harvest data on a particular topic (most of which is unlikely to fall in to the 1% sample). (ii) Here, the Streaming *(filter)* API allows users to harvest real-time data for specific search terms. This is likely to return a far greater proportion of the tweets for a given search term (Gaffney and Puschmann 2015) but there is no guarantee that it will return all tweets for that search term, even if these constitute less than 1% of the firehose. Twitter also offers a Search API, *'… to search against a sample of Tweets published in the previous 7 days'* (https://dev.twitter.com/rest/public/search) (until recently this was 14 days). The sample received is *'focussed on relevance not completeness'* (ibid) but we do not know how this is sampled. Experiments suggest that the Search API returns far fewer tweets than Streaming (filter) API, at a ratio of approximately 1:4 (Gonzalez-Bailon et al 2012), whilst our own experiments confirmed this and showed far fewer retweets in the historic data sets[6]. Furthermore, Driscoll and Walker's (2014) suggest that Search API data are skewed heavily towards central users and more clustered regions of the network. Meanwhile, the amount of data received may also be shaped by the client software (see discussion at https://news.ycombinator.com/item?id=4795052 Accessed 05/10/16).

In addition, Twitter and many other social media platforms impose rate limits on the number of calls that can be made to an API during a given timeframe. This may be driven by practical limitations e.g. to load manage the network to try and keep a consistent service, or it could be a business decision to stratify the service offered (pay more get more), or it might be the result of decisions – ethical or otherwise – about what to make public. For example, on Twitter, each account can only make 180 queries to the Search API every 15 minutes (as of August 2016). On Facebook, each user can make 200 calls per hour, whilst the Instagram API (which is not public but available at a cost) allows 500 calls per access token in a one hour sliding window. Whether this matters depends on the nature of the data being queried. Small, regular data streams may not be affected at all by rate limits, whereas large data streams will be incomplete and clustered in the first part of time windows. In either case, it will not be clear what percentage of the potential data available has been returned, unless there is also robust data supplied on the total number of messages sent during a given timeframe. Note too that web scraping methods are subject to a different form of rate limiting as there are restrictions on how many times a HTTP GET request can be called (the technical method used to retrieve Web data) before the server denies the request. More generally, the capacity of the social media company's servers at any given point in time may impact on the amount of data delivered, whilst the geographical location of servers may affect the nature of data (if, for example Safe Harbour arrangements don't exist between countries it may not be possible to deliver personal data on individuals.

[5] http://blog.falcondai.com/2013/06/666-and-how-twitter-samples-tweets-in.html Accessed 05/10/16

[6] We collected data from both the Search API and also data (at 1%, 10% and 100% of the Twitter Firehose) from an official Twitter data reseller; both queries were based on a specific set of hashtags, during the same time period. We found that the Search API contained significantly less 'Retweet' statuses compared to the data obtained from the official data provider, similar to the figures stated by Gonzales-Bailon et al. (2012).

These technical descriptions are not intended to provide a technical 'manual'. Rather, our intention is to illustrate what APIs *do* (Busher 2013) to support our claims about how they shape the data that are collected, what can be harvested and how, in turn, this shapes the research that can be done and the claims that can be made. Furthermore, it is important to note that APIs are only temporarily stabilised. They evolve over time as new functionalities are added to the platform, as additions are made to underlying data models, or as political decisions are made to change access models, for example with regard to the functionalities and rate limits of particular APIs. This has particular implications for research that seeks to replicate previous studies or to take a longitudinal perspective. However, returning to our literature review, we found that 23% of papers offered no description at all of how the data were harvested, 46% stated that they used the Twitter API, but only 43% of these explained which API specifically, and of these few considered the implications of this for their data or findings. A further 24% of papers used data from third parties, including publically available data sets and data broking services, and 45% used web crawling methods but none included significant discussions of the implications for data sampling or, in turn for research findings[7].

3.3 Method

Social research has a rich repertoire of methods through which to 'capture' data including, for example, questionnaires and interview schedules. Similarly, social media platforms are not a mirror of social life 'out there' but designed artefacts that record particular types of information, and not others. The data generated by social media present the world for us according to designed features of these platforms - posts, comments, 'friends' or 'likes' - and the emergence of associated cultural practices and norms of sociality. These are not unconnected to the social but nor do they simply reflect an independent sociality. Social media functionalities show significant convergence across platforms over time, for example the major platforms operate with a version of profile, timeline, followers/friends, likes/favourites and location. Using social media data, we come to know the social through these features, retrofitting meaning to functionality. Although, of course what particular actions mean is far from evident.

