Qual Quant
DOI 10.1007/s11135-015-0162-8

Does 'bigger' mean 'better'? Pitfalls and shortcuts associated with big data for social research

Paolo Giardullo

Abstract 'Big data is here to stay.' This key statement has a double value: is an assumption as well as the reason why a theoretical reflection is needed. Furthermore, Big data is something that is gaining visibility and success in social sciences even, overcoming the division between humanities and computer sciences. In this contribution some considerations on the presence and the certain persistence of Big data as a socio-technical assemblage will be outlined. Therefore, the intriguing opportunities for social research linked to such interaction between practices and technological development will be developed. However, despite a promissory rhetoric, fostered by several scholars since the birth of Big data as a labelled concept, some risks are just around the corner. The claims for the methodological power of bigger and bigger datasets, as well as increasing speed in analysis and data collection, are creating a real hype in social research. Peculiar attention is needed in order to avoid some pitfalls. These risks will be analysed for what concerns the validity of the research results 'obtained through Big data. After a pars distruens, this contribution will conclude with a pars construens; assuming the previous critiques, a mixed methods research design approach will be described as a general proposal with the objective of stimulating a debate on the integration of Big data in complex research projecting.

Keywords Big data · Digital methods · Socio-technical assemblage · Actor-network theory · Mixed methods

1 Introduction

Big data is here to stay'. This is the key statement on which Bill Franks' book (2012) is based. As chief data analyst in marketing for several firms and private companies, Franks explains the importance of being prepared for the increasing amount of sources of information and

P. Giardullo (✉)
Department of Economy, Society and Politics (DESP),
University of Urbino Carlo Bo, Via Saffi, 15-61029 Urbino, Pesaro-Urbino, Italy
e-mail: paolo.giardullo@uniurb.it

Published online: 29 January 2015

data storage that are available for business purposes.[1] Big data (from now on, BD) are, at present, pervasive and unavoidable for anyone who wants to work in marketing in this Internet era. As we are aware, BD are not only interesting for marketing purposes, they also have the potential, as well as the claim, to be a true revolution in science and research (Anderson 2008). They represent the frontier, and possibly the future; as Neelie Kroes, former European Commissioner for Digital agenda, stated, 'Knowledge is the engine of our economy. And data is its fuel (quoted in Barland 2013, p. 12).

This paper aims to foster further discussion on socio-material interactions typical of science and technology studies (STS); this frame certainly appears useful, because of the double-folded nature of the process of BD production. Describing it as an example of mutual shaping between technology and practices in everyday life, I will mark out the importance of BD for social science in a different way. I will use a typical STS approach based on actor-network theory (ANT; Latour 1987, 2005; Callon 1987; Law 1987) in order to achieve this, accounting for both material elements and linked social practices. With regard to the former, the diffusion of Internet connections has quantitatively risen to important proportions in recent years; the global number of Internet users has increased to 2.3 billion, which is a third of the world's population. Moreover, this number is still rapidly growing; there has been a 40 % increase in only 4 years, from 2009 to 2012, which is five times the increase since 2000.[2] In the EU, one of the most well-connected areas in the world, the percentage of active Internet users increased by 22 % in the same period, reaching 70 % of people aged up to 74 years.[3] Moreover, as mentioned above, it is not just the spread of broadband connections that matters; rather, an increase in the use of different practices, as well as the birth of new practices, on the Web has been observed in several studies (Consalvo and Ess 2011; Bakardjieva 2005). These practices can be recorded and stored as data useful in gaining information on how the practices themselves are performed (González-Bailón 2013; Lazer et al. 2009).

BD are gaining visibility and success in social sciences, even overcoming the division between humanities and computer sciences (Snijders et al. 2012); this appears to be an additional step in the so-called digital methods, which emerged earlier (Rogers 2013, 2009). Nonetheless, from a qualitative point of view, the point is pretty different; BD, as I will argue below, overcome the mere question of a jump of scale from small to very big. Indeed, discussing BD is not only a matter of volume, as different authors have agreed (boyd and Crawford 2012; Franks 2012). Therefore, the opportunities linked to such interaction between practices and technological acceleration for social researchers are numerous, accordingly (Manovich 2012b, a; Lazer et al. 2009). I will outline some examples, stressing the fact that we are now not just applying more powerful tools, but are also facing something that is different and new.

Despite a promissory rhetoric developed since the birth of BD as a label (Anderson 2008; Cukier 2010), some risks are just around the corner. Particular attention is required in order to avoid some criticisms, most of which have been recorded with regard to ethical issues (Ohm 2013; Neuhaus and Webmoor 2012; boyd and Crawford 2012). However, it is not only a matter of ethics in social research; the claims for the methodological power of bigger and bigger datasets, increasing speed in analysis and data collection, are creating genuine hype in social sciences, an explosion of self-declared BD research (Tinati et al. 2014), and I will analyse the connected methodological risks. The concerns raised primarily involve the

[1] The book title, not by chance, is: "Taming the tidal wave of Big data".

[2] Internet World Stats, usage and population statistics, checked on June the 5th 2014.

[3] Comparison between Internet use in households and by individuals edition 2009 and 2012; data coming from Eurostat: statistics in focus.

validity of the results that researchers can obtain when BD are the main, or (even worse) are a unique, source of information (Tufekci 2014).

After a pars distruens I will conclude with a pars construens, and I will describe a mixed methods research design approach (Morgan 2007; Lieberman 2005) as a general proposal. The objective of this second part is to stimulate a debate on the integration of BD in more complex research design.

2 Socio-technical assemblage

The equivalence of BD and web data is very often taken for granted, and this is especially true with regard to that which concerns the social sciences. Although such an assumption might ease numerous arguments in favour of, or against, their use, some clarifications are required. Indeed, I would consider some further distinctions in order to underline the qualitative steps forward that the Internet permits. Such steps forward are the same as those on which the triumphant rhetoric of BD is based, with regard to that which is of concern to the social sciences, but I will detail these subsequently.

To begin with, the Web is not a unique source of BD, in fact it is one of several, for instance, industry, healthcare and mobility systems (Demchenko et al. 2013). These are even capable of meeting the requirements of all the characteristics that have been individualised in the literature, such as the three ‘Vs’: Volume, Velocity and Variety (Brynjolfsson and McAfee 2012; Zikopoulos and Eaton 2011).[4] Volume, the size of information collected in databases, is the most evident, but it is not the major characteristic; Velocity and Variety should also be considered. In this configuration of the object, there is something latent and taken for granted; the technological component. Perhaps the most important element distinguishing BD from other huge collections of data, that is, census data, is the fast and automatic generation of a high volume of information, which means delegating data collection to an automatic device. In fact, huge databases are ‘populated’ through specific scripts that are nested in servers and types of counter machinery. This is possible and affordable on a large scale thanks to various technological improvements, such as the increasing rate of operations per second and the speed of collecting data through computing machinery (Hilbert 2012). Computational capacity has also increased enormously. In 2007, it reached a global capacity of 6.4 trillion instructions per second for personal computers, which is equal to an increase of 58 % per year since the 1980s (Hilbert and López 2011).[5] Adopted measures of estimation are providing us with insights regarding world storage capacity, which has increased since the 1980s, reaching 295 exabytes in 2007 (Ibidem), and is still increasing further.

In contrast, the heterogeneity (variety) of data collected is of added value for those who are interested in data analytics. It is precisely this feature that enhances the importance of web data as BD sources; a single source, for example, Facebook, allows the management of datasets that include information such as gender, age, geographical location, interest in a certain topic, etc. According to this, BD is something we encounter almost every day, several times per day. More precisely, we produce BD sources during many of our daily activities, that is, every time we access a browser.

