The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts

Article *in* Science and Engineering Ethics · May 2015

DOI: 10.1007/s11948-015-9652-2 · Source: PubMed

Brent Mittelstadt
University of Oxford

Luciano Floridi
University of Oxford

Sci Eng Ethics
DOI 10.1007/s11948-015-9652-2

REVIEW PAPER

The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts

Brent Daniel Mittelstadt[1] · Luciano Floridi[1]

Received: 11 February 2015 / Accepted: 13 May 2015

Abstract The capacity to collect and analyse data is growing exponentially. Referred to as 'Big Data', this scientific, social and technological trend has helped create destabilising amounts of information, which can challenge accepted social and ethical norms. Big Data remains a fuzzy idea, emerging across social, scientific, and business contexts sometimes seemingly related only by the gigantic size of the datasets being considered. As is often the case with the cutting edge of scientific and technological progress, understanding of the ethical implications of Big Data lags behind. In order to bridge such a gap, this article systematically and comprehensively analyses academic literature concerning the ethical implications of Big Data, providing a watershed for future ethical investigations and regulations. Particular attention is paid to biomedical Big Data due to the inherent sensitivity of medical information. By means of a meta-analysis of the literature, a thematic narrative is provided to guide ethicists, data scientists, regulators and other stakeholders through what is already known or hypothesised about the ethical risks of this emerging and innovative phenomenon. Five key areas of concern are identified: (1) informed consent, (2) privacy (including anonymisation and data protection), (3) ownership, (4) epistemology and objectivity, and (5) 'Big Data Divides' created between those who have or lack the necessary resources to analyse increasingly large datasets. Critical gaps in the treatment of these themes are identified with suggestions for future research. Six additional areas of concern are then suggested which, although related have not yet attracted extensive debate in the existing literature. It is argued that they will require much closer scrutiny in the immediate future: (6) the dangers of ignoring group-level ethical harms; (7) the importance of epistemology in assessing the ethics of Big Data; (8) the changing nature of fiduciary relationships that become increasingly data saturated; (9) the need to distinguish between

✉ Brent Daniel Mittelstadt
brent.mittelstadt@oii.ox.ac.uk

[1] Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, UK

Published online: 23 May 2015

‘academic’ and ‘commercial’ Big Data practices in terms of potential harm to data subjects; (10) future problems with ownership of intellectual property generated from analysis of aggregated datasets; and (11) the difficulty of providing meaningful access rights to individual data subjects that lack necessary resources. Considered together, these eleven themes provide a thorough critical framework to guide ethical assessment and governance of emerging Big Data practices.

Keywords Ethics · Big data · Bioethics · Information ethics · Medical ethics · Ethical foresight

Introduction

The amount of data being amassed by humanity is growing exponentially (Bail 2014, p. 465). Digital technologies, including online services and emerging ubiquitous computing devices, can track behaviour to a greater degree than ever possible (Markowetz et al. 2014). At the same time, policies as well as ethical, social and legal understanding of such “technological capabilities to merge, link, re-use and exchange data” lag behind the growth of technical capacities for data storage and analysis (Collingridge 1980; Safran et al. 2006, p. 6), the potential benefits of which are already being hailed in mass media (Markowetz et al. 2014, p. 407). Big Data (see below) has contributed to the definition of modern life as the ‘information age’.

The technologies producing and processing data “provide destabilising amounts of knowledge and information which lack the regulating force of philosophy which…ensures that institutions remain rational.” (Berry 2011, p. 8). New examples of the problems faced by Big Data systems appear regularly in mass media. Facebook’s Beacon software, which was rolled out in 2007 to connect automatically external purchases to Facebook profiles, provided an early example of the ethically problematic nature of linking datasets. The service, intended to improve personalised advertising, inadvertently revealed sensitive characteristics of a person’s life such as sexual preference (Oboler et al. 2012a, p. 7). Big Data can similarly ground controversial forms of research, as demonstrated by the much discussed Facebook ‘emotional contagion study’ in 2012 (Schroeder 2014) A more recent and as-of-yet comparatively uncontroversial example are location awareness systems such as Apple’s iBeacon software which connects information from a user’s Apple profile to in-store systems and advertising boards, allowing for a ‘personalised’ shopping experience and tracking of (profiled) customers within physical stores (Apple 2014). Such systems effectively link online and offline personalities, supporting an “onlife” environment that may become invasive (Floridi 2014a).

As shown by these attention-grabbing examples, increasing interconnectivity in data-rich contexts can challenge accepted social and ethical norms. Practices centred on the mass curation and processing of personal data can quickly gain a negative connotation which, in a way similar to what has happened in the public debate over genetically modified organisms (cf. Devos et al. 2008), places potentially beneficial applications at risk through association with problematic applications. A 'whiplash effect' can occur, by which overly restrictive measures (especially legislation and policies) are proposed in reaction to perceived harms, which overreact in order to re-establish the primacy of threatened values, such as privacy. Such a situation may be occurring at present as reflected in the debate on the proposed European Data Protection Regulation currently under consideration by the European Parliament (Wellcome Trust 2013), which may drastically restrict information-based medical research utilising aggregated datasets to uphold ethical ideals of data protection and informed consent (see "Informed Consent").

Ethical foresight may reduce the probability of 'regulatory whiplash' by informing public debate through improved understanding of the 'moral potential' of emerging technological applications and data practices. To contribute to this process, a systematic and comprehensive review of academic literature discussing the ethical implications of Big Data was conducted to identify the issues of emerging importance for this novel form of data curation and analysis. "Background" section provides a brief background on 'Big Data' as a concept. "Methodology" section describes the methodology of a systematic and comprehensive review of academic literature discussing the ethics of Big Data, before presenting a narrative synthesis of the "Results" section. Shortcomings and further relevant issues not currently addressed sufficiently in the literature are then highlighted in "Discussion" section, before concluding by reflecting on directions for further research in "Conclusion" section.

Background

'Big Data' covers a vast variety of phenomena focused on the analysis of large datasets. Data types and applications can be found in areas such as intelligence analytics (Mahajan et al. 2012), behaviour and preference modelling (Coll 2014, p. 1257; Lomborg and Bechmann 2014), sustainability studies (Mahajan et al. 2012), online and offline commerce, biomedical research and healthcare, and various other forms of scientific and social research and commercial pursuits (Costa 2014; Markowetz et al. 2014) based around mining vast datasets (Bail 2014, pp. 466–7). Data can be quantitative and textual, much of which is now user-generated via online behaviour that is revealing in terms of personal preferences and behaviours (Puschmann and Burgess 2014, p. 1694). The perceived value of such practices is variable, and may stem from characteristics such as the ability to collect data for research 'unintrusively' or perhaps, covertly (e.g. Lomborg and Bechmann 2014, p. 256), to track and profile fine-grained behaviours, preferences and other characteristics (e.g. sexual orientation or political opinions) of individuals (Coll 2014, p. 1257; Mahajan et al. 2012; Pariser 2011), to predict future behaviour (as used in law enforcement or credit, insurance and employment screening); or more broadly to search for connections across vast datasets for a variety of research purposes (Floridi 2012).