For example, 'like' is a common feature but motivations for and meanings of the action are not conveyed by the vocabulary of the interface. Indeed, Meier, Elswiler and Wilson (2014) identify 25 different uses, ranging from indication that an item is topically relevant, acknowledging a family member, bookmarking, agreement with a statement, accident, or trying to engage others. Furthermore, how the 'like' is added to the database may be significant, for example, whether we click a 'widget' on an online news source or a shopping website ('like us on Facebook') or whether this is a like in the user's own client device application. In sociological terms these may indicate different things, but the data generated rarely distinguishes between them.

Similarly, the number of account 'followers' may be used to indicate popularity and/or influence, but the meaning of 'following' is complicated. For a start, the more followers an account has, the more likely it is that these are bots, since bots are often programmed to follow accounts with greatest popularity and/or influence (Chu et al 2012). Bots can be passive followers, significant if the focus is social influence, or more actively push information across networks, important if the

[7] NB. Some papers used more than one method of data harvesting.

focus is information diffusion. Meanwhile, even human followers aren't guaranteed to actually read content posted to timelines, despite social media companies' investment in metrics to encourage users with (rather vague) information about 'impressions or 'engagements'[8] . Moreover not all user activity or information flow is captured by formal metrics for example, users search and follow hashtags and keywords and content gets shared through alternative channels.

Furthermore, functionalities, and our use of them, change over time. As users, we adjust our practices to social media platforms – doing things we may never have done before *and* over time new practices may emerge, only possible because the platform is there but that were never envisaged by the designers. Much Twitter research has focussed on the 'retweet' function – to explore information flows and network formations. But prior to 2009 there was no formal retweet functionality, and until 2012 the Streaming API did not deliver retweets made with the retweet button because they were not identifiable in Twitter's internal data structure (Bruns and Stieglitz 2012). At the same time, if the focus is information flows, there are plenty of other ways to pass on information. Indeed, boyd et al (2010) suggest that 'dark retweets' (re-posts made not using the formal RT convention or retweet button) may account for up to 40% of total re-posts and that these are domain specific, so our knowledge of information flows in particular parts of the network may be especially limited. The underlying data model and counting mechanisms may treat these actions as different, but whether this is so, and the nature of their significance is far from clear.

Finally, the methods and design decisions of data management, through the organisation and configuration of servers and databases, may have significant effect on what data is returned to queries, shaping the types of analysis that can be performed. Whilst data may be received as unstructured streams of user generated content, engineers decide how this is stored and managed, with consequences for how it can be searched and is delivered in response to queries. Since Twitter and Facebook first launched, their API, the richness and structure of the data made available has changed considerably. Twitter's data structure did not originally contain geolocation, retweet, or hashtag data, and has only recently incorporated the 'like' feature. Moreover, fields which have been present since Twitter first launched have also changed; the 'created_at' and 'text' field has gone through several iterations, with changes in format and markup. These changing schemas and data formats makes it very difficult to assure consistency in data derived at different points in time, important if the aim is to replicate experiments or conduct longitudinal analysis.

Overall, social media data are produced through specific methods and metrics of data collection and circulation. As Marres and Weltervrede (2013) have argued in a broader context, new forms of digital data *'…tend to come with external analytics already built in'* (p.313), which requires reflection if we are to make the most of these data (see also Heer and Verdegem 2015; Marres and Gerlitz 2015). This was not considered in the papers that we reviewed.

[8] https://unionmetrics.zendesk.com/hc/en-us/articles/201201636-What-do-you-mean-by-Twitter-reach-exposure-and-impressions- Accessed 05/10/16

4: Discussion and Conclusions

The previous section highlighted some of the key methodological challenges that arise when working with social media data, as these appear along the pipeline of data construction and circulation. For ease of explanation we have presented these in a linear way, beginning with the subjects on the far right and ending with the databases at the far left of Figure 1. So it is important to emphasise here that the processes along the pipeline are iterative: for example, changes in the client device may impact on what users (can) do, changes in storage may impact on how the API can be searched and – at the heart of data construction and circulation – changes to the API may impact along the pipeline in both directions, perhaps with rebounding effects, for example as researchers turn to web scraping methods or new kinds of widgets are developed. Relatedly, we should mention that ethical challenges arise iteratively along the pipeline. For example, how should we treat personal data that users post on public pages? What data should social media companies release? What implications do data structure and format have for personal data linkage? And so on. In the papers we reviewed that included data, only three reported on completing an ethical review process and one other explained why ethical review had not been necessary. There is, finally, also a set of related issues regarding computational analytics tools that raise similar questions, as social scientists import black boxed methods to work with social media data (Author 2016). Whilst it is beyond the scope of this paper to review and explore this latter issue, it is quite clear – as any good researcher would expect – that the tools chosen have consequences for the analytical outcomes.