[4] Or five if we include value and veracity (Demchenko et al. 2013), but in the interests of economy they will not be accounted for in this paper.

[5] According to Hilbert and López (2011), Communication capacity is certainly growing but they admit that still “(…) the digital age increased our capacity to store information […] much more than our capacity to transmit information through broadcast and telecommunication networks.” Anyway this element will be further analysed later on as one of the main criticalities of BD.

If we consider the proliferation of different Internet-based practices, we can easily recognise how varied the kind of information available through the Web can be, and how almost all our activities on the Web are generating information. For instance, the McKinsey Global Institute (2011) estimated that an average of 30 billion contents are shared on Facebook per month. This example contains the two key factors that are necessary for the understanding of web data and its relevance for the social sciences. Tweets, Facebook and Blog posts, as well as Google queries, are all useful elements in allowing researchers to measure behaviour (McAfee and Brynjolfsson 2014), and indeed this is what happens. These are all practices of human interaction that occur thanks to the Web. Social networks are maybe the most evident and well-known places where such types of data are collected; however, they are part of a broader digital turning point in social practices. Indeed, as many scholars have pointed out, the Internet provides the opportunity to reshape previous practices, such as communication, one-to-one relationships, group and community interactions and networks of help and support (Wellman 2011; Ellison et al. 2007; Boase et al. 2006; Wellman and Gulia 1999). New practices are also emerging, including cultural consumption and production of artistic goods, such as digital texts (Giardullo and Lazzer 2014) music (Magaudda 2011, 2013) and photography (Schwarz 2009, 2010).

Since all of these activities occur on the Web, they can be recorded, traced and, hence, analysed through specific tools. I will discuss this in greater depth in the next paragraph, but at this point some more attention must be paid to the relationship between the social and the technological aspects of the relationship that I have sketched. Describing the expansion of digital technology and the emergence of brand new practices, as well as the online redefinition of typical off-line practices, is not sufficient. The two partners in this relationship (practices and the expansion of technological infrastructure) are necessary conditions for the argument I have developed here, but they are not enough. In order to provide some useful elements of theoretical reflections, the relationship between the two partners must be analysed. Indeed, the main argument of this section is that BD are qualitatively different from mere huge out-of-scale collection of data. As already stated, it is not only a matter of volume; the shape of BD and their subsequent added value inherently goes beyond the jump of scale. Describing them separately, as carried out earlier in this paragraph, is a necessary starting point, but is not enough. The fact that heterogeneous elements exist is not sufficient for adding a step forward in reflection. According to the ANT approach (Venturini 2010; Latour 2005), a third key element of the interaction between different entities is the interaction between the various entities themselves. Networks are composed of an array of heterogeneous elements (Law 1987) that include both humans and non-humans; they are types of actors that should be considered symmetrically (Latour 1987; Callon 1987) avoiding a deterministic starting point, both socially and technologically (Latour 2005). The interaction process is fundamental, as Bijker and Law have clearly stated:

> Only when the self-evident and unambiguous character of such assemblage has been deconstructed does the quest for the origins of their obduracy become relevant. (Bijker 1993, p. 292).

Of course, these two scholars were not referring to the Internet, as they were writing in a period in which the Web was an instrument for the use of highly skilled people only; the so-called age of internet wizards (Wellman 2011, 2004). Between the end of the 1980s and the beginning of the 1990s, Bijker and Law, together with their colleagues, were engaged in a theoretical struggle in order to affirm the importance of technology and science as elements through which to observe and understand society. This struggle was centred on socio-technical assemblage as a proper object for social inquiry. They were working on material elements that, now, as then, are part of the everyday lives of individuals.[6] In attempting to recover this legacy, after the already affirmed importance of the Internet in so many social practices, a deconstruction of its naturalised, and taken-for-granted, use is required. An adoption of this approach, which relies on the general framework of STS, may be useful in order to reinforce the epistemological strength of BD, avoiding some triumphalist rhetoric that often affects the use of huge data bases in the social sciences. As already stated, it is not only a matter of volume.

I have hinted above that BD might be considered as the most evident expression of these transformations in recent years: on one side, the increase of informatics' capacity, and on the other the change in everyday practices. However, I have not yet discussed the mutual shaping of the automatic collection schemes that allow informatics instruments to create BD, which means analytically addressing the interactions between them.

An informatics architecture, on which BD are based, is normally built according to the needs and wills of engineers. Nonetheless, they are not built in a vacuum, creating from nowhere users willing to complete records on a database through their surfing sessions. Here is an example. In 2001, the first patent for monitoring shopping carts on Amazon was filed by the research team lead by Jennifer A. Jacobi.[7] That recommendation algorithm development represents a key feature for online shopping, as we may read from the patent document:

> The user is commonly faced with the onerous task of having to rate items in the database of items in the database to build up a personal rating profile. This task might be frustrating particularly if the user is not familiar with many of the items that are presented for rating purposes.

By providing suggestions based on previous choices, it aims to ease Web-users' experiences while surfing a website looking for a purchase of any type of content. Amazon, founded in 1994, was already an affirmed online book store, with thousands of goods in its catalogues, when the recommendation algorithm was proposed. This means that there was already a market, because the constellation of actors was already aligned in an established network. An electronic market,[8] composed of the Internet and Internet users that also surf the web with the aim of making purchases, was already there, which justifies efforts and investments made in that direction. In order to ease the buying experience, Amazon and its competitors further implemented such algorithms. In the same way, Google and Facebook often make suggestions and recommendations to us, through advertisement banners that describe similar objects, goods and people that might interest us. As MacKenzie (2014) showed, algorithms are not passive elements, nor are they purely deterministic triggers, rather, their feature is that they are also capable of functioning in environments[9] that are different from the one in which they have been created, fostering a new kind of interaction. In this case, such algorithms are

[6] In the chapter in his edited book, Bijker analysed the fluorescent lamp; later on he dealt with bakelites, bicicles and incadescent bulbs (1997).

[7] http://google.com/patents/US6317722, Retrieved on February the 12th, 2014.

[8] An attempt to apply ANT approach to the description of the raise of electronic market in the case of EUREX has been conducted in Baygeldi and Smithson (2004). The ability of Actor Network Theory (ANT) to model and interpret an electronic market. Creating Knowledge-based Organisations, 109–26.

[9] Or 'Ecology' as meant by Abbott (2005). In contrast with deterministic concepts such as mechanisms and organisms, he argued that "When we call a set of social relations an ecology, we mean that it is best understood in terms of interactions between multiple elements that are neither fully constrained nor fully independent." p. 248.

capable of working both for the proper and optimised technical functioning of the website and the users' needs.

As is well known, the entire 'recommending process' is possible precisely because our behaviour on such websites, can be monitored and cross checked with our data, for example, as language, through HTTP cookies.[10] Such algorithms also work thanks to other types of data (demographic data, such age, gender, etc.). Users insert these when they create or update their own profile, making them available in major server farms. As boyd argued (2012), 'Today, information about people can be easily accessed with just a few keystrokes'. This is basically possible because people, or rather users of the Internet and social network sites, add their own and personal information. The algorithms nested there, through the requirements they need to work, actually discipline the practices of users that want to have access to such services. The latter all follow the technological scripts (Akrich, in Bijker 1992): 'technical objects define a framework of action together with the actors and the space in which they are supposed to act' (Ibidem, p. 209). Transposing the process described for general technical objects to the present Internet context, we might consider the rules for user registration with such Web services as the script provided by the informatics infrastructure of websites, and for social network sites in particular. In this case, the former allow the latter to collect a huge amount of social data. It is not by chance that the main social network sites explicitly ask newcomers to use their own real name (Omernick and Sood 2013; boyd 2012). The deal is easy: being connected through such infrastructure, social network sites and webmail, implies the following of scripts that require the use of real names, surnames, place of residence, past experiences, interests, etc.

