Background: A Social Framework for Big Data

1. Introduction

This document is an outcome of an Economic and Social Research Council (ESRC) funded project that took place from June 2013 to Sept 2014. *Socialising Big Data* was an interdisciplinary collaboration between social scientists from a range of backgrounds (sociology, anthropology, and science and technology studies), many of whom are or were affiliated with the Centre for Research on Socio-Cultural Change (CRESC Manchester and The Open University) and the Centre for Economic and Social Aspects of Genomics (Cesagen Lancaster), but since expanded to include other institutions.[1] The project aimed to advance the social scientific understanding of Big Data to benefit academics, students, practitioners and policy makers. It did this through *collaboratories* with practitioners from three specific contexts with different trajectories, understandings and working relations to Big Data – genomics, national statistics and waste management.

Based on the findings of the collaboratories, the Project Team completed a working paper, *Socialising Big Data: From concept to practice* and developed this Social Framework for Big Data.[2] The framework arose out of a concern that technical, legal, economic and political frameworks for understanding Big Data do not attend to broader social meanings, which are necessary in order to address potentially intractable policy problems such as data ownership and the ethics of data use. If Big Data will replace other ways of knowing and provide the basis for new kinds of evidence, then understanding its distinctively social implications is vital for informing the actions and decisions of government, business and researchers.

2. Context

In a very short time what was initially referred to as the data deluge, information overload or tsunami of data has come to known as 'big data.' While variously defined, Big Data generally refers to digital content stored in social, commercial, scientific, and governmental databases and often generated as a by-product of digital transactions, communications, interactions, and so on.[3] According to the most popularly referenced definition, what makes this data distinctive is not only its *volume* but its *velocity* of generation (the speed of collecting data in 'real time') and *variety* of data sources and formats (increasing array of data types from audio, video, and image data, and the mixing and linking of information collected from diverse sources). It is in relation to the accumulation of Big Data that a number of national and international government digital policies, agendas and frameworks have been written that speak to its values and potentialities. They typically have two key foci: Big Data as a driver of *economies* and as a driver of *societies*. On the former, Big Data are referred to as a resource with qualities to be mined and capitalized, the new oil to be tapped to spur economies: 'Data has become a key asset for the economy and our societies similar

---

[1] PI: Evelyn Ruppert, Goldsmiths, University of London. Co-Is: Penny Harvey, Manchester, CRESC; Celia Lury, Warwick; Adrian Mackenzie, Lancaster; Ruth McNally, Anglia Ruskin. Researchers: Stephanie Alice Baker, Goldsmiths, University of London; Yannis Kallianos and Camilla Lewis, University of Manchester, CRESC.

[2] Ruppert E, Harvey P, Lury C, et al. (2015) 'Socialising Big Data: From Concept to Practice.' CRESC Working Paper 138. Available at http://www.cresc.ac.uk/publications/working-papers/. Socialising Big Data Project (2015) *Background: A Social Framework for Big Data.* CRESC: The University of Manchester and The Open University.

[3] This definition is from Stapleton, Lisa K. 2011. Taming Big Data. *IBM Data Management*. Kitchin adds the following qualities: exhaustive in scope (e.g., covering 'whole populations'); fine-grained in resolution and uniquely indexical; relational by being made up of common fields that enable linking; and flexible and scalable.[3] The growing list of qualities attests to the diversity of what is being defined as Big Data but also that the relevance and degree of each is highly variable depending on the particular data in question. Kitchin, Rob. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: SAGE.

to the classic categories of human and financial resources.'[4] Through a variety of practices of valuation, data is envisaged as something to be capitalized and traded like other commodities, and government digital agendas seek to facilitate such valuations through policies such as open data and open access policies. As a driver of societies, agendas and frameworks address issues of access and skills, public service provision, development, sustainability, healthy and smart living, trust, privacy and cybersecurity: 'Having digital skills and knowledge is just as important as reading, writing and arithmetic in today's society'.[5] Social values then are about promoting 'a people-centered Big Data revolution' by leveraging data to 'improve decisions and empower people.'[6]