Against this broad context, biomedical Big Data has gained significant attention due to a combination of two factors. On the one hand, there is the huge potential to advance the diagnosis, treatment, and prevention of diseases as well as foster healthy habits and practices (Costa 2014). On the other hand, there is the obvious, inherent sensitivity of health-related data and the implicit vulnerability and needs of those potentially requiring treatments (Pellegrino and Thomasma 1993). Academically and commercially[1] valuable biomedical big data can exist in many forms, including aggregated clinical trials (Costa 2014), genetic and microbiomic sequencing data[2] (Mathaiyan et al. 2013; McGuire et al. 2008; The NIH HMP Working Group et al. 2009), biological specimens, electronic health records and administrative hospital data.[3] Such data can be held in biobanks, cyberbanks and virtual research repositories[4] (Costa 2014, p. 436; Currie 2013; Majumder 2005, p. 32). Compared with traditional forms of storage, such repositories tend to assemble aggregated datasets explicitly for research purposes with "virtually unlimited opportunities for data linkage and data-mining" (Prainsack and Buyx 2013, p. 73) due to the sheer scale of the datasets (Steinsbekk et al. 2013, p. 151).

Data can also be generated explicitly or covertly via social media applications and health platforms[5] (Costa 2014; Lupton 2014, p. 858), emerging 'personal health monitoring' technologies (Mittelstadt et al. 2011, 2013) including wearable devices (Boye 2012), home sensors (Niemeijer et al. 2010) and smart phone applications, and online forums and search queries. The latter, for example, enable public health and outbreak tracking[6] (Butler 2013; Costa 2014, p. 435). Other data come from 'data brokers' which collect, process, store and sell intelligence based on a variety of medical and health-related data sourced from social media, online purchases, insurance claims, medical devices and clinical data provided by public health agencies and pharmacies, among others (Terry 2012, 2014).

Analysis of these data types can be undertaken for numerous purposes, including development of clinically useful predictive models (Choudhury et al. 2014, p. 3), longitudinal and cross-sectional effectiveness and interaction studies of pharmaceuticals (Tene and Polonetsky 2013, p. 246), and long-term 'personal health monitoring' (Boye 2012; Mittelstadt et al. 2014; Niemeijer et al. 2010). Broadly, these data may foster understanding of health disorders and the efficiency and

[1] For example, the identification of the presence of diabetes can support targeted marketing (Terry 2012, p. 392).

[2] For an overview of sample companies providing such services, see Costa 2014.

[3] In some contexts, such as the USA under Health Insurance Portability and Accountability Act, administrative data will be afforded less protection than genomic and similar biobank data despite possessing similar capacities for revealing sensitive aspects of a person's health. This may be due partly to the possibility of removing identifiers from administrative data without 'ruining' the data (Currie 2013) as is an apparent limitation with anonymisation of genomic data (Hansson 2009, p. 10).

[4] These forms of biomedical data are incredibly varied and complex, consisting of data produced from a wide variety of sources, including "laboratory auto-analyzers, pharmacy systems, and clinical imaging systems…augmented by data from systems supporting health administrative functions such as patient demographics, insurance coverage, financial data, etc.…clinical narrative information, captured electronically as structured data or transcribed 'free text'…electronic health records" to name but a few (Safran et al. 2006, p. 2).

[5] For instance, Facebook has recently announced plans for "support communities" and "preventative care applications" (Reuters 2014), while Google and Apple have recently released platforms for health and fitness data aggregation (Google Fit and Apple HealthKit/ResearchKit).

[6] However, the efficacy of such platforms remains questionable (Butler 2013).

effectiveness of treatments and health systems and organisations. They also create repositories for public health and information-based research (Safran et al. 2006, p. 2; Steinsbekk et al. 2013, p. 151). With that said, clinical applications are not guaranteed (Lewis et al. 2012). While promising on many fronts, biomedical Big Data, and the findings derived from it, may raise a host of ethical concerns stemming from the sensitivity of data being manipulated and the seemingly limitless potential uses and repurposing, and implications of data that concern individuals as well as groups.

Methodology

In order to understand what ethical issues have already been identified and discussed in the context of Big Data, a comprehensive and systematic meta-analysis of academic literature was conducted in October 2014. Six databases were searched (Web of Science, Scopus, Global Health, Philpapers, PubMed and Google Scholar) to identify literature discussing ethical aspects of Big Data. Search terms (with wildcards) were chosen to limit the review to articles explicitly mentioning 'Big Data' and ethics or morality, rather than searching for individual related concepts such as 'biobanks' or 'informed consent', thus allowing for a comprehensive review of Big Data literature. In recognition of the prevalence of biomedical applications, searches were divided between 'Big Data' and 'biomedical Big Data' to facilitate comparison. A breakdown of the search by database, search terms and results returned can be found in Table 1.

The title and abstract of each returned article was reviewed by the authors to determine relevance. Inclusion was based solely on the discussion of ethical issues in the article, with the goal of identifying themes in the literature. Limitations were not placed on the quality or length of the discussion, but rather on the mere presence of ethical concepts and issues. Further sources were also located through hand-searching and backtracking of citations provided within the reviewed articles.

The search was limited to English language articles. Although most of the reviewed literature consisted of peer-reviewed journal articles, other types of publications including commentaries, working reports, white papers and scientific books were also located. Date restrictions were not enforced.

Data Analysis

Each article was analysed and key passages highlighted for further interpretation and grouping into themes existing across multiple sources. These themes were allowed to emerge from the literature rather than starting from a pre-defined theoretical framework. However, the review was intended to address two questions:

(1) How is 'Big Data' conceptualised within discussions of its ethics?
(2) What types of ethical issues are raised by Big Data?

Table 1 Search queries

Database	Search string	Returned
Ethics of biomedical big data		
Web of science	*TOPIC*: ((ethic* OR moral*) (health* OR *medic* OR bio*) "big data")	23
Scopus	*TITLE-ABS-KEY*: ((ethic* OR moral*) AND (health* OR *medic* OR bio*) AND "big data")	18
Global health	"Big data" AND (ethic* OR moral*)	145
PubMed	"Big data" AND (ethic* OR moral*)	19
Ethics of big data		
Web of science	*TOPIC*: ((ethic* OR moral*) "big data" NOT (*medic* OR health* OR bio*))	18
Scopus	*TITLE-ABS-KEY*: ((ethic* OR moral*) AND "big data" AND NOT (*medic* OR health* OR bio*))	28
Philpapers	"Big data"	19
Google scholar	"Big data" ethics OR ethical OR ethic OR moral OR morality OR morals	50[a]
Philosopher's index	"Big data"	6

[a] 11,000 Returned, first 50 reviewed

To start, phrases and passages were highlighted that appeared to refer to ethical issues or concepts, understood as areas of 'right' and 'wrong' or the clash of competing values or normative interests among stakeholders. Highlighted segments were then coded to reflect the author's interpretation of the text (cf. Gadamer 2004; Patterson and Williams 2002). Similar codes were then grouped and assigned to ethical themes. Once themes had emerged from the literature, a second systematic analysis was performed using the NVivo 10 software package. All sources were re-checked via a text search for the presence of the themes that emerged. The following terminological convention was adopted in discussing 'Big Data stakeholders' below: 'data subject' refers to the individual described by the data, 'data custodian' refers to any individual or organisation responsible for hosting or archiving the data in either its individual or aggregated form, and 'data analyst' refers to any individual or organisation analysing the data, but not necessarily hosting it.

Results

A total of 365 non-unique sources were identified for review across the databases, with 78 title/abstract combinations reviewed in full. Rejected sources were either off-topic or duplicates as determined by assessing the title and, in some cases, abstract. Once fully reviewed, a further ten were excluded for being off-topic, leaving 68 sources that met the inclusion criteria of explicitly discussing ethical aspects of Big Data.

In terms of the types of Big Data discussed, 36 sources primarily addressed 'biomedical' data, or data with a health or medical connotation. The remaining 32 sources primarily discussed non-medical types of Big Data, referred to as 'general' data. For article types, 43 described original research or in-depth analyses of 'peer review' quality, while 25 were 'commentary' sources, which included consultancy documents, editorials and opinion pieces, section introductions and other short pieces that do not always require peer review, empirical research, or extensive referencing. Only 7 of the 68 sources included empirical research. Finally, 23 of the 68 sources featured an in-depth discussion of ethics or ethical aspects of Big Data.