This review may drive the reader to consider BD as the result of a technology push, rather than a co-shaping process. As some critical scholars have argued, this is equal to the commodification of users (Fuchs 2014). In reality, this is more likely a side effect, certainly not a deliberate action, of a co-shaping process. As already mentioned, an algorithm can work by exporting its performance outside the context in which it was originally conceived. In this case, the performativity of algorithms is definitely positively greeted by users. As Geser recognised for mobile phones, a communication technology cannot be viewed as a causal determination, 'but rather as a tool providing a set of specific functional capacities which may be more, less or not at all exploited under various socio-cultural (…) conditions' (Geser 2005, p. 255). Indeed, balance between the social and technical components can be further restored if we consider why people add their information, their pictures and their socio-demographic details. People recognise in the Web an array of opportunities (Bakardjieva 2005, p. 187) that goes far beyond the simple 'let's keep in touch'. As well-known scholars have discussed, web-users are creating horizontal networks of communication (Castells 2009; Hampton and Wellman 2003).On one side, they are encouraged in doing so for political and information purposes (Castells 2009; Lin 2001), while on the other, they adopt the Internet, and more specifically social-network sites, as a strategy to reinforce their social ties and even to increase them (Pew Internet Research 2014; Hampton and Wellman 2003; Wellman 2002). In order to do so, they create profiles for others (boyd 2008), creating a socially mediated publicness (Baym and boyd 2012), one of the main features of Internet 2.0. Moreover, this publicness is not an idle entity; rather, it is dynamic, interactive and mobile. In a nutshell, users create public profiles in order to interact through the Web, adding multimedia contents to conversations, such as pictures or videos, giving comments or looking for advice. In addition, we should not forget that scripts disciplined by algorithms can also be modified by users.

[10] As defined by Internet Engineering Task Force (IETF) of Berkeley University in the document available on http://tools.ietf.org/html/rfc6265#section-3. Checked on August 9th 2014.

As studies in participatory web cultures (Beer and Burrows 2010) have shown, users and members of particular communities often bend technological scripts to their own interests, as demonstrated by Apple products’ jail-breakers communities, for example (Magaudda 2010).

To conclude this section, we can remark that BD are an example of socio-technical assemblage. Material elements, such the informatics infrastructure interwoven of prescriptive algorithms, and the performances encoded within it, are producing information. The scale actually does not really matter per se, although the socio-technical process that underlines BD creation is qualitatively different. Instead of researchers interrogating people, we have people acting, and in doing so they are leaving traces. On the Web they act by following the encoded scripts within the technological infrastructure of research engines (Google is the most popular example) and sites (Facebook, Amazon or Twitter, just to name the major sites), which create the complexity of the Internet. This way, they can be automatically tracked in what they do, and they are, because of the script encoded in the infrastructure. Reflection was accordingly required in order to better specify the type of source of data to which we are referring when we talk of BD. As already stated, it is may be the most evident expression of these transformations in recent years: on one side, the increasing of informatics’ capacity, and on the other, the change in everyday practices. This is a key element on which there is presently no doubt.

The added value of BD is not only in terms of scale, but also in terms of the type of data collected and the subsequent level of abstraction for the analysis. On one side we have interactions that take place online, and which are peculiar to those environments, Internet-based social phenomena, as someone referred to them (De Paoli and Teli 2011), while on the other we have a way to detect, and therefore make inferences concerning, the offline interactions, power relationships, etc. that reverberate on the web, the typical approach of digital methods (Rogers 2013; Rogers and Marres 2000). In a nutshell, BD appear to provide us with the insight and the opportunity to create figures of both. In the next section, I will detail several examples of BD adoption after discussing some more insights regarding the implied methodological jump.

3 Promises and great expectations: some undeniable shortcuts

The socio-technical assemblage made possible by the interaction of people adopting the Internet and the technological infrastructure is continuously reproduced by computer-mediated social practices. I have already pointed out that such practices can be recorded, adding metadata of the practitioners, such as geographical location, gender, age, etc. These continuing interactions among users, and between users and the Internet itself, are producing added value for social research. The key element is that we are capable of observing what people do almost directly. Instead of interrogating, interviewing or surveying them, we can just observe the outcome of their activities. Rather than answering questions related to what we want to know, they are simply doing things. They are acting. Moreover, they are leaving us their fingerprints on a huge scale. Further, the shared content, and the queries regarding a certain topic, can be the unit of analysis for researchers who are addressing online interaction.

As is clear, the sociotechnical assemblage is leading to some undeniable shortcuts for social researchers. Before going further through some examples from recent international literature, I will develop the epistemological and methodological promissory horizon of BD. In doing so, I will recover the rhetoric discourse fostered by enthusiastic supporters. Once again, the most engaged BD supporters are coming from marketing. In 2005, Eric Schmidt, former CEO of Google, stated:

> We would also store everything; we would store all of your information. How many of you have more than one PC? I would guess everybody. You have one at home, or a Mac, you have one at home, one at the office, maybe you have a portable, maybe you have two. (I counted one day, I had like 15, but—OK, I'm not normal.) How many cell phones do you all have? One, plus you have a back up, and if you're a CDMA network you have the GSM one to work in the non-CDMA and so forth. How much information do you have trapped in these client devices? Why don't you put it all on your server? What server? The server that you'd have if networks were free and storage were infinite—it would all be there. The transparent personalization would come with it—the machines would immediately know you. I want to emphasize this is with your permission, of course. We would know enough about you to give you targeted information, the targeted news, the targeted advertising, to make instantaneous, and seamless, happen.[11]

A decade ago, the idea of cloud-stored information for all types of customers who had a connected device would have sounded almost like science fiction. Now, it has become real, cheap and easy, and is naturally adopted. Again, new computers bring with them the scripts envisaged by the algorithms of the cloud, since the main commercial operative systems, Microsoft Windows and Mac OS, require an account for their cloud storage application. However, as already stated, comments on blog-posts, such as on Facebook, Google queries and tweets and re-tweets on Twitter, are all out there and available for us as social researchers. They are all embedded in our daily life, and are increasingly interconnected through this technological infrastructure called the Internet.

Beginning with these assumptions, new issues have arisen for social scientists, especially from an epistemological point of view. More generally, this is the perspective that has been fostered since the beginning of its use; the concept of BD is charged with a promissory rhetoric. The potential added value of the availability of such an amount of information is considered an actual revolution by most enthusiastic supporters. In a key article in Wired, Anderson (2008), foresaw the 'end of theory' and the failure of 'the old way of doing science'. This is a general discourse that intersects several aspects of scientific research, but might ring especially true for social sciences and the Web. On one side, since a growing share of the global population is gaining access to the Web and to connected services, the potential basis for representation is increasing as a consequence. As already stated above, a growing number of practices also occur online, while others are directly emerging on digital highways; in this manner, the potential for different kinds of proxies for social inquiry is greatly increased.