Taken together, this problematization of social media data may appear only to underscore the concerns expressed by those who have doubted their promise for robust social scientific research. This is not our intention. To the contrary, our tactic is to suggest that those of us using social media data should seek to address these challenges in our research. Certainly, we must accept that social media data are not like earlier generations of data, and consequently that the exact same methodological frameworks will not be appropriate. However, we should seek to position this new form of data methodologically, and develop new frameworks that will ensure its future value for researchers. In this paper we have suggested that attention to the 'data pipeline' offers one approach to this, drawing our attention to how the production and circulation of data shape population, sample and the method of data collection. This is summarised in Table 1 below.

-- Table 1 about here --

Taken together our literature review and our own experience in the field of social media research suggest that the issues raised in this paper are rarely considered. Whilst we are critical of this, we are also sympathetic. Social media are a new source of data and none of us was, from the beginning, an expert. To use a Norwegian expression we are all 'paving the road as we walk'[9]. Certainly, we recognise the omissions in our own work, as well as the work of others.

Looking forward, we suggest three key steps towards a more robust methodological approach, based on familiar principles. First is transparency, basic diligence in reporting how data are

[9] 'Veien blir til mens du går'.

harvested, when, using which data streams and what search terms. In addition, as we would with any other method, we should record key metrics of the resultant data streams, including size and any other notable characteristics. These details matter in a number of ways. Most obviously, they underpin comparative and longitudinal research and the possibility of reproducibility. If the intention is to pursue such research, then we need to know whether we are comparing like with like, and if not what the differences are as well as if and how they might be significant. The API changes are a particular issue here, since these may have a significant effect on the data that are returned even to identical queries. Whilst we will not always know what these changes have been, we will need to investigate these possibilities and/or – perhaps more likely - caveat the claims that we make accordingly. For example, if claims are being made with regards to a particular temporal phenomenon occurring, e.g. the speed to which a comment is spread across the network, we should bear in mind where temporal representations have changed. This might include being aware that if the timestamp format and time zone record has changed in the Twitter data structure since the API was first made public, this means that comparisons between data sets harvested at different points in time may yield inaccurate and misleading results.

Second, it is important that we consider if and how data construction might matter for the particular research questions under consideration. Some of the issues that we have highlighted above will matter a great deal for some research questions, and not at all for others. For example, the presence of bots may not matter at all if the aim is to explore information diffusion across a social media network but may matter a great deal if claims are made about human forms of influence in these networks. Claims about temporal patterns of social media activity should bear in mind the potential presence of cyborg accounts, whilst geographical questions and mapping methodologies will need to consider the very low proportion of geotagged data and the potential biases of those who enable this. These considerations in turn may moderate the kinds of questions that are asked and the claims that can be made. Take the vexed issue of demographics for example. One response to the well-known biases in Twitter data has been to 'convert' these data to more conventional social science data, for example by developing methods to make demographic biases explicit and to create demographically representative sub-samples of data (Sloan et al 2016). Alternatively, it has been suggested that particular demographic biases might be harnessed to explore populations that are under-represented in other sources of research data, young men in epidemiological research, for instance. We have to be clear about this and not infer claims from a social media data set to the general population without careful methodological controls or infer claims about all users of a particular social media platform from a sub-sample of data unless similar steps to match the sample with the wider population can be taken. However, we should also recognise that *a priori* demographic categories (sex, ethnicity, social class or age, for instance) may not the most important variables for working with these data, for example, we might be interested in the ebb and flow of public debate over time, or between different types of social media account (corporate news accounts, political parties and individuals, for instance). Or we might want to see how emergent social networks produce collectivities based on online activities, rather than reflecting external demographic characteristics. In short, the approach depends on the question that is being asked and the claims that we want to make.