These agendas and frameworks are important in providing particular perspectives on the potential of Big Data to generate social goods. But their economic and societal valuations of social goods inevitably encounter opposing objectives or 'social bads': openness and sharing of data compete with concerns about privacy and data rights; better governing decisions through data compete with concerns about data surveillance and control; valuations and the commodification of data compete with concerns about data ownership and consent; and so on. It is our contention that these competing objectives are in part a consequence of a narrow utilitarian framing of what is *social* about Big Data and the social goods it can deliver. Instead, we propose a broader understanding of *both* the social and of Big Data. Put succinctly our understanding recognises that Big Data involves social relations: it is inherently social because it is a product of and has a capacity to establish social relations. As we set out below, this calls for reframing conventional debates that typically focus on individual rights and ownership to an ethic that recognises the connectedness and interdependent relations that make up Big Data.

3. Socialising through Big Data

One of the starting hypotheses of our Project was that Big Data does not have one meaning but has multiple histories and contexts of use, which are becoming part of the formation of what Big Data *is*. It is for this reason that we initially organised collaboratories that brought together a diverse set of social scientists, national statisticians, bioscientists and waste management practitioners. The collaboratories confirmed that Big Data is multiple in part because it is emergent and in the process of being shaped and composed in myriad ways by diverse social and technical practices of collection, analysis, interpretation and storage.

At the same time, Big Data is the very thing that enabled our diverse group to come together. That is, Big Data connected us socially and in this way constituted a boundary object between our multiple interpretations and contexts. Such multiplicity though was not incidental but key to recognising and managing tensions across our different contexts because a single definition did not need to be settled or stabilised.[7] Indeed, tensions were productive in opening up different ways of thinking but also corresponded to the nature of our object of interest, which is multiple, unstable and changing. Yet, while there were great differences in histories and trajectories of working with Big Data across our contexts, we identified common themes such as the economies and ethics of Big Data.

But secondly and perhaps more significantly was our collective recognition that Big Data not only brought us together but also brings others together across diverse practical contexts because of its capacities to be mixed, divided, and reused for myriad purposes and ends. This is in

---

[4] European Commission (2015) *Digital Agenda for Europe: Making Big Data work for Europe*. Available at: http://ec.europa.eu/digital-agenda/en/big-data.

[5] Ibid.

[6] See for example, the Data-Pop Alliance where the social value is in the uses of Big Data for especially humanitarian and development oriented policies: http://www.datapopalliance.org.

[7] We took up the concept of boundary object from Bowker GC and Star SL. (1999) *Sorting Things Out: Classification and its Consequences,* Cambridge, Massachusetts: The MIT Press.

part captured in proposals to centralise patient records, which are described as part of an 'era of socialised big NHS data' that can benefit the health of both societies and individuals.[8] However, we extended socialising to include the different actors and technologies involved in its generation (digital platforms, mobile devices, sensors, sequencers), formatting (cleaned, linked, packaged, stored, curated) and analysis (mined, visualised, correlated). In this way, Big Data connects myriad distributed people (computer scientists, data handlers, mathematicians, platform designers) and technologies (computers, devices, software, algorithms). In the social sciences these distributed associations are sometimes referred to as sociotechnical arrangements through which people get connected and related to each other.

People are also attached to and socialised by Big Data in another way.[9] Through their bodies and their actions, interactions and transactions they are part of sociotechnical arrangements that generate Big Data and through which they become data subjects. That is, people are part of, attached to, and become subjects through sociotechnical arrangements such as social media platforms, sensors and genomic sequencers, which then come to generate Big Data. But there is a second aspect of this: both data subjects and sociotechnical arrangements are formed and changed through their mutual attachment. Platforms such as social media or search engines are calibrated and recalibrated in relation to what subjects do, and subjects adjust and change their actions and interactions in relation to those calibrations. There are, in other words, feedback loops between the two and Big Data is an outcome of these relations. Put simply, subjects and sociotechnical arrangements do not exist without the other and change and modulate in relation to each other.[10]