Given the two premises, almost in a syllogistic way, the conclusion appears to be that we are facing a new scientific revolution. According to Anderson, an unprecedented technological puissance now would allow researchers to directly apply a correlation matrix to populations. In a nutshell, the aim of being representative of the target universe might no longer be an issue. This would be the new way of conducting science; rather than hypotheses to be tested on a surrogate of the real world, now, BD supporters affirm, it is possible to work directly on the real world itself. This means avoiding the logic of sampling; rather than creating correlations between randomly selected units of analysis, we look at what 'actually' happens. Conversely, this is translated in so-called 'data-driven research'; since there is no need for sampling before beginning data collection, yes, we are already working on populations; there is no need to assume a hypothesis. A researcher can directly observe whatever s/he wants. Nonetheless,

[11] Eric Schimdt's speech to the Association of National Advertisers, 2005, retrieved on Google press on May 21 2014. http://www.google.com/press/podium/ana.html.

BD supporters recognise the difficulty that lies behind this new challenge, and at the same time they argue that it is a way to keep up-to-date along the journey of innovation

> The evidence is clear: data-driven decisions tend to be better decisions. Leaders will either embrace this fact or be replaced by others who do. In sector after sector, companies that figure out how to combine domain expertise with data science will pull away from their rivals. We can’t say that all the winners will be harnessing big data to transform decision making. But the data tell us that’s the surest bet.[12]

This boost of overconfidence is primarily linked to business and economic decision-making rhetoric; in social science there is no such strong discourse. However, the equivalence of data, information and knowledge is spreading pretty rapidly (Barland 2013). Furthermore, the fascinating mantra of ‘the more data available the more accurate you can be in your argument’ is gaining importance in social science. The fact that the concept of BD is spreading in the international community (conferences and journals) demonstrates a certain interest in the topic.

The increasing speed in collecting and analysing data through informatics tools, at relatively low research costs, as conducted in much web-based research (Bainbridge 2007), is definitely an effective way of obtaining success, for example, some tools are becoming fairly well-known among research teams, providing access to thousands of pre-analysed data sets, such as Trendspottr.com. With relatively low economic efforts, the access to its API[13] is capable of providing insights on trends regarding emerging topics on Twitter. Other researchers (Hale 2014), those who are the most skilled, are using the so-called ‘Spritzer’ Twitter access, which gives back subsets of tweets. By creating their own API, they may directly access hundreds of thousands of contents shared on the popular social network, almost in real time.

Considering the characteristics of BD, the ones listed above (embeddedness, velocity, automaticity and relatively low costs of use in terms of time and money) indubitably represent a genuine shortcut for social research. It is not only a matter of sparing resources, but also a different way to observe reality. Some processes, for instance, the spread of a topic, may be observed directly instead of by asking for people’s opinions on such. Some others might be reconstructed ex-post because, as stated several times, the most popular practice among Internet users, generally speaking, is generating data, which enable analyses such as:

- Real-time monitoring—in a closed environment, such as a website, every click can be monitored and recorded (Amazon with shopping carts, and, more recently, Facebook)
- Topic trends—through keywords analysis (or hashtags on Twitter). In the international literature, topic-trending is a very common example of research that we can find both in computer science and Internet studies, as well as in the conference proceedings of new media studies
- Creation of historical series’ and cross-country comparison—recording data from the Web allows scholars to have sufficient data to create longitudinal analyses; the variety of geographical locations of Internet users even allows comparisons between countries.

[12] Big Data: the Management revolution, McAfee and Brynjolfsson in Big data’s management revolution, The Promise and Challange of Big Data, Harvard Business Review Insight Center Report, September 11, 2014.

[13] In computer programming Application Programming Interface, is exactly a software interface able to regulate software interaction. In this case, APIs are useful interfaces that allow interaction between different web applications as for instance social networks sites.

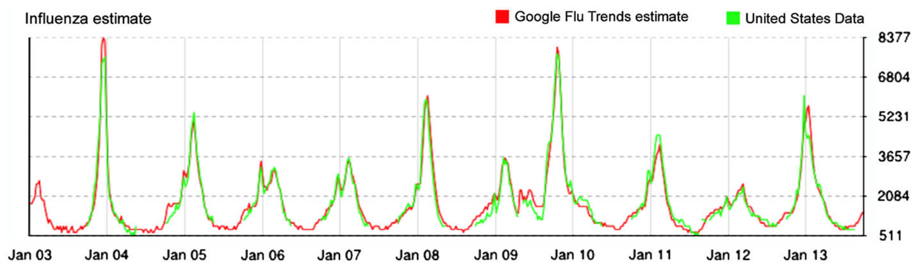

Fig. 1 Graph comparing Google Flu Trend estimation and official CDC data. *Source* Ginsberg et al. (2009)

With regard to this latter opportunity linked to BD, the most influential example of potential for this kind of analysis has been directly produced by Google. The contribution of the famous Google Flu Trends has been a pluri-quoted article, following publication in Nature (Fig. 1).[14]

In this well-known article, the Google Inc. researchers demonstrated how the queries typed by users looking for information about influenza symptoms on the Web can be a valid indicator for epidemiologists: the trend cases of flu registered by the Centre of Disease Control and Prevention (CDC) matched the data produced by Google services. Moreover, the index of Google Flu Trends was conceived as a forecasting tool for CDC epidemiologic reports. It has been considered as one of the most important achievements in BD research. In fact, as Rogers argued (2013),[15] a key challenge for those who are working with web data, and BD I would say, is the reliability of their outcomes. A valid way of assessing an analysis obtained through online data is an offline benchmark. Google Flu Trends is possibly the best example of how the gap between the online and offline has been proficiently filled; online forecasts regarding flu trends were exactly matched by the actual epidemiological data. In this manner, BD supporters claim, the reliability of web data, has been proved.

Another example of how to relate BD obtained through the web is the graphically elaborated proposal put forward by the data artist Erik Fisher. He transposed all the pictures uploaded on Flickr.com (up to 2010) with a geo tag in several cities of the world. He worked both on European and US cities; London, Philadelphia, San Francisco, etc. This exercise shows the potential of using such totally native web data on a fairly large scale, with a high level of detail. For example, in tourism management, an image like this can be useful in order to visually understand where tourist guides lead travellers and tourists coming to the capital of Italy (Fig. 2).

In the picture, the red dots are the geo-references of the pictures uploaded to Flickr.com by users. The blue dots on the map are pictures taken by locals (people who have taken pictures in this city for over a month or longer). The red dots, the majority, are pictures taken by tourists (people who declared that they were a local of a different city in their personal information, and who took pictures in this city for less than a month). The yellow points are pictures from which whether or not the photographer was a tourist (because they have not taken pictures anywhere for over a month) cannot be determined. They are probably tourists, but just might not post many pictures. Superimposing the map of the city, we can pinpoint the areas where tourists are most concentrated. Instead of conducting a survey asking people

[14] Ginsberg et al. (2009) Detecting influenza epidemics using search engine query data, Nature vol. 457, 19 February 2009.

[15] Actually Rogers quotes Mike Thelwall, an affirmed webometrician, who said that the challenge "is to demonstrate the web data correlate significantly with some non web data in order to prove that web data are not wholly random" (p. 205).