Third, and finally, we must consider what these data are, what they can tell us and what they cannot. We have already suggested that functionalities cannot be conflated with human meaning or relationships: likes, re-posts, friends – may be indicative but are, in the end, designed functionalities of commercial data companies. To describe the patterns that they produce it may be more appropriate to refer to activity, information flows and networks, without making claims about their social significance. Making social claims about social media data will be more robust if we draw on 'wide data' – that is, multiple sources of digital data – and, as some other social media data researchers increasingly coming to conclude (e.g. Freelon and Karpf 2015; Hall et al 2016), if we employ mixed methods, to include offline as well as online data, qualitative as well as quantitative information. Indeed, it may be in this assemblage of multiple data sources, harnessed by theoretical understanding and methodological clarity that social media data find their most powerful contribution for understanding the social world.

FIGURES AND TABLES

Figure 1: Basic Pipeline of Social Media Data Production and Circulation

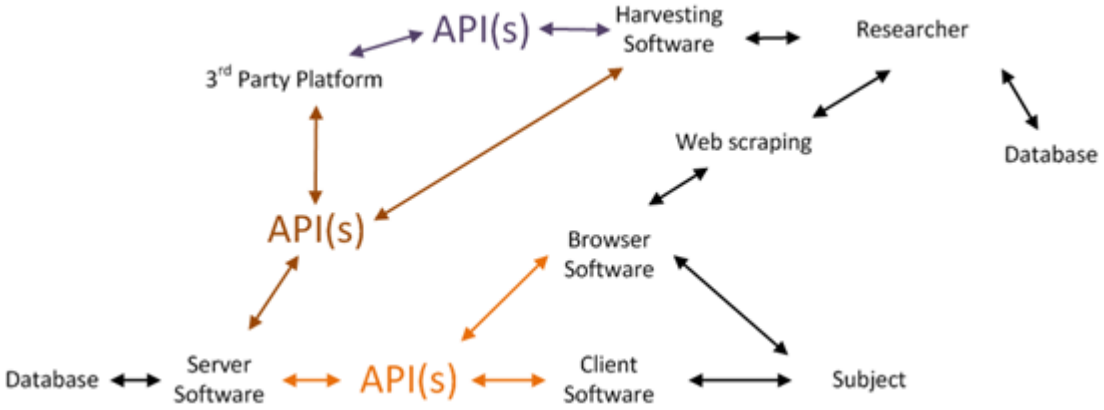

Figure 2: The Research Data Pipeline

	Methodological Considerations		
	Population	*Sample*	*Method of data production*
Database	Storage design and method shapes the types of information recorded about users.	Historic data storage decisions and technical query limitations may shape what data are included in samples.	Considerations of cost, performance and business requirements for data storage may shape what data are collected and stored and how.
Server Software	Determines who or what has access to the service, and what information is required to set up an account.	Server capacity may restrict data volume delivered; geographical location of server may affect data delivered.	Operates data management (e.g. spam removal and moderation, load balancing) shaping what data are collected.
API	APIs may not recognise all characters (languages) effectively; or be available to all operating systems/software development toolkits	A variety of differently structured samples may be available.	Defines the scope and volume of what data can be collected, stored and queried.
Harvesting Method	Harvesting methods construct different views of the populations. Web scraping may be more likely to access the population of currently active users, which could be different to the population accessed via historical searches using an API.	Web scraping will by-pass 'official' data samples, offering data from a sample of web pages. This sample may be affected by the 'filter bubble' of the person accessing the web pages. Use of third party data may introduce additional sampling effects.	Different harvesting methods have access to different types of data about the population and sample.
Client Software	Different clients may generate different information about the population. On some platforms you may know what client generated the content (this used to be the case on Twitter), on many though you can't know this.	Some clients (apps) may receive more data than others (if harvesting through a client).	Different clients may produce distinctive forms of data and metadata e.g. some may add geographic data by default, some might link directly to shared or reshared material.
Subject	Different subjects – human/non-human, demographically distinct – may characterise particular platform populations.	User activities may shape sampling methods (e.g. official samples may focus on central or highly active users.)	User practices and meanings shape the data generated and the claims that can be made from these.