The results of the meta-analysis are presented as a narrative overview, which highlights and comments upon key themes and topics in the literature. A breakdown of all reviewed sources by ethical theme (see "Ethical Themes") and data type is provided in "Appendix 1". This overview is intended to address the two aforementioned questions in order to provide a starting point and reference for future discussions concerning the development, regulation, and ethical evaluation of Big Data practices.

Conceptualising Big Data

To identify how Big Data is conceptualised in ethics literature, it is necessary to consider how it is defined and which applications are discussed as posing potential ethical harms. A commonly accepted definition of 'Big Data' was not reflected in the literature. While definitions vary, some common characteristics and frameworks can be found. The first and most influential definition of Big Data was provided by Laney (2001) in terms of three dimensions: (1) Volume, or the scale of data; (2) Velocity, or the analysis of streaming data; and (3) Variety, or different forms of data. Later, another V was added: (4) Veracity, or the uncertainty of data (IBM 2014), giving rise to an influential framework (Andrejevic 2014; McNeely and Hahm 2014; Nunan and Di Domenico 2013). Accordingly, Big Data is unique in terms of the size and "speed of data generation and processing and the heterogeneity of data that can be dumped into combined databases" (Andrejevic 2014, p. 1676). Aggregation is justified by the idea that "things can be learned from a large body of data that cannot be comprehended from smaller amounts," revealing the implicit link between Big Data and complexity (McNeely and Hahm 2014, p. 305).

Broadly, Big Data can refer to (1) the *process* of analysing 'big' data sets, and (2) the *datasets* themselves. 'Big' can be defined variably in terms of quantities of electronic size (gigabytes, terabytes, petabytes, etc.), entries, individuals or events represented by the data, or alternatively in relation to the techniques and technologies currently available for analysis. The latter approach defines 'big' in procedural rather than quantitative terms, by connecting the size of the dataset to its complexity, understood in terms of the computational or human effort necessary for analysis (e.g. Costa 2014; Dereli et al. 2014; Fan and Bifet 2013; McNeely and Hahm 2014; National Science Foundation 2014; Terry 2012, p. 389). In other words, the data is 'Big' because it is difficult to sort and analyse with existing computing technologies.

While helpful for bridging the space between analysis processes and datasets, this approach suggests data that is ‘Big’ now may not be so in a year or a decade due to advances in computing technology and analysis procedures (Floridi 2012; Liyanage et al. 2014, p. 27). Although not semantically problematic (as adjectives describing technology tend to be relative, e.g. fast internet 10 years ago is slow internet today), this nevertheless poses a technological solution to an epistemological query by making the definition of ‘Big Data’ relative in relation to technical and analytical capacities. ‘Big Data’ becomes data that is difficult to analyse due to its size and complexity. This also suggests that more or better computing will enable us to ‘get ahead’ of the data and analyse all of it meaningfully again, as we did prior to the current era of Big Data. However, the exponential growth of data (Bail 2014, p. 465) suggests this is unlikely to occur, a point that further reinforces the view that Big Data describes a break with prior practice. Explicit consideration of historical context reduces the fluidity of the definition; in other words, labelling a study as ‘Big Data’ recognises the technical and analytical barriers faced at the time it occurred. Such fixed labelling may be important in ex-post ethical analysis (see “Ownership of Intellectual Property”).

Recognising these implications of a purely technical definition, it may be helpful to consider also the perceived value of Big Data as suggested in the types of analysis it allows. Boyd and Crawford (2012, p. 663) suggest Big Data is valuable due to the “capacity to search, aggregate, and cross-reference large data sets.” Similarly, according to Floridi (2012), a unique feature of Big Data is the possibility of identifying small patterns and connections in quantitatively large (and often aggregated) datasets. ‘Small patterns’ refer to connections between entries within the dataset, meaning connections are found within a subset of entries in a much larger dataset.

State of Deployment

As an emerging concept, defining fixed boundaries or practices which are ‘Big Data’ is perhaps impossible. Despite this, to avoid unbridled speculation over future ethical implications of Big Data practices, it is useful to consider first the state of development and deployment of different practices as reflected in the literature. Figure 1 shows a timeline describing the current state of various Big Data applications and practices categorised according to likelihood for deployment or commercial/research applications, ranging from: (1) real world, or currently in use; (2) on the horizon, with high likelihood of materialising due to the existence of the necessary technologies or data and empirically demonstrable motivation for use; or (3) not currently possible, meaning deployment may theoretically be possible but of limited likelihood or frequency due to access limitations or technical, ethical and other constraints. The boxes were populated to reflect Big Data technologies and practices currently in use, on the horizon in terms of research and development, and imagined but not yet possible. Each category and its position is reflected in the reviewed literature: for example, biobanks, cyberbanks, loyalty cards and personal health monitors are all widespread technologies, enabling large-scale data collection and assessment. At the other end of the scale, linking of offline medical records with

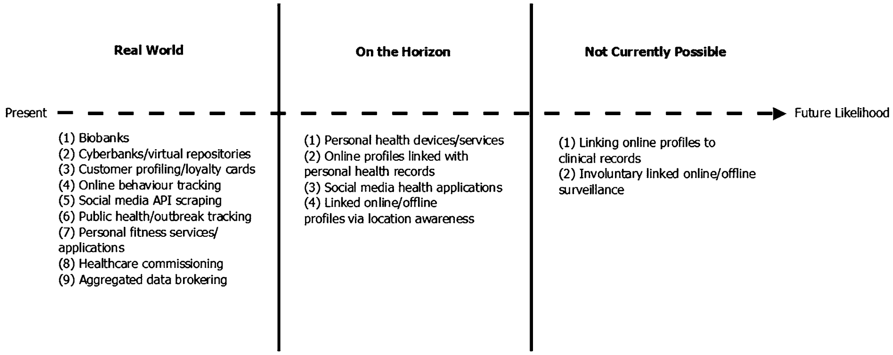
Real World
On the Horizon
Not Currently Possible
Present
Future Likelihood
(1) Biobanks
(2) Cyberbanks/virtual repositories
(3) Customer profiling/loyalty cards
(4) Online behaviour tracking
(5) Social media API scraping
(6) Public health/outbreak tracking
(7) Personal fitness services/ applications
(8) Healthcare commissioning
(9) Aggregated data brokering
(1) Personal health devices/services
(2) Online profiles linked with personal health records
(3) Social media health applications
(4) Linked online/offline profiles via location awareness
(1) Linking online profiles to clinical records
(2) Involuntary linked online/offline surveillance

Fig. 1 Estimated timeline of big data applications

online profiles (such as a Facebook user account) is currently seen as potentially valuable but unlikely due to ethical concerns, which would likely also preclude linked online/offline tracking of individuals without consent. With that said, the table is intended as an informed estimate of Big Data deployment rather than a description of the state-of-the-art. The latter was not deemed possible given the varied age and empirical evidence-base of the reviewed papers. Furthermore, technical development likely precedes commentary on social and ethical aspects (cf. Collingridge 1980; Moor 1985). Importantly, the figure reflects the likelihood of *widespread* deployment, meaning devices and uses currently in development, while testing or limited public release are considered 'on the horizon' rather than 'real world', e.g. profiles linked via location awareness, see (Apple 2014).