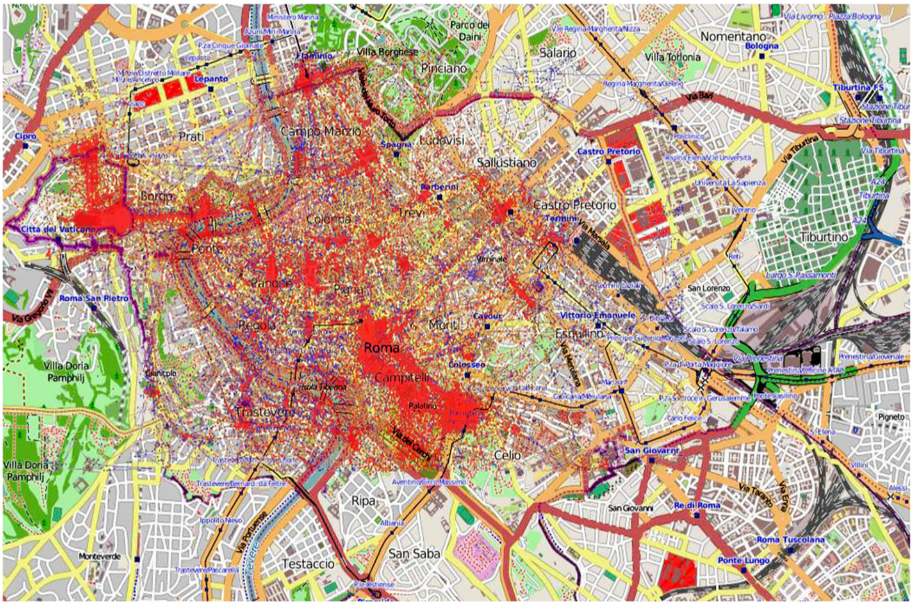

Fig. 2 This image has been overlapped to a traditional map of Rome, with the aim of showing the precision of the data. Original one is available with other cities' images at https://www.flickr.com/photos/walkingsf/4671567441/in/set-72157624209158632/

where they focus their attention in the Italian capital, we can save money and time being more accurate.[16] Furthermore, in this case, the offline benchmark is available, even if in a less rigorous manner.

Considering once more the property of monitoring, this time in real-time, the spread of some specific content, detectable through the so-called hashtag,[17] can be followed in time and space thanks to web geotagging techniques (Elwood 2010), as was carried out by a group of human geographers following the hashtag #LexingtonPoliceScanner (Crampton et al. 2013). The researchers detected the spread of this hashtag during the celebrations of the Wildcats' (the University of Kentucky's men's basketball team) victory in the 2012 NCAA championship. The celebrations lasted until the morning after, and involved some trouble in the city of Lexington (KY). The geographers were able to follow this hashtag throughout the entire night. They recorded 12,590 tweets generated by 6,564 users using the same hashtag, and showed how the topic spread far outside the city limits (Fig. 3).

As readers may have noticed, I have already quoted Twitter several times in this paper. Studies and research reports regarding Twitter and its hashtags' ranking, tracing and monitoring can be found very frequently. It is true to say that these hashtags are very trendy objects for BD, and Twitter is also a highly versatile source of data; in fact, it is common to find pictures such as the one shown below.

In addition to geotagging, Twitter also allows researchers to adopt network analysis techniques because of the tweet and re-tweet interaction (once more, a script) among users made

[16] A global map produced by Eric Fisher based on Twitter is available here: https://www.mapbox.com/labs/twitter-gnip/locals/#.

[17] The celeb way of labelling a content in order to assure that the content is referred to a specific object, topic or event.

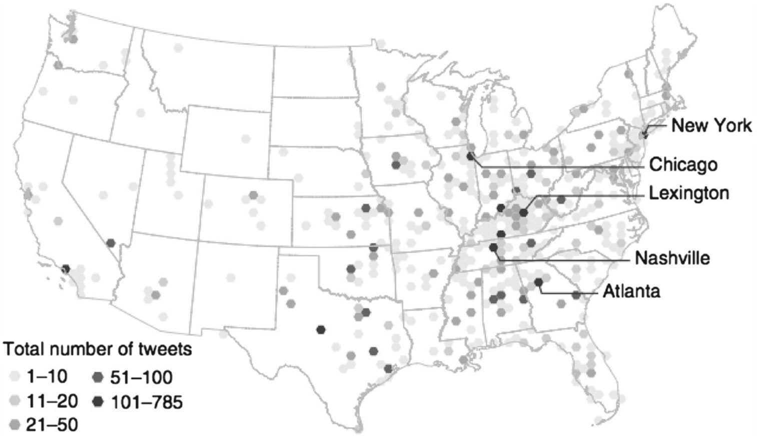

Fig. 3 Distribution of all #LexingtonPoliceScanner across the USA. *Source* Crampton et al. (2013)

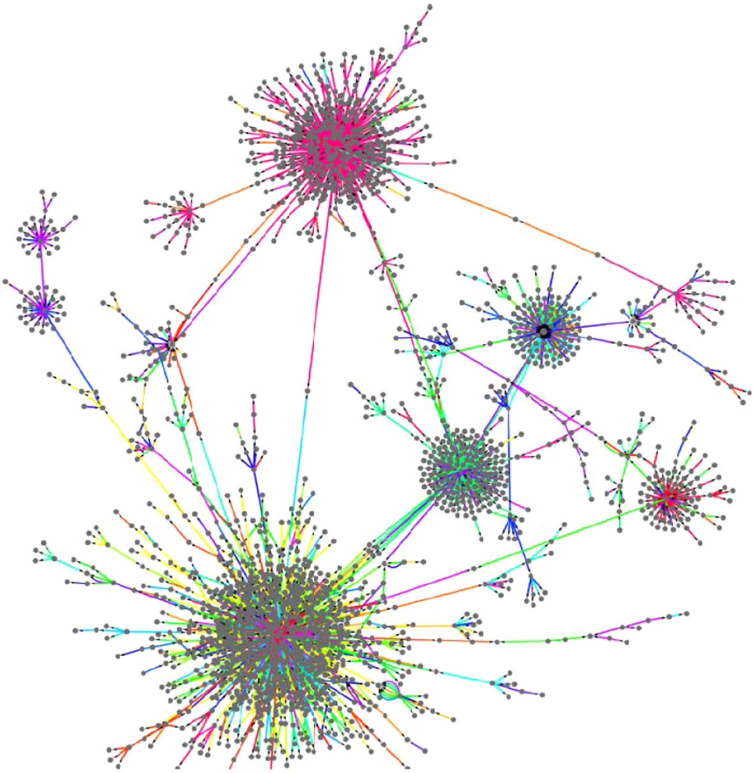

Fig. 4 Network analysis of the re-tweet chain for #Airfranceflight. *Source* Kwak et al. (2010)

possible by the hashtags. In this case, Fig. 4 shows the configuration of the topic Air France Flight 447, which went down in 2007 (Kwak et al. 2010). Another example of Twitter monitoring derives from a fairly well-known contribution on the riots in northern African countries; the so-called Arab Spring. Between 2010 and 2011, during the protests, researchers (Lotan et al. 2011) monitored over 90,000 unprotected Twitter accounts in Egypt and Tunisia. They were able to describe the networked production and dissemination of news during the riots across activists, bloggers, journalists, mainstream media outlets and other engaged participants that, arguably, lead the protests. They detected alternative news streams to the main media channels, which confirmed the hypothesis that, in cases of crisis and dramatic events, journalism is a type of direct conversation between different parties that are directly involved in such events.

All these examples sound very promising. The claims for the methodological power of increasingly large datasets, as well as increasing speed in analysis and data collection, are creating a real hype in social research. Nonetheless, particular attention is required in order to avoid some pitfalls.

4 Validity and reliability: some pitfalls

Although the examples sketched above are successful, a critical point of view should be kept in mind. As every research technique benefits from specific strengths, BD also have their own weakness, and this should also be considered. As BD may potentially appear capable of solving some typical social research problems in a brilliant manner, some problems, such as lists of populations or criteria for selecting sources, actually still remain.