It is also through Big Data that subjects get attached and connected to each other. Being socially attached through Big Data happens in a variety of ways. People can identify affiliations and form communities of mutual support through biological and cultural data; from genomic to social media data, people can identify with and become attached to each other. Researchers, businesses and governments can also identify attachments between people. Indeed, the predominant analyses of Big Data are not focused on the data of specific subjects or their identities but on patterns in Big Data through which networks, groups, profiles and publics can be identified.[11] While there can be much uncertainty about the validity, veracity, meaning and implications of these attachments (which can only be evaluated in relation to specific instances) the point is that Big Data has the potential to join up and connect people socially in new and novel ways. At the same time, attachments can and do come to have an impact on people through targeted interventions or the differential treatment of identified groups.

Uses of Big Data are thus double-edged. In medicine, genetic profiles can aid in the identification of risk and in turn improve interventions; in social or environmental policy, collective benefits such as efficiency and improvements can be achieved through services targeted to particular identified groups or communities; etc. However, they can also lead to potential discriminatory, manipulative, and stigmatising practices; in consumer finance risky groups can be identified and denied credit; in social policy, specific communities associated with particular behaviours can be targeted for increased police surveillance. Furthermore, social attachments

---

[8] Bell, Alice. 2014. Why You Should Be Angry About Changes to NHS Patient Data Policy. *Guardian Comment Network*, 20 January 2014 [cited 4 February 2014]. Source: http://bit.ly/1KnpOvO.

[9] Our use of the term attachment draws from how it has been defined in the study of markets and consumers. Products and consumers do not exist as separate and independent entities but are generated and changed in relation to each other through their embeddedness in and attachment to sociotechnical arrangements. See Callon, Michel and Law, John 2005. "On qualculation, agency, and otherness." *Environment and Planning D: Society and Space* 23(5) 717 – 733.

[10] This challenges assumptions that device and platform owners are the designers and makers of data and merely collect and thus own it.

[11] Indeed, digital actions are often disconnected from identifiable subjects, which is the working logic of data mining where inferences and predictions are made based on associations between actions across multiple aggregated data sets.

through Big Data can possibly undo or devalue other ways in which people get connected or identified. These issues suggest that Big Data has social consequences such as its constitutive (identifying) and distributive (targeting) effects that are typically not addressed by utilitarian valuations of social goods.

4. Data Socialities and their implications

The double-edge is in part a consequence of three types of relations described above and which we summarise here as *data socialities*:[12]

- Relations between people and sociotechnical arrangements generate and get attached to Big Data.
- Relations between people can be created out of patterns of similar attachment to Big Data.
- Relations between people are practically created across diverse contexts when they share, re-use, mix, engage and analyse Big Data.

Socialities are thus generated by people's mutual use and attachments to Big Data. They introduce different understandings of social goods than those typically identified in digital agendas and frameworks. These agendas and their attendant debates are documented in numerous policy papers and proposals, which we do not seek to review or critique here.[13] Instead we focus on how data socialities highlight that Big Data is inherently social because it comes out of and has a capacity to establish social relations. The implications are that conventional debates need to be reframed towards an ethic of care that recognises connectedness or interdependence.[14] We suggest in a *preliminary* way how data socialities could reframe debates on two questions:

1. If Big Data is a product of interdependent and connected relations then who has rights over it?
    a. Data ownership is usually reserved for platform owners and people are treated as data subjects rather than interdependent co-producers. When the data rights of subjects are promoted then rights are considered the property of individuals. This is the working logic of digital platforms that require consent to data collection, analysis and trading of data as conditions of use. Such conditions are often buried in end-user license agreements with limited opt-in/opt-out possibilities, and require inordinate effort or do not apply to all possible forms of sharing and re-use.
    b. Data socialities suggest that data subjects are already and always attached to Big Data through their bodies and everyday actions and that Big Data is an interdependent and collective accomplishment and good.[15]
    c. Big Data brings people, social groups, organisations and institutions into new proximities and relations and the social goods of Big Data can be most effectively

[12] Data socialities picks up from Paul Rabinow's term 'biosocialities', which captures how biomedical knowledge shapes the making of social identities through not only biological but also social forms of association and attachment Rabinow P. (1996) Artificiality and enlightenment: from sociobiology to biosociality. *Essays on the Anthropology of Reason*. Princeton, NJ: Princeton University Press.