Ethical Themes

Through content meta-analysis five major ethical themes emerged from the literature: informed consent, privacy, ownership, epistemology, and the 'Big Data divide'. Interpretation and designation of themes were discussed and agreed upon by the authors. Although the ethical themes emerged according to frequency (Table 2), the overview does not merely highlight this frequency. Rather, the results discussed in the following sections were chosen for one of four reasons: (1) to draw attention to common interpretations of ethical themes and concepts, (2) to emphasise individual cases and issues that reveal unique ethical aspects of Big Data, (3) to highlight studies with an in-depth analysis of ethical concepts and issues, and (4) to identify gaps in the discussion in need of further research. The presentation of results therefore focuses on the authors' analysis and interpretation of the literature.

Informed Consent

Half of the literature addresses issues of informed consent. The concept (cf. Angrist 2009; General Medical Council 2008) does not cleanly transfer to research involving Big Data for a variety of reasons. Historically, consent is taken for

Table 2 Ethical themes

Theme	No of sources
Informed consent	34
Privacy	44
Anonymisation	20
Data protection	14
Ownership	12
Epistemology	14
Power/control (***Big data divide***)	22
Digital divide (***Big data divide***)	22

participation in a single study, not covering unrelated investigations resulting from sharing, aggregating, or even repurposing data within the wider research community (Choudhury et al. 2014, p. 4). This form of consent is problematic because Big Data is intended by design to reveal unforeseen connections between data points. This means that both what the data reveals about the subject and its utility in future research present greater uncertainty than normal at the time of consent. For example, secondary effects of pharmaceuticals can be identified by comparing data not only from multiple clinical trials, but 'informal sources' as well, such as incidental self-reporting via social media and search engine queries. In this type of research the connections that can be revealed through linking multiple data sets cannot be accurately predicted prior to carrying out the research. As a result, 'consent' cannot be 'informed' in the sense that data subjects cannot be told about future uses and consequences of their data, which are unknowable at the time the data is collected or aggregated.

Recognising this, calls to reform single-instance consent based on the belief that it is a barrier to 'necessary' research and innovation can be found in the debate (Larson 2013). 'Broad' and 'blanket' consent mechanisms, which pre-authorise future secondary analyses, are sometimes used in place of single-instance consent (Clayton 2005; Ioannidis 2013), if only due to the impracticality of renewing consent for each new analysis (Currie 2013; Lomborg and Bechmann 2014, p. 262). Such barriers often lead to biobanks employing a broad type of consent covering all future research activities. However, this approach has been recognised to limit the autonomy of data subjects (Master et al. 2014, p. 1). Tiered consent can also be used, which provides the data subjects with options for permitting specific uses of their data—for example to allow the data to be used in cancer research but not in genomic research—or to require specific re-consent for future uses rather than blanket consent for all potential uses (Majumder 2005, p. 33). Exclusion clauses can be used for a 'line-veto' type of tiered consent, which can increase confidence in data subjects that custodians are actually respecting their beliefs and values as translated into prohibitions of specific re-uses (Master et al. 2014). Where such formats are used, governance mechanisms, such as review councils and committees, help distinguish *'bona fide'* and problematic requests for access to data.[7] Note that

[7] See for example the UK Biobank Ethics and Governance Council: http://www.egcukbiobank.org.uk/.

differences in broad or blanket consent across national borders may also complicate sharing and re-use of data, an issue that is likely to become more pressing in the future, especially at the European level.

The impossibility of certainty concerning future uses of data highlights a key aspect of ‘Big Data’, namely the desire for openness and creativity in identifying novel connections between data sets. For data collected explicitly for aggregation into a ‘Big Data set’, the openness of the format does not create difficulties, although open data sharing may require a global type of consent, due to data travelling across the political and electronic borders of institutions and nations (Majumder 2005, p. 33). The same cannot be said for historical data for which consent was granted for a specific purpose. By the ideal of explicit single-instance consent such data should not be used without explicit consent for secondary uses by the person providing it. However, obtaining such consent years after a trial has been conducted can be very difficult, if not impossible (Clayton 2005; Wellcome Trust 2013). A tension therefore exists between the potential benefits of ‘Big Data’ analysis of historical datasets and the need for consent. Technical compromises are not obviously the solution—for instance, anonymising the data may not be sufficient to eliminate the need for consent due to the possibility of re-identification (Mello et al. 2013, p. 1653; Tene and Polonetsky 2013, p. 251; Terry 2014, p. 837). The temptation to conduct research on datasets beyond the boundaries of the initial consent agreement is at the heart of Big Data—in some cases, not only is consent no longer sought, but instead a ‘duty to participate’ in secondary research, thought of as ‘research that is not research’, is described (Ioannidis 2013, p. 40).

These two issues of consent are clearly applicable to data obtained from scientific (in particular, biomedical) research for which informed consent has long been standard. However, new issues are raised by the collection and analysis of data from potentially ‘unwilling’ participants, for example data scraped from social media platforms, smart phone applications, or open web forums (e.g. Krotoski 2012; Lomborg and Bechmann 2014; Markowetz et al. 2014). Terms of Service and other end-user agreements governing the usage of these applications tend to allow for collection, aggregation and analysis of such data. Indeed, social networking platforms such as Facebook and Twitter rely heavily upon advertising revenues generated through just such practices. However, as social scientific and other forms of research begin to utilise data collected from the unwilling or uninformed (cf. Enjolras 2014; Lazer et al. 2009), the lack of an explicit *informed* consent mechanism in end-user agreements gives cause for concern (Fairfield and Shtein 2014), even when ‘participants’ are ‘de-identified’ (Ioannidis 2013).

Modifying Consent for Big Data An unintended consequence of overly restrictive data protection and distribution policies is that a barrier can be erected to sharing data between researchers that is otherwise acceptable to data subjects (Choudhury et al. 2014, p. 5). In doing so, researchers may be missing opportunities to derive valuable information and innovations from the samples and data offered by research participants. Such a situation is currently materialising in relation to the European Data Protection Regulation under debate in the European Parliament. The regulation may severely restrict 'Big Data sciences' by requiring "specific, informed and explicit consent" for each instance of processing or analysis of 'personal data' (as specified in Article 83). The potential existence of Big Data research is therefore currently in jeopardy in Europe due to consent requirements (EURORDIS 2013; Wellcome Trust 2013). As argued by a consortium of biobanks, research councils and trusts (Wellcome Trust 2013), such a requirement creates barriers for 'big' datasets containing data from (hundreds of) thousands of individuals that are impractical and perhaps impossible to overcome in practice; indeed, most such repositories use blanket, broad or tiered consent.

If the Regulation were passed in its current form, each request by researchers for access to records held by biobanks would require re-contacting and re-consenting the data subject (EURORDIS 2013). In many cases, this will be impossible due to changes of contact details or death. Even where possible, it presents a significant financial and bureaucratic barrier to research (Wellcome Trust 2013). In this context, explicit, single-instance informed consent is causing rather than solving ethical problems by creating barriers for legitimate forms of research in addition to those rightly viewed as challenging, thus preventing researchers from advancing scientific knowledge, from deriving beneficial applications, and more generally from fulfilling the moral obligation to data subjects that have volunteered their time, bodies, and data for research.[8]

For the problems faced in Europe, these difficulties may be solved by introducing a number of clarifications and modifications to Article 83 of the Regulation, focusing mainly on exemption of pseudonymized data from the need for 'explicit' consent without which a "disproportionate regulatory burden" would be created, equivalent to that governing identifiable data (EURORDIS 2013; Wellcome Trust 2013, p. 2). More broadly, numerous approaches to consent have been proposed to overcome such barriers in purely information-based research, which re-use existing datasets (e.g. Prainsack and Buyx 2013; Rothstein and Shoben 2013; Schadt 2012). These 'fixes' tend to eliminate the need for consent to some degree in two main ways: pragmatically, by relying upon altruism, or substantively, by using an 'opt-out' approach, emphasising solidarity (see for example Prainsack and Buyx 2013) or the public good rather than individual autonomy (Rothstein and Shoben 2013).