To begin with, and in connection to the last examples cited, I will further analyse Twitter, underlining some of the possible problems involved. Between micro-blogging and acting as a social network platform, this popular site is a flexible source of data, as we have observed, and it is apparently highly efficient. As in the previous examples, we can monitor what occurs on Twitter in real-time, with very low costs. It allows for the collection of different types of data: georeferenced data, network analysis measures, demographic data (age, gender) and textual data. Having access to such a single source actually opens an array of diversified opportunities for a single researcher or a research team. Therefore, in social sciences, as well as in trans-disciplinary, journals, Twitter is a highly popular way of collecting data.[18] As some scholars have pointed out, it is considered desirable because of its own properties (Snijders et al. 2012). The hidden pitfall is to assume that it is representative of online interaction and, therefore, to consider it as the best way to infer offline processes. In analysing this effect Tufekci (2014) denounced this risk using the metaphor of model organisms for biological research. In fact, in biosciences, most of the experiments are conducted on specimens of certain biological species, such as *Drosophila melanogaster*, or *Escherichia coli*. The choice of these particular species lies in the assumption that they possess several general biological characteristics, meaning that the outcomes of experiments can be successful generalised to other species. Tufekci has argued that this type of choice has some epistemological limits for biologists, which are linked to the tight array of options that might affect the generalisability of their outcomes. Such limits are even greater for that which concerns BD and social research. Assuming that Twitter is the prominent method of web interaction might be misleading; when compared to the entire set of possible combinations that the socio-technical assemblage actually offers, Twitter is just one of a bundle. The fact that Twitter offers such varied types of data may drive researchers to focus their research on this site alone; however, paradoxically, it offers just one main type of interaction, tweet and re-tweet, and only concerns a certain type

[18] Since 2010, on Sage’s Social Science Computer Review, 53 papers (57.6 % of total) has been published basing their empirical data on Twitter. Site consulted last time on 17th July 2014. As Zeynep Tufekci noticed (2014), at the 2013 edition of International Conference of Weblogs and Social Media (ICWSM) almost more than thirty papers (the half of total accepted) were based on some kind of Twitter analysis.

of users. Indeed, even if tweets and re-tweets are accepted as significant forms of Internet interaction, Twitter's users are characterised by a certain profile. Twitter use is polarised by certain groups, for instance, a recent study (Smith and Brenner 2012) showed that US Twitter's users are in the youngest section of the population (26 % of them is under 29 years old) and primarily belong to a specific ethnic group (28 % are black non-hispanic). Hence, the claims for general validity of such research results are undermined if based on a single type of interaction. Therefore, a further reflection should be considered: concentrating mainly, or, even worse, only, on a certain manner of online interaction equates to considering an issue as being confined within a single social network, with no possibility of expanding outside it (Tufekci 2014). Considering the characteristics of online interaction, such a creeping assumption is unrealistic, and therefore is not acceptable.

Moving away from Twitter to become more general, another criticism that has been recorded regarding BD concerns the figures relating to the Internet users that benefit from a certain service (boyd and Crawford 2012). Such figures should be the denominator for the calculation of users' share of certain webpages, or social network sites; these figures are calculated from the data available through API connections, but there is currently no direct way to ascertain exactly how many users are interacting in a certain online environment. Data are owned by firms that have created technological infrastructure on the Internet, and it is they who actually have such data available. Consequently, being external to these firms, we must trust or assume the good quality of the data to which firms grant us access; in a nutshell this means depending on the expertise of someone else. As externals, we have no control over a large set of variables that might affect the data we are analysing. We can pick the hypothetical case of a particular shared content, that might reach some dozens of thousands of users, but numerous people connected through their own account are not necessarily taking the interaction (i.e. sharing content), rather, some are just listeners (Twitter 2011), who are maybe sharing the 'listened' content vocally offline, or in another way that BD are unable to record. Similarly, we have some large and complex measures regarding the intensity of certain links between sites, or shared contents; however, only little information regarding the quality of such links is available –at the moment.[19]

Another order of pitfalls consists of the sneaking risk for the interpretation of BD. As Lazer and colleagues showed, BD might fail as a source for forecasts (2014); the shining example of Google Flu Trends is a double-edged sword. Forecasts should meet three requirements: when an episode may occur (time dimension), where (space) and how intense it is (intensity). Recent analysis of the data produced by Google about users' queries for flu showed that a dimension was missing: indeed, data overestimated the actual intensity of epidemiological peaks. Although this had no actual consequences, it showed that some phenomena, very similar to over-represented facts in the newspapers, such as 'over-querying', might affect the reliability of forecasts.

Correlations between huge amounts of data and the consequent inferences might also be erroneous. The so-called risk of apophenia, 'seeing patterns where none actually exist' (boyd and Crawford 2012, p. 2), is a possible side effect of BD analysis, embracing critically automated instruments for analysis. We are certainly capable of observing a great number of

[19] It should be mark out that computer sciences applied on content analysis is proceeding very fast. Information retrieval is a cutting edge field of research for informatics: some of the key articles are about machine learning and sentiment analysis. The former is Blei et al. (2003), on the so called LDA (Latent Dirichelet Allocation) that allows to automatically individuate on probabilistic basis the topics contained into a complex corpus of texts; the latter is Pang and Lee (2008), about sentiment analysis, a way to automatically calculate partisanship of texts about a certain topic.

people expressing themselves saying something and sharing something else, but by basing our inferences on BD only, we risk losing the sense of action (example of irony). This is a pitfall that could affect not only Twitter research, but also the georeferenced information of images, as shown in the Flickr example; without the ‘voice’ of the individual who has snapped those pictures, for instance, we are losing the reason behind why some parts of the city were forgotten.

The last order of pitfalls is linked to the concrete practice of research. Not infrequently, those who want to use such BD sources are actually using pre-configured BD analysis, deriving directly from the firms’ services. The use of pre-configured API, or other services, forces a researcher to be dependent on the expertise of others who produce black boxes. These might be convenient but, in parallel, they might affect the validity of our research design, and without the researcher being aware of it. This process of dependency locks one of the most important phases of research, the data construction, in black boxes. This is not a risk per se; through the practices that a tool might discipline, the script reinforces theoretical assumptions that are not neutral. As frequently shown by STS scholars (Crabu 2014; Pickering 2010; Clarke and Fujimura 1992), tools for analysis actually simplify researchers’ daily jobs, creating order and sense, by disciplining their practices. In this manner, dependency on the black boxes of others can drive the reproduction of some assumptions that are part of the tool. The script encoded into a material actor might foster the spread of a theoretical assumption; nonetheless, this could expose unaware researchers to undesirable biases.

5 Conclusions

As may be clear at this point, BD are raising more questions, rather than providing answers to existing questions. Considering the pitfalls listed above, some specific questions have been raised, and in attempting to summarise the main questions we can ask: is such an amount of data useless?

This question is so big that it requires an adequate and immediate answer: no. The huge volume of data is not useless per se, but nor is it the panacea for all our research needs. As discussed in the second paragraph, as the result of a convergence of practices and technology, the origin of such data is necessarily something useful and precious. The advantages these data provide lie precisely in their origin; they are produced directly from the ecology of the social action on the Web. They are providing new insight, enriching empirical perspectives, and not only because of the increasing scale of the empirical material available. Despite the fact that web data are not the only source of BD, they represent the most common source for the social sciences, since they have already been previously managed with new media studies and Internet studies. As shown above, this has driven us to certain specific choices in selecting where collection of data should be concretely conducted; these choices have seldom reflected the actual array of opportunities allowed by the Web, probably favouring the most convenient. Nonetheless, the convenience parameter might not be sufficient for selecting adequately perspectives through which to observe social phenomena. As discussed previously, the choice of certain source of data is crucial. A singular perspective might be misleading; however, a pure BD approach is not necessarily sufficient for making inferences and analysis. This apparently looks like a failure of BD rhetoric or, at least, a downscaling of the validity and reliability of the BD approach compared to its premises.