Table 1: Methodological Challenges along the Data Pipeline

REFERENCES

Authors 2014
Authors 2016
Anderson C. (2008) 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete' *Wired* 23/06/08 http://www.wired.com/2008/06/pb-theory/
Bowker J. and Starr S.L-, (1999) *Sorting Things Out: Classification and its' consequences* Cambridge, MA., MIT Press.
Bowker G. (2005) Memory Practices in the Sciences Cambridge, MA., MIT Press.
Bowker G. (2013) 'Data Flakes: an afterword to raw data is an oxymoron' in Gitelman, L. (Ed) *"Raw Data" is an Oxymoron* Cambridge, MA., MIT Press.
boyd D, Golder S, and Lotan G (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. *Hawaii International Conference on System Sciences*, pp.1--10, Los Alamitos, CA, IEEE Computer Society.
Brine K and Poovey M (2013) 'From measuring desire to quantifying expectations: a late nineteenth century effort to marry economic theory and data' in Gitelman, L. (Ed) *"Raw Data" is an Oxymoron* Cambridge, MA., MIT Press.
Bruns A and Stieglitz S (2012) 'Towards a more systematic Twitter analysis: metrics for tweeting activities' *International Journal of Social Research Methodology* 16(2): 91-108.
Burgess J and Bruns A (2012) 'Twitter archives and the challenges of big social data for media and communication research' M/C 15 (5) http://journal.media-culture.org.au/index.php/mcjournal/article/view/561
Burgess J and Bruns A (2016) 'Easy Data, Hard Data: the politics and pragmatics of Twitter research after the computational turn' in Langlois. G., Redden, J., and Elmer, G. (2015) *Compromised Data: from social media to big data* London, Bloomsbury.
Burnap P, Avis N and Rana O (2013) 'Making sense of self-reported socially significant data using computational methods' *International Journal of Social Research Methodology* 16, 3: 215-230.
Busher T (2013) 'Objects of Intense Feeling: The case of the Twitter API' Computational Culture http://computationalculture.net/article/objects-of-intense-feeling-the-case-of-the-twitter-api
Chu Z, Gianvecchio S, Wang H and Jajodia S (2010) 'Who is tweeting on Twitter: human, bot, or cyborg?' *Proceedings of the 26th Annual Computer Security Applications Conference*. Austin, Texas: ACM.
Chu Z, Widjaja I and Wang H (2012) 'Detecting social spam campaigns on Twitter' *International Conference on Applied Cryptography and Network Security*. Springer Berlin Heidelberg.
Davis Jr, C Pappa, G De Oliveira, D & Del Arcanjo F (2011) 'Inferring the Location of Twitter Messages Based on User Relationships'. *Transactions in GIS* 15: 735–751.
Doré B, Ort L, Braverman O and Ochsner K (2015) 'Sadness Shifts to Anxiety Over Time and Distance From the National Tragedy in Newtown, Connecticut' *Psychological Science* 27, 4: 363–373
Driscoll K and Walker S (2014) 'Working within a Black Box: transparency in the collection and production of big Twitter data' *International Journal of Communication* 8:. 1745-1764.
Elmer G (2016) 'Scraping the first person' in Langlois. G., Redden, J., and Elmer, G. (2015) *Compromised Data: from social media to big data* London, Bloomsbury.
Evans L and Saker M (2017) *Location-Based Social Media: Space, Time and Identity* Basingstoke, Palgrave Macmillan.
Freelon D and Karpf D (2015) 'Of big birds and bayonets: hybrid Twitter in the 2012 Presidential debates' *Information, Communication and Society* 18, 4: 390-406.
Gaffney D (2010) ' #iranElection: quantifying online activism'. *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, April 26-27th, 2010, Raleigh, NC: US.