Concerning pragmatic solutions, in the context of genomic sequencing, research is sometimes restricted to "information altruists" (Choudhury et al. 2014), or individuals willing to openly share their data (and sometimes, identity) on the basis that they possess the social status or economic resources to be sufficiently protected from future discrimination or harmful consequences. 'Radical honesty' models are similar, through which individuals volunteer de-identified genetic information for public sharing (Hayden 2012). Another approach is to establish "honest broker" and "stewardship" consent models by which impartial third parties mediate broad

[8] By some accounts moral obligations exist for medical research. As suggested by accounts of solidarity-based governance of biomedical Big Data (e.g. Prainsack and Buyx 2013), patients may have a moral duty to participate in research due to the value generated through advances in medical knowledge and treatments (Harris 2005; Schaefer et al. 2009). As participation inherently includes risks, researchers may similarly have a moral obligation to minimise risks as far as possible by extracting maximum value from existing datasets through re-purposing and aggregation (Currie 2013; Harris 2005).

consent agreements to protect the interests of data subjects (Choudhury et al. 2014, p. 7; Goodman 2014). Emphasising professionalism or enacting punitive measures for misuse of data can shift some of the burden to researchers benefiting from access to the data and promote feelings of responsibility to data subjects (Fairfield and Shtein 2014). The hope here is that forbidding unacceptable forms or research, such as re-identification of anonymised data, will minimise potential negative impacts on data subjects (Hayden 2012, p. 314).

Concerning substantive solutions, an 'opt-out' approach to consent (e.g. Hoffman 2014; Rothstein and Shoben 2013; Tene and Polonetsky 2013; Terry 2012) should not be seen as ethically equivalent to informed consent. Opt-out consent models may take advantage of people in vulnerable moments (Hayden 2012), for example if consent is taken during a clinical encounter in which the data subject is seeking treatment (cf. MacIntyre 2007; Pellegrino and Thomasma 1993). However, the weaknesses of such approaches do not suggest that explicit consent for each instance of data use is the correct path either; rather, a revision of ethical standards which strikes a balance between the requirement for consent and the practical requirements of 'Big Data science' may be appropriate. Tene and Polonetsky (2013, p. 262) suggest as much in calling for debate on the "merits of a given data use" as a broader societal issue, wherein distinctions can be drawn between 'types' of data uses requiring full informed, opt-in, opt-out or no consent at all.

It may be possible to reduce or eliminate the need for consent by focusing on the concept of solidarity and the alleged reduced risks to data subjects in data-based research. Prainsack and Buyx (2013) suggest that a solidarity-based approach to biobank governance, focused on harm mitigation, can be used in place of informed consent,[9] which recognises an empirically supported sentiment among the general public (in Europe) to want to participate in biobanking research (Kaye et al. 2012; Steinsbekk et al. 2013). Rather than tweaking consent as such, their approach seeks to re-define the relationship between biobanks and data subjects by emphasising the willingness to share data or assist others to support research and innovation (Prainsack and Buyx 2013, p. 74). In contrast to autonomy-based consent approaches, biobanks would instead model consent on solidarity by providing data subjects with a 'mission statement', information on potential areas of research, future uses, risks and benefits, feedback procedures and the potential commercial value of the data, so as to establish a "contractual" rather than consent basis for the research relationship (Prainsack and Buyx 2013, p. 84). Such an approach is claimed to be acceptable given the relatively low risks in genomic research. According to the authors, few examples of discrimination based on biobank-facilitated research exist and are incomparable in

[9] The shift to solidarity is also said to free up the "significant resources" currently spent on (re-)consenting procedures for primary and secondary uses of data held in biobanks for research, innovation and infrastructural improvements including interoperability between repositories (Prainsack and Buyx 2013, p. 80). This position rests on the assumption that significant resources are currently being spent on re-consent procedures in particular, which are a central concern for consent and Big Data (e.g. Wellcome Trust 2013), and that these resources would instead be spent on valuable research and structural improvements.

quality to the bodily harm possible in other types of medical research.[10] One of the most commonly cited fears of insurance discrimination based on disease susceptibility is dismissed due to the "limited predictive value" of genetic markers at the individual level (Prainsack and Buyx 2013, p. 79).

As a counterpoint to solidarity, the alleged 'responsibility' to give up consent rights to prevent "hindering progress" in scientific research and thus social good can be seen as an unethical burden placed on the individual (Crawford et al. 2014, p. 1666). Given the extensive uncertainty over what collected data may reveal in the future, eliminating or reducing the need for informed consent based on the solidarity of the general public cannot be accepted uncritically and seemingly without public debate, particularly if democratic ideals are valued. With that said, the acceptability changes drastically with timescale—possible implications decades in the future are unlikely to outweigh the potential benefits of data sharing now. Another possibility not relying on such a problematic form of responsibility may be to emphasise trust between governance bodies wherein data is shared only between 'trusted' bodies (cf. Hansson 2009, p. 9); however, this may only be a tenable alternative for research data repositories where the extent and initial purpose of data collected is known to data subjects.

Privacy

Unsurprisingly, privacy features very frequently in the literature, often in parallel with anonymisation and confidentiality. Commentary pieces often address privacy issues of Big Data (e.g. Craig 2011; Goodman 2014; Schadt 2012), presumably due to the prevalence of the concept in international legislation and related discussions in applied ethics. In the reviewed literature, numerous concerns were described in terms of privacy, some of which relate to alternative concepts such as autonomy or freedom of information. Links are frequently made with confidentiality, understood as "the duties that accompany the disclosure of non-public information within a fiduciary, professional or contractual relationship" (Majumder 2005, p. 33). Others discussed privacy in terms of the 'invasiveness' of Big Data analysis. Invasiveness was connected in particular to analysis of combined data sets, particularly from geolocation and internet-based sources, even when such data is anonymised (e.g. Markowetz et al. 2014; Moore et al. 2013; Shilton 2012).

Where explicit theories and frameworks of privacy were applied, the OECD's Fair Information Principles and Nissenbaum's 'contextual integrity' (2004) were influential (see for example Andrejevic 2014; Helbing and Balietti 2011; Tene and Polonetsky 2013). Nissenbaum's context-sensitive approach to privacy norms is clearly relevant for emerging forms of participatory data generation such as social media, where data subjects may not be aware of the extent to which data can be publicly 'scraped' and analysed outside of the "highly context-sensitive spaces" in

[10] The relative lack of reporting on harms stemming from abuses of biomedical data has been noted in a recent Nuffield Council report on the ethics of linking biomedical datasets for research (Nuffield Council on Bioethics 2015). The lack has been largely attributed to a lack of robust reporting mechanisms and empirical research on underreporting, with most cases coming from anecdotal accounts and notable media stories. As a result a lack of evidence of harms should not be considered evidence for a lack of harms.

which it is created. Such uses may violate subjects' expectations of data privacy (Boyd and Crawford 2012, p. 673) and expose the data to acontextual interpretation (e.g. Andrejevic 2014, p. 1685). A conceptual link can be drawn to the distinction between 'being in public', in the sense that data communicated via the internet is publicly visible by default, and 'being public', or asserting one's agency purposefully to make something publicly known. This distinction is often ignored in Big Data (Boyd and Crawford 2012, p. 673), insofar as being able to do something becomes synonymous to being justified in doing it. To facilitate formation of realistic privacy norms in Big Data contexts it may be necessary to reinforce the distinction in digital spaces between 'being in public' and 'being public'. 'Offline' privacy barriers such as physical walls can be replaced by raising awareness among data subjects of the uncertain but broad value and seemingly limitless lifespan of the data outside of the original context in which it was authored. Awareness may inhibit authorship or dissemination of sensitive or particularly context-sensitive data.