In common with every type of data, BD are primary, data, so they must be managed and considered as such. Certainly, as expressed above, because of the manner of their construction, a socio-technical assemblage, they might be richer than other data with regard to certain phenomena; however, they are constructed through certain process types and should be considered within these. A proper generalisation of the outcomes of BD analysis likely requires complementary insights on the same phenomenon or object.

In the light of these last considerations some more questions arise:

- How to select sources of data?
- How to make inferences regarding such data?
- Are social scientists sufficient to address the BD challenge?

These three questions are probably not new; indeed the first two are similar to more traditional social research problems. Some problems remain, enhanced rather than solved. Since BD data are here to stay, the key is probably to move on pragmatically, avoiding extremisms from both sides. Once agreed that it would be short-sighted to consider BD as either useless or an actual panacea, a pragmatic approach is progressing in our minds. According to Morgan (2007), being pragmatic in social research means to refuse paradigms as totalising worldviews; refusing paradigms, reveals to deconstruct rhetorical discourses; rather that defending a specific metaphysical worldview (p. 73), a pragmatic approach should start precisely from the methodological choices that characterise the social researcher's job. In greater depth, this would be possible by mixing the outcomes available from different research techniques and tools. As pointed out by Lieberman in his seminal contribution (2005), this would mean corroborating the validity of research designs, filling the gaps in certain methodological choices with the strengths of other tools. Projections regarding the future adoption of BD as a source of information are probably linked to their adoption in more complex research designs. As claimed by several scholars with critical perspectives on BD (Tufekci 2014; Snijders et al. 2012; Manovich (2012b); boyd and Crawford 2012), the combination of different types of data is the next challenge in order to properly address BD. For instance, selecting cases for qualitative in-depth analysis on the basis of large-scale indications could be useful in order to obtain as much insight as possible and to obtain trustworthy outcomes. As Lazer and colleagues recently argued (2014), 'The internet has opened the way for improving standard enabling more traditional data collection. Instead of focusing on a 'Big data revolution', perhaps it is time to focus on an 'all data revolution' (…). This pragmatic reply has actually left one background question still unresolved: are social scientists sufficient to address BD? According to what has been discussed above, the answer is: no. There is a growing need for informatics and computer sciences skills for managing data and programming tools for analysing such data. Further, collecting data without being black-boxed by others tools is a complex task that, in the absence of pure hybrid scholars,[20] cannot be satisfied by a single area of expertise. Therefore, an increasing interaction among groups is highly recommended. Indeed, there have been some attempts, and some recent European calls have proposed the convergence of social and computer sciences. It appears that new research questions should arise from interdisciplinary teams; they should be able to create research tools in order to furnish adequate answers to trans-disciplinary and shared research questions.

Therefore, should we still talk about data-driven research?

[20] A pure mix of social sciences and informatics might probably be a new frontier for high education but this would open an other kind problems in the academic world.

References

Abbott, A.: Linked ecologies: states and universities as environments for professions*. Sociol. Theory **23**(3), 245–274 (2005)

Akrich, M.: The description of technical objects. In: Bijker, W.E., Law, J. Shaping Technology/Building Society: Studies in Sociotechnical Change. MIT Press, Cambridge pp. 205–224 (1992)

Anderson, C.: The end of theory. Wired Mag. **16** (2008)

Bainbridge, W.S.: The scientific research potential of virtual worlds. Sci. Mag. **317**(5837), 472–476 (2007)

Barland, M.: Big Data Special Report—Data-driven decision-making: interpreting the digital exhaust of everybody, everywhere, Volta Science, technology and society in Europe- **2013**(5), 6–14 (2013)

Baygeldi, M., Smithson, S.: The ability of Actor Network Theory (ANT) to model and interpret an electronic market. Creat. Knowl.-Based Organ., 109–26 (2004)

Baym, N., Boyd, d: Socially mediated publicness: an introduction. J. Broadcast. Electron. Media **56**(3), 320–329 (2012)

Beer, D., Burrows, R.: Consumption, prosumption and participatory web cultures. J. Consum. Cult. **10**(1), 3 (2010)

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)

Bijker, W.E., Law, J.: Shaping Technology/Building Society: Studies in Sociotechnical Change. MIT Press, Cambridge (1992)

Bijker, W.E.: Of Bicycles, Bakelites, and Bulbs: Toward a Theory of Sociotechnical Change. MIT Press, Cambridge (1997)

Boase, J., et al.: The strength of internet ties. Pew Internet and American Life Project (2006)

boyd, d.: Social network sites: the role of networked publics in teenagesocial life. Youth, identity, and digital media. In: Buckingham, D. (ed.) The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning, pp. 119–142. The MIT Press, Cambridge (2008)

boyd, d: The politics of ’Real Names’: power, context, and control in networked publics. Commun. ACM **55**(8), 29–31 (2012)

boyd, d., Crawford, K.: Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. **15**(5), 662–679 (2012)

Brynjolfsson E., McAfee, A.: Big data’s management revolution. The Promise and Challange of Big Data, Harvard Business Review Insight Center Report, September 11 (2012)

Callon, M.: Society in the making: the study of technology as a tool for sociological analysis. In: Bijker, W.E., Hughes, T.P., Pinch, T.J. (eds.) The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology, pp. 83–103. MIT Press, Cambridge (1987)

Castells, M.: Communication Power. Oxford University Press, New York (2009)

Clarke, A.E., Fujimura, J. (eds.): The Right Tools for the Job: At Work in Twentieth-Century Life Sciences. University Press, Princeton (1992)

Consalvo, M., Ess, C. (eds.): The Hanbook of Internet Studies. Wiley, Oxford (2011)

Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W., Zook, M.: Beyond the geotag: situating ‘big data’ and leveraging the potential of the geoweb. Cartogr. Geogr. Inf. Sci. **40**(2), 130–139 (2013)

Crabu, S.: Give us a protocol and we will rise a lab, The shaping of the infra-structuring objects. In: Mongili, A., Pellegrino, G. (eds.) Information Infrastructure(s): Boundaries, Ecologies, Multiplicity, pp. 120–144. Cambridge Scholars Publishing, Cambridge (2014)

Cukier, K.: Data, data everywhere: a special report on managing information. The Economist, 25th February 2010 (2010)

Demchenko, Y., Ngo, C., Membrey, P.: Architecture framework and components for the big data ecosystem. J. Syst. Netw. Eng., 1–31 (2013)

De Paoli, Stefanoe, T., Maurizio : New Groups and New Methods? The Ethnography and Qualitative Research of Online Groups, special issue Etnografia e Ricerca Qualitativa, **4**(2) (2011)

Ellison, N.B., Steinfield, C., Lampe, C.: The benefits of Facebook friends: social capital and college students’ use of online social network sites. J. Comput. Mediat. Commun. **12**(4), 1143–1168 (2007)

Elwood, S.: Geographic information science: emerging research on the societal implications of the geospatial web. Prog. Hum. Geogr. **34**(3), 349–357 (2010)

Franks, B.: Taming the Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, vol. 56. John Wiley and Sons, New York (2012)