Garvey E (2013) "'facts and FACTS": abolitionists database innovations' in Gitelman, L. (Ed) *"Raw Data" is an Oxymoron* Cambridge, MA., MIT Press.
Gehl R (2015)'Critical reverse engineering' in Langlois G, Redden J and Elmer G (Eds) C*ompromised Data: from social media to big data'* London, Bloomsbury.
Gitleman L and Jackson V (2013) 'Introduction' in Gitelman, L. (Ed) *"Raw Data" is an Oxymoron* Cambridge, MA., MIT Press.
Goldthorpe J (2016), *Sociology as a Population Science*, Cambridge, Cambridge UP
Gonzalez-Bailon S, Wang N, Rivero A, Borge-Holthoefer J and Moreno Y (2012) 'Accessing the Bias of Samples in Large Online Networks' *Social Networks* 38: 16–27
Hacking I (2006) *Kinds of People, Moving Targets* 10th British Academy Lecture, http://nurs7009philosophyofinquiry.weebly.com/uploads/6/0/4/0/6040397/hacking_20071.pdf Accessed 2 August 2016
Hall M, Mazarakis A, Peters I, Chorley M, Caton S, Mai J-E and Strohmaier M (2016) 'Following User Pathways: Cross Platform and Mixed Methods Analysis in Social Media Studies' *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '16). ACM, New York, NY, USA, 3400-3407.
Hardaker C (2016) 'Misogyny, machines, and the media, or: how science should not be reported' http://wp.lancs.ac.uk/drclaireh/2016/05/27/misogyny-machines-and-the-media-or-how-science-should-not-be-reported/ Accessed 2 August 2016
Heer E and Verdegem P (2015) 'What social media mean for audience studies: a multidimensional investigation of Twitter use during a current affairs TV programme' *Information, Communication and Society* 18, 2: 221-234.
Highfield T (2016) 'News via Voldemort: Parody accounts in topical discussions' *New Media and Society* 18(9): 2028—2045
Highfield T, Harrington S and Bruns A (2013) 'Twitter as a technology for audiencing and fandom' *Information, Communication and Society* 16, 3: 315-339.
Humby (2014) https://www.marketingweek.com/2015/11/19/dunnhumby-founder-clive-humby-customer-insights-should-be-based-on-passions-as-well-as-purchases/ Accessed 5/09/16
Kitchin R and Lauriault T (2013) Towards Critical Data Studies: Charting and unpacking data assemblages and their work' *The Programmable City Working Paper 2*, University of Ireland Maynooth.
Langlois G, Redden J and Elmer G (2015) (Eds) C*ompromised Data: from social media to big data'* London, Bloomsbury.
Latour B (2007) 'Beware, your imagination leaves digital traces' *Times Higher Literary Supplement*, (April).
Leetaru K, Wang S, Cau G, Padmanabhan A and Shook, E (2013) 'Mapping the global Twitter hearbeat: The geography of Twitter' *First Monday* 16, 5: April 2013.
Manovich L (2001) *Software Takes Command* London, Bloomsbury.
Marres N and Weltevrade E (2013) 'Scraping the Social? Issues in live social research' *Journal of Cultural Economy* 6,3: 313-335.
Marres, N and Gerlitz C (2015) 'Renegotiating relations between digital social research, STS and the sociology of innovation' *Sociological Review* 64, 1: 24-46.
Mayer-Schönberger V and Cukier K (2013) *Big Data: a revolution that will transform how we live, work and think* London, John Murray.
Meier F, Elsweiler D and Wilson, M (2014) '+More than Liking and Bookmarking? Towards Understanding Twitter Favouriting Behaviour' *International AAAI Conference on Web and Social Media*, North America, May. 2014.
Morstatter F, Pfeffer J, Liu,H and Carley K (2013) Is the sample good enough? Comparing data from Twitter's streaming API with Twitter Firehose. In Proceedings of the 7th International Conference on Weblog fs and Social Media (ICWSM'13). The AAAI Press
Pariser E (2012) *The Filter Bubble: what the internet is hiding from you* New York, Penguin.

Patterson D and Hennesey J Computer Organization and Design: The Hardware/Software Interface, Second Edition, San Francisco, CA: Morgan Kaufman (1998).
Rodríguez-Amat J and Brantne C 2016 'Space and place matters: A tool for the analysis of geolocated and mapped protests'*New Media & Society June 2016 18: 1027-1046*
Ruppert E, Law J and Savage M (2012) 'Reassembling Social Science Methods: the challenge of digital devices' *Theory, Culture and Society* 30, 4: 22-46.
Savage M and Burrows R (2007) 'The coming crisis of empirical sociology' *Sociology* 41, 5: 885-899
Scott J (1998) *Seeing Like a State* Yale University Press
Sloan L, Morgan J, Burnap P and Williams M (2015) 'Who Tweets? Deriving demographic information from Twitter' in *Meta-Data*. PLoS ONE 10(3): e0115545.
Wang Y, Callan J and Zheng B (2015) 'Should we use the Sample? Analysing Datasets sampled from Twitter's Stream API' *ACM Transactions on the Web* 9,3.
Watts D (2011) *Everything is Obvious: how common sense fails* London, Atlantic Books
Weller K, Bruns A, Burgess J, Mahrt M and Puschmann C (Eds) (2013) *Twitter and Society* New York, Peter Lang.
Zhao L, Lu Y and Gupta S (2012) 'Disclosure intention of location-related information in location-based social network services' *International Journal of Electronic Commerce* 16, 4: 53–90
Zhang Y and Lu J (2016) 'Discover millions of fake followers in Weibo' *Social Network Analysis and Mining* 6, 1: 1-15