[13] For instance, Alex Pentland at MIT, proposes a 'New Deal on Data' to 'rebalance of the ownership of data in favor of the individual whose data is collected.' Data is understood as individual rather than socially interdependent and engages subjects as customers and owners. Harvard Business Review Staff (2014) 'With Big Data Comes Big Responsibility,' *Harvard Business Review*. Retrieved 21 April 2015 (https://hbr.org/2014/11/with-big-data-comes-big-responsibility).

[14] The ethic of care derives from feminism, in particular Carol Gilligan's 1982 book, *In a Different Voice: Psychological Theory and Women's Development*, Harvard University Press. An ethic of care offers an alternative approach to mainstream ethics through a theory of the self as relational and in a web of interconnection. The ethic of care we outline for Big Data attends to issues of power, relationships, responsibility and experience.

[15] For example, that call data records belong to mobile operators has been disputed: see Letouze E and Vinck P. (2014) *The Politics and Ethics of CDR Analytics* (draft). New York: Data-Pop Alliance.

realised through practices of sharing, mixing and combining data. As such, all forms of ownership can have the consequence of disconnecting data from not only existing but also future relations. Generating the potential social goods of Big Data thus calls for opening data to governments, businesses, researchers and individuals, who are part of its value-making. Such openings would not only be generative of new analyses and insights but also contribute to the development of trust.

How might data rights - including those of consent and data protection - be reframed to recognise that Big Data is a product and generative of connected and interdependent relations?[16]

2. If Big Data has the capacity to be generative of new social relations then who is accountable for ethical effects?
    a. Data rights are often reduced to questions of confidentiality and data protection; if data does not disclose identity and cannot be linked to an individual then data rights have been addressed. De-personalised data in other words is presumed to resolve ethical issues though de-anonymisation continues to be a major concern as a result of the possibility of joining up ever-increasing datasets that can lead to re-identification.
    b. Because data subjects can be attached to each other through Big Data, group identifications can lead to adverse group profiles and effects.[17] Even if a subject can opt-out this does not protect them from being subject to these effects, and opting out can itself become another attribute to be modelled. If data are the 'raw material for accountability' then securing and making accountability transparent are key.[18]
    c. That sociotechnical arrangements to which Big Data are attached involve myriad distributed people and technologies challenges the possibility of identifying and allocating responsibility and accountability for social effects and consequences.

How might ethical responsibilities for the making of social relations through Big Data be reframed and accountability and answerability for effects be attributed?

Collectively, data socialities suggest that power relations and questions of ownership between platform owners, data subjects, and data users over the generation, circulation and analysis of Big Data need to be reframed in ways that go beyond existing digital policies, agendas and frameworks.

4. Conclusion

To summarise, data socialities arise when people share and reuse Big Data, when they get attached to Big Data and when they get attached to each other through Big Data. Each of these has meanings specific to particular practices and contexts but generally speaking they establish a fundamentally different understanding of what is the social of Big Data. By not attending to these socialities and the transparency, accountability and collectivising that they demand, utilitarian pursuits of social 'goods' (better health, environment, economy) will be generative of intractable problems such as distrust. Rather, it is through an ethic of care and social ownership that recognise the connectedness and interdependence of people inherent in Big Data that collective goods can be both recognised and realised.

[16] 'Data cooperatives' for example have been proposed for the sharing of genomic data: www.midata.coop.
[17] For this reason Andrej Zwitter suggests that Big Data leads to questions of group privacy. Zwitter A. (2014) 'Big Data ethics', *Big Data & Society* 1: 1-6. DOI: 10.1177/2053951714559253.
[18] United Nations Independent Expert Advisory Group (2014) *A World that Counts*. Available at: http://www.undatarevolution.org/report/.