Fuchs, C.: Social Media: A Critical Introduction. Sage, London (2014)

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature **457**, 1012–1014 (2009)

González-Bailón, S.: Social science in the era of big data. Policy Internet **5**(2), 147–160 (2013)

Geser, H.: Towards a sociological theory of the mobile phone. In: Zerdick, A., Picot, A., Schrape, K., Burgelman, J.-C., Silverstone, R. (eds.) E-merging Media Communication and the Media Economy of the Future. Springer, Heidelberg (2005)

Giardullo, P., Lazzer, G.P.: Digital texts: writing and reading at the digital turning point. In: Mongili, A., Pellegrino, G. (eds.) Information Infrastructure(s): Boundaries, Ecologies, Multiplicity, pp. 166–189. Cambridge Scholars Publishing, Cambridge (2014)

Hampton, K., Wellman, B.: Neighboring in Netville: how the Internet supports community and social capital in a wired suburb. City Community **2**(4), 277–311 (2003)

Hale, S.A.: Global connectivity and multilinguals in the Twitter network. In: Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems, pp. 833–842 (2014)

Hilbert, M.: How much information is there in the information society? Significance **9**(4), 8–12 (2012)

Hilbert, M., López, P.: The world's technological capacity to store, communicate, and compute information. Science **332**(6025), 60–65 (2011)

Hilbert, M., López, P.: Info capacity: how to measure the world's technological capacity to communicate, store and compute information? Part I: results and scope. Int. J. Commun. **6**, 24 (2012)

Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media?. In: Proceedings of the 19th international conference on World wide web, pp. 591–600. ACM, New York (2010)

Latour, B.: Science in action: How to follow scientists and engineers through society. Harvard university press, Cambridge (1987)

Latour, B.: Reassembling the social: an introduction to actor-network-theory., by Bruno Latour, Oxford University Press, New York (2005)

Law, J.: Technology and heterogeneous engineering: the case of Portuguese expansion. In: Bijker, W.E., Hughes, T.P., Pinch, T.J. (eds.) The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology, pp. 111–134. MIT Press, Cambridge (1987)

Lazer, D., Pentland, A.S., Adamic, L., Aral, S., Barabasi, A.L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., Van Alstyne, M.: Computational social science. Science **323**(5915), 721–723 (2009)

Lazer, D., Kennedy, R., King, G., Vespignani, A.: Big data. The parable of Google Flu: traps in big data analysis. Science (New York, NY) **343**(6176), 1203–1205n (2014)

Lieberman, E.S.: Nested analysis as a mixed-method strategy for comparative research. Am. Polit. Sci. Rev. **99**(03), 435–452 (2005)

Lin, N.: Social Capital: A Theory of Structure and Action. Cambridge University Press, London (2001)

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I.: The Arab Spring| the revolutions were tweeted: information flows during the 2011 Tunisian and Egyptian revolutions. Int. J. Commun. **5**, 31 (2011)

Magaudda, P.: Hacking Practices and their Relevance for Consumer Studies: The Example of the 'Jailbreaking'of the iPhone. Consumers, Commodities and Consumption (2010)

Magaudda, P.: When materiality 'bites back': digital music consumption practices in the age of dematerialization. J. Consum. Cult. **11**(1), 15–36 (2011)

Magaudda, P.: What happens to materiality in digital virtual consumption? In: Denegri-Knott, J., Molesworth, M. (eds.) Digital Virtual Consumption. Routledge, London (2013)

Manovich, L.: Trending: the promises and the challenges of Big Social Data. In: Gold, M.K. (ed.) Debates in the Digital Humanities. University of Minnesota Press, Minneapolis (2012a)

Manovich, L.: How to compare one million images. In: Berry, D.M. (ed.) Understanding Digital Humanities, pp. 249–278. Palgrave Macmillan, Basingstoke (2012b)

MacKenzie D.: A sociology of algorithms: high-frequency trading, boundary work, and market configurations, Paper discussed at MaxPo Conference on Monday, March 3rd, SCOOPs Seminar MaxPo (2014)

McAfee, A., Brynjolfsson, E.: Big data. The management revolution. Harv. Bus Rev. **90**(10), 61–67 (2012)

McKinsey Global Institute, Big data: the next frontier for innovation, competition, and productivity, May 2011 http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation. Accessed 29 May 2014

Morgan, D.L.: Paradigms lost and pragmatism regained methodological implications of combining qualitative and quantitative methods. J. Mixed Methods Res. **1**(1), 48–76 (2007)

Neuhaus, F., Webmoor, T.: Agile ethics for massified research and visualisation. Inf. Commun. Soc. **15**(1), 43–65 (2012)

Ohm P.: The underwhelming benefits of big data, University of Pennsylvania Law Review, 161 U. Pa. L. Rev 339, (2013)

Omernick, E., Sood, S.O.: The impact of anonymity in online communities. In: Social Computing (SocialCom), 2013 International Conference on (pp. 526–535). IEEE (2013)

Pang, B., Lee, L.: Opinion mining and sentiment analysis. Found. Trends Inf. Retr. **2**(1–2), 1–135 (2008)

Pew Research Center, Older Adults and Technology Use Available at: http://www.pewinternet.org/2014/04/03/older-adults-and-technology-use/ (2014). Accessed 1st Aug 2014

Pickering, A.: The Mangle of Practice: Time, Agency, and Science. University of Chicago Press, Chicago (2010)

Rogers, R.: Digital methods. MIT press, Cambridge (2013)

Rogers, R., Marres, N.: Landscaping climate change: a mapping technique for understanding science and technology debates on the World Wide Web. Public Underst. Sci. **9**(2), 141–163 (2000)

Schwarz, O.: Good young nostalgia: camera phones and technologies of self among Israeli youth. J. Consum. Cult. **9**(3), 348–376 (2009)

Schwarz, O.: Going to bed with a camera: on the visualization of sexuality and the production of knowledge. Int. J. Cult. Stud. **13**(6), 637–656 (2010)

Smith, A., Brenner, J.: Twitter use 2012. Pew internet and american life project, 4. http://pewinternet.org/Reports/2012/Twitter-Use-2012.aspx (2012). Accessed 9th Aug 2014

Snijders, C., Matzat, U., Reips, U.: Big data: big gaps of knowledge in the field of internet science. Int. J. Internet Sci. **7**(1), 1–5 (2012)

Tinati, R. et al.: Big data: methodological challenges and approaches for sociological analysis. Sociology 1–19 (2014)

Tufekci Z.: Big questions for social media big data: representativeness, validity and other methodological pitfalls. In: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (2014)

Twitter: One hundred million voices, Twitter blog, http://blog.twitter.com/2011/09/one-hundred-million-voices.html (2011). Accessed 12 Sep 2011

Venturini, T.: Diving in magma: How to explore controversies with actor-network theory. Public Underst. Sci. **19**(3), 258–273 (2010)

Wellman, B.: Designing the Internet for a networked society. Commun. ACM **45**(5), 91–96 (2002)

Wellman, B.: The three ages of internet studies: ten, five and zero years ago. New Media Soc. **6**(1), 123–129 (2004)

Wellman, B.: Studying the internet through the ages. In: Consalvo, M., Ess, C. (eds.) The Handbook of Internet Studies. Wiley, Oxford (2011)

Wellman B., Gulia M.: Net surfers don’t ride alone: virtual communities as communities. Netw. Glob. Village 331–366 (1999)

Zikopoulos, P., Eaton, C.: Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. McGraw-Hill Osborne Media, New York (2011)