Following from this, the scope of data being collected can also be conceived of as a privacy issue. Traditionally, data collection has been limited by human perception and cognition. However, with automated and autonomous collection by information technologies, the scope of data, as can be seen over the past two decades, has grown exponentially. More personal and highly detailed data can be collected and analysed than at any other time in history (Nunan and Di Domenico 2013, p. 5). This is a unique characteristic of the 'age of Big Data' (Andrejevic 2014; Puschmann and Burgess 2014). Furthermore, these data are designed to be stored in perpetuity, meaning that traditional limitations of memory no longer apply; data collected today may, in theory, be equally accessible and of the same quality in the future. Other issues are now emerging as important for the preservation of data, such as the obsolescence of software, the presence of malware, and the potential fragility of physical supports. While not a privacy issue per se, extending the lifespan of data describing phenomena that would otherwise be forgotten does increase the risks that privacy violations may occur.

Anonymisation Anonymisation and privacy were closely linked in the literature, wherein privacy concerns raised by Big Data practices can be addressed merely by removing identifying information. Anonymisation was frequently seen as the minimum requirement necessary to protect data subjects' privacy in aggregating data, despite the possibility of re-identification through cross-referencing with data concerning ethnic background, locational data, other metadata, health records or even small pieces of identified genetic data (Choudhury et al. 2014, p. 6; Hayden 2012, p. 313; Joly et al. 2012).

For medical research, data is often anonymised or de-identified to gain consent from data subjects and in accordance with data protection legislation. Beyond explicit 'biobanking' research, study data collected in this way populate Big Datasets. Interestingly, some authors hold that "when analysing data at an aggregate level, a waiver of consent" is acceptable (Krotoski 2012, p. 30), directly linking the acceptability of analysis and re-use of data to the issue of anonymisation (see

"Modifying Consent for Big Data"). However, this position is problematic because it portrays consent as a concept relevant only to the identifiable individual, whereas group-level harms from analysis of aggregated data are clearly possible (Floridi 2014b). Whereas traditionally research ethics has focused on the harms to individual participants, Big Data operates on and impacts groups (of anonymised individuals) (Fairfield and Shtein 2014). Where anonymised data subjects are grouped according to geographical, socioeconomic, ethnic or other characteristics, the anonymisation of individuals matters little if outcomes affect the groups to which they belong (Choudhury et al. 2014, p. 6). Problematic discrimination and stigmatisation of affected groups (see "Group-Level Ethics") is therefore a real risk (Docherty 2014), even in anonymised datasets. Such effects impact on all members of the community, not only those who gave consent (Fairfield and Shtein 2014, p. 45).

As shown at the group-level, anonymisation can be criticised when presented as a 'silver bullet' that avoids, or at least minimises, the risk of being 'singled out' for discrimination or preferential treatment (McGuire et al. 2012). An 'ethics of care' approach may be appropriate when working with Big Data collected from groups based on, for example, indigenous, demographic, ethnic or cultural features, to avoid possibilities of discrimination (Lewis et al. 2012, p. 3).[11]

Data Protection Current data protection legislation in the USA and EU may not protect all medically-relevant or health-related Big Data, or afford such data the protections granted to sensitive health data. As a result, usage of these data will largely be governed by the ethical systems and values governing particular databases or custodians (Liyanage et al. 2014, p. 33), such as institutional or ethical review boards. This situation is particularly concerning for privatized and internet-based health data sources, such as patient-driven databases (e.g. PatientsLikeMe) (Liyanage et al. 2014, p. 33), which are likely to be subjected to less stringent requirements when compared to biobanks and repositories of clinical trial data, where restrictions can be enforced by governance bodies.

Ownership

Ownership is a complex concept, as it can refer to rights regarding the redistribution and modification of data, along with benefiting from intellectual property and innovations developed from its analysis. Redistribution and modification of data may be restricted by the data 'owner' to maintain data integrity, while access is still allowed to data 'analysts' for analysis, innovation, and development of intellectual property. Different databases will have different restrictions in place. A key

[11] The applicability of theories on the ethics of care (e.g. Gilligan 1982; Noddings 2013; Slote 2007) to Big Data likely extend beyond discrimination against marginalised groups. For example, emphasising responsiveness and relationships between data subjects, custodians and analysts may provide avenues for development of new privacy protection mechanisms and group-level ethics which acknowledge the network ethical effects possible through Big Data (see "Group-level ethics"). While a full account of this and related topics concerning ethics of care goes beyond the scope of this paper, existing work on the applicability of the ethics of care to public health (e.g. Kass 2001) may provide a starting point for future enquiries.

distinction is that there are two forms of ownership, as rights to 'control' data, and as rights to 'benefit from' data. The former of these two conceptions was primarily discussed in the literature, perhaps due to the nascent development of intellectual property and products from Big Data thus far.

Understood in terms of 'control', ownership grounds empowerment of data subjects through mechanisms to track and check the existence and manipulation of their data, which can help prevent "the existence of "secret" databases and leverage societal pressure to constrain any unacceptable uses" (Tene and Polonetsky 2013, p. 242). Here, a link can be seen between discrimination and surveillance (see "The Big Data Divide"). When the possibilities of re-identification and 'hidden' analysis exist, data subjects' control over uses of their data acquires greater importance (Choudhury et al. 2014, p. 6), as control allows subjects to restrict undesired uses. For biobanks, control is relevant to considering the permissibility of using research data for commercial pursuits as is made possible by allowing private and third party companies access (e.g. NHS England 2014). Permissibility can be described as an ethical issue when 'human dignity' precludes commodification of humans, or selling one's body or data describing one's body (Steinsbekk et al. 2013).

Understood as a 'benefit', ownership can also require data custodians to enable data subjects to benefit and utilise Big Data for personal uses by being offered "meaningful rights to access their data in a usable, machine-readable format" (Tene and Polonetsky 2013, p. 242). Such steps allow subjects to find individual benefits from the data they produce and communities (or aggregated datasets) in which it resides (Lupton 2014, p. 866).

Accessibility in both contexts is not without risks and necessary limitations. For instance, providing data subjects with unrestricted access to raw data may be harmful in the sense that it is practically useless or open to misinterpretation without the presence of a trained clinician or analyst to explain its significance (Watson et al. 2010). Furthermore, revision rights open datasets to mistakes and inaccurate modification by data subjects, while not addressing questions of accuracy of interpretations or the completeness of the data representations.

Epistemology

Interestingly, despite the search being restricted to literature discussing 'ethics', a number of sources revealed a connection between the ethics and epistemology of Big Data. The connections stem from the perceived complexity of Big Data and the algorithms used to analyse it (Callebaut 2012, p. 70), which may exceed human comprehension—"the intelligent citizen cannot read the programs that run our data sets." In other words, "the natural world and its human observers are being ever more instrumented with intelligent machines … as people we are, in Olga Kuchinskaya's memorable phrase, becoming our own data" (Bowker 2013, p. 170). The problem is not new: patients are usually unable to interpret radiographies, for example. But it has become more significant because of its size, opacity, and pervasiveness. Complexity now refers both to the inherent difficulty of analysing vast datasets, and to the complicated reasoning or rationale of the algorithms (or analytical processes) that make discoveries in Big Data. As a result, questioning the validity of relationships and findings based upon analysis of Big Data becomes increasingly difficult not just for the general public but also for experts, whose critical investigations may become comparable to questioning the outputs of a 'black box'.

Objectivity The aforementioned complexity leads to several related problems. One is a tendency, particularly in mass media and industry, to view Big Data as 'objective'(Crawford 2013; Crawford et al. 2014) or as revealing objective truths without the need for human interpretation. This 'mythological' view of Big Data as the 'end of theory' creates ethical concerns regarding justification of increasingly pervasive and unbounded secondary manipulation and aggregation of data when Big Data practices are seen as the future of science and scientific discoveries. Data are (mistakenly) said to "speak for themselves", creating the possibility of science being driven entirely by induction and reduction, or 'data-driven science' without a need for theory or hypotheses (Callebaut 2012, p. 70; Crawford et al. 2014; Fairfield and Shtein 2014, p. 47). For proponents, the "sheer abundance of information" is seen as providing a "degree of scientific authority" to Big Data practices, which can seemingly be used to explain any "natural and social phenomena" (Puschmann and Burgess 2014, p. 1691) because its meaning is "already there, just waiting to be uncovered" (Puschmann and Burgess 2014, p. 1699).[12]

This idea of data-led objective discoveries entirely discounts the role of interpretive frameworks in making sense of data which, according to non-objectivist ontologies (e.g. Gadamer 1976; Habermas 1984, 1985; Heidegger 1967; Schwandt 2000), is a necessary and inevitable part of interacting with the world, people and phenomena (and thus, data). In the increasingly abstract and complex practices that make up Big Data (Puschmann and Burgess 2014, p. 1697) "data is extracted, collected, cleaned and transformed, stored and managed, analysed, indexed and searched, as well as visualized" (Markowetz et al. 2014, p. 407). In unstructured searches for 'patterns in the data', 'noise' is eliminated as the dataset's boundaries are modified to facilitate the search (Puschmann and Burgess 2014, p. 1699). At each step the data undergoes a transformation by passing through an interpretive framework, yet custodians act as though it remains an objective analogue of reality. What is or may be relevant depends on the questions being asked, which in turn depend on the purposes for which the investigation is being developed. Only a clear understanding of the purposes can ground a rational determination of the levels of abstraction at which the data are queried. The need for human intelligence is actually increasing the more data become available, in order to know which sensible questions to ask and what answers actually make sense (Floridi 2008, 2013).

The tendency to rely on mere Big Data furthermore ignores the variable quality of datasets. For instance, electronic health records typically consist of data written by clinicians for clinical work without the interests of researchers, standardisation and interoperability in mind, while aggregation of observational data for purposes of identifying causal links is prone to selection, confounding and measurement biases

[12] With these tendencies noted, the capacity of Big Data to provide scientific explanations of particular types of social phenomena or human behaviours should not be rejected (e.g. Schroeder 2014).

(Hoffman and Podgurski 2013; Ioannidis 2013, p. 40). If data come to be processed automatically without "human checks" (Hoffman and Podgurski 2013, p. 56) or by algorithms beyond the capabilities of human understanding, the variable quality of the data undermines justification of the actions taken on their behalf.

Some may argue in favour for a distinction between 'objective' or 'raw' data about the physical world and necessarily subjective or interpretive data about human behaviour or social reality, which is nevertheless treated as similarly objective when labelled as 'Big Data' (cf. Lupton 2014, p. 859; Schroeder and Cowls 2014). However, regardless of the position taken on this issue, the key message is that the objectivity of Big Data describing social reality (which includes biomedical data) is often falsely represented by those treating data as inherently neutral and capable of explaining complex phenomena (e.g. 'data-driven science') without need of further contextual knowledge, meaning, or interpretation.

Context When knowledge is seen to 'emerge' from 'raw' data, the need for understanding the contextual meaning or 'situatedness' of the data is seemingly dismissed. Even where the need is acknowledged, contextual understanding may be impossible in aggregated datasets; for instance, as recognised in cultural sociology, "big data often does not include information about the social context in which texts are produced" (Bail 2014, p. 477). Context and meaning may also be purposefully stripped from behaviour and actions, for instance when tracking behaviour via Big Data is viewed as a 'less biased' way to collect behavioural data compared to self-reporting and questionnaires (e.g. Markowetz et al. 2014, p. 406). For research studies, data is collected and interpreted in a particular way to "solve a specific problem, characterized by a limited focus and functionality," limiting possibilities for interoperability between 'stovepipes' or datasets (McNeely and Hahm 2014, p. 306). A tangible loss of methodological and scholarly context occurs through the aggregation of data unless extensive precautions are taken to preserve the 'assumptions' that helped generate the data: "the categories to be used in collecting data, the procedures for handling missing data, the specific subjects of data collection, the nature of the sampling methods used, and the means by which to construct and aggregate the data" (Busch 2014, p. 1730). In other words, aggregation obscures the complex methodological decisions and ontological assumptions that ground the research that produced the data to be aggregated. Busch (2014) describes the following characteristics of aggregated datasets which contribute to a loss of context:

- *Lossiness*: Aggregation, case construction, standardisation and simplification of data to enable cross-sectional analysis may 'lose' certain aspects of the phenomena studied.
- *Drift*: Phenomena change over time, but the data representing them does not. The same can be said for the methods underlying the primary data collection and analysis.
- *Distancing*: Large datasets facilitate identification of patterns or 'clarity' by distancing oneself from the phenomenon.

- *Layering*: A 'realist' ontology is pre-supposed in that an assumption is made that the relations underlying the phenomena will remain over time as the data is aggregated and manipulated. The 'situatedness' or contextual meaning of each phenomenon must be removed for the (data representations of the) phenomena to be treated as sufficiently similar for aggregation. Context is lost by reducing the phenomena to a set of variables: "those aspects of things that are not amenable to numerical or statistical analysis—that situate particular phenomena—are systematically downgraded or removed from consideration" (Busch 2014, p. 1735).
- *Errors*: How are errors within the dataset identified and addressed?
- *Standards*: "The process of creating uniformity through standardization," or fitting data to discipline conventions or categories, "may obfuscate phenomena of considerable importance" (Busch 2014, p. 1736).
- *Disproportionality*: Outlying data may be deleted or treated as 'errors' to enable simplification and standardisation of the dataset.
- *Amplification/Reduction*: Aspects of phenomena amenable to quantified measurement are amplified in importance, while those that are not are reduced.
- *Narratives*: Large datasets can hide the role of interpretation in seeing data *as* something and obscure alternatives to the preferred interpretation.

Contextual aspects which do not fit the structure or classification framework of a database appear to be (sometime irreversibly) lost in Big Data, in some cases through computer-led 'interpretation' of the data (Bowker 2013, p. 170), particularly in social research. This can be referred to as a "signal problem" wherein data is treated as an accurate representation of social reality despite lacking signals from particular communities (Crawford 2013) or interpretive frameworks. For instance, blog posts or tweets can be analysed out of the context in which they are posted (Boyd and Crawford 2012, p. 672), such as responses to a news item or as part of a sarcastic dialogue. This may in part be a technical limitation: the 'sensitivity' of social media data gathered via Application Programming Interfaces (API) is largely unclear at the time of collection for researchers, meaning seemingly innocuous 'chatter' can, when connected to other pieces of data, become highly sensitive and revealing about the data subject (Lomborg and Bechmann 2014, p. 261). The reviewed literature goes so far as to acknowledge the potentially problematic epistemic implications of such acts of interpretation, while stopping short of making a connection with normative or ethical implications.

The Big Data Divide

As with nearly any modern information and communication technology or practice, 'digital divides' can exist within Big Data practices. For example, individuals that 'opt out' of data collection may experience "exclusion from the digitally connected world in which they reside" (Nunan and Di Domenico 2013, p. 7). However, the term 'Big Data divide' is used to describe related but qualitatively different phenomena, understood as the inequalities between data subjects providing the material for Big Data analytics and the organisations with the necessary infrastructure and resources to analyse and understand the data (Andrejevic 2014; Crawford et al. 2014). A divide is created in the terms of the 'forms of knowing' made possible (Andrejevic 2014, p. 1676). The divide concerns 'haves' and 'have nots' of Big Data, where the ability and thus opportunities to assess and utilise Big Data are located within the few organisations possessing the required access, knowledge, computational, and organisational resources necessary to analyse and understand Big Datasets (Berry 2011, p. 2011; Boyd and Crawford 2012; Fairfield and Shtein 2014, p. 46; McNeely and Hahm 2014, p. 308; Puschmann and Burgess 2014, p. 1694). Such a divide can already be seen for research via social media, where access to data from APIs is greatly restricted for individual researchers when compared to organisations or research groups that can pay for access (Lomborg and Bechmann 2014, p. 256; Schroeder 2014).

Big Data is increasingly becoming the sole domain of large organisations, despite calls to allow data subjects to benefit from and manipulate their data (Boyd and Crawford 2012; Tene and Polonetsky 2013). This situation can be troublesome for several reasons, foremost due to the inability of 'underprivileged' individual data subjects and organisations both to understand and have access to the methods, logic or at least "decisional criteria" behind Big Data analysis and decision-making processes (Tene and Polonetsky 2013, p. 243). Furthermore, it is often unclear which individuals and organisations can access or buy one's data (McNeely and Hahm 2014, p. 308).

The divide can also be conceived in terms of access to modify the data (Boyd and Crawford 2012, p. 674), or whether data subjects are empowered to be notified when data about them are created, modified or analysed, and given fair opportunities to access the data and correct errors or misinterpretations in the data and knowledge and profiles built upon it (Coll 2014). Superficially, such potential 'rights' can be connected to the 'right to be forgotten'[13] (Higuchi 2013), insofar as similar rights to modify privately held personal data (rather than publicly available links) could conceivably be granted as an oversight mechanism. Hypothetically, a right to 'self-determination' can ground such connected data rights (Coll 2014, p. 1258) to combat the 'transparency asymmetry' that exists when consumers lack information about how data about them is "collected, analysed and used" (Coll 2014, p. 1259; Richards and King 2013). Broader social "inequalities and biases" can therefore have uninhibited influence over data analysis where subjects lack oversight (McNeely and Hahm 2014, p. 308; Oboler et al. 2012a, p. 3).

Profiling and Surveillance A lack of oversight means data subjects are unaware of the decisions made about their data, and the criteria and categories into which their data fit. Decisions made on the basis of Big Data in some way may restrict the treatment, information or opportunities offered to data subjects (Tene and Polonetsky 2013, p. 252). These decisions made on the basis of aggregated data affect the individual behind the (de-identified) profile as a member of a group or category; "the profile and the person intersect" (Andrejevic 2014, p. 1677) quite

[13] For further details on the specification of the right to be forgotten by Google in the EU, see: Advisory Council to Google on the Right to be Forgotten, 2015.

apart from the individual’s identity. Understanding when and why one’s data have been ‘categorised’ as a particular type or instance of a particular phenomenon is therefore key to reinforcing self-control of data and reducing the imbalance of power characteristic of the ‘Big Data divide’ (Lyon 2003). The ‘data poor’ are caught in a position of weakness wherein the ability to understand the data and methods used to make decisions about them as individuals and members of groups is beyond their means (Andrejevic 2014, p. 1678). Even where discrimination does not occur, “the relegation of decisions about an individual’s life to automated processes” (Tene and Polonetsky 2013, p. 252) is itself troubling due to the imbalance in knowledge and decision-making power inherent in this setup.

Lupton (2014) describes this phenomenon in terms of analytic metrics used to sort individuals and groups and highlight specific aspects or characteristics to ‘understand’ them. The implicit interpretation behind supposedly ‘objective’ Big Data analysis can be seen in these metrics used within aggregated datasets. Metrics “make visible aspects of individuals and groups that are not otherwise perceptible, because they are able to join-up a vast range of details derived from diverse sources” (Lupton 2014, p. 859). These metrics provide different ways of ‘seeing’ the groups and interpreting their behaviours; whether a particular interpretation is correct or reflective of the meaning, identities or motivations given to acts by members of the group is unclear. Following on from the inability to modify or correct one’s data (see “The Big Data Divide“), a ‘right to be forgotten’ according to which data subjects can request deletion or correction of particular pieces of data is thought to be more empowering and privacy-protecting than a blanket right to have a person’s profile or entire data set deleted (Oboler et al. 2012a, p. 9). Correcting the underlying data means future metrics will ideally be applied to a more ‘accurate’ or representative picture of the data subject in her terms.

Profiling can quickly take on surveillance implications (Bonilla 2014, p. 265); Big Data has been compared to an omniscient ‘transparent human’ capable of mass surveillance (Markowetz et al. 2014, p. 410). However, profiling need not be seen as a surveillance practice for concerns over profiling to be relevant—it is the act of interpreting the data through a particular framework of understanding or metric to ‘make sense’ of it, rather than any (problematic) actions taken once this sorting has occurred, which constitutes profiling.

Once profiled, actions taken towards particular groups may be problematic. To take an example from biomedicine, the extent to which data subjects are informed about research results, such as disease proclivity, may require new policies of professional conduct concerning when and how results are released to data subjects sorted into particular disease groups (McGuire et al. 2008, p. 1862, 2012).[14] Discrimination and benefits of Big Data may become localised around groups that present easy or interesting analysis opportunities. Crawford et al. (2014, p. 1667) argue that Big Data leads to new concentrations of power, ‘blind spots’ and problems of representativeness because it “cannot account for those who participate

[14] Regulatory action may be required, as Big Data creates new opportunities for “data aggregators and miners to…run around health care’s domain-specific protections by creating medical profiles of individuals” not subject to existing legislation (Terry 2012, p. 386), as was the case with the Google Health platform which operated outside of HIPAA restrictions in the United States (Mora 2012, p. 373).

in the social world in ways that do not register as digital signals." Correcting these gaps is unlikely, as "big data's opacity to outsiders and subsequent claims to veracity through volume…discursively neutralizes the tendency to make errors." These 'blind spots' mean that analysis will tend to focus on data subjects and phenomena amenable to digitisation and measurement, meaning that the benefits and ethical burdens of Big Data will be placed, for better or worse, on specific social, cultural and economic groups (Majumder 2005, p. 37; McGuire et al. 2008). For instance, analysis of social media datasets will necessarily affect social media users and their underlying demographics in the first instance.

Justice It may be possible to express such divides as ethically problematic in terms of justice. Interventions and knowledge developed from Big Data, particularly genomic and microbiomic data (Lewis et al. 2012), may favour populations from whom data is collected, further exacerbating existing gaps in medical practice and knowledge between "Euro-Americans of middle to upper socio-economic status" and others (Lewis et al. 2012, p. 2). Even where studied populations are diverse, formal benefit sharing agreements may be required between data subjects and custodians or researchers to ensure data are not taken from one context purely to benefit individuals in another, similar to the issues faced with pharmaceutical research in the third world (Mathaiyan et al. 2013, p. 103). As much should be done to facilitate benefit sharing as possible (Choudhury et al. 2014, p. 4), as Big Data can allow researchers to meet the moral obligation to maximise the value of data collected from research participants without the need for further data collection which places participants at risk (Currie 2013; Mello et al. 2013, p. 1653).

Discussion

Reviewing literature is a first step to conduct ethical foresight, in the sense that it allows one to distinguish between issues and implications that are currently under consideration, and those that are not yet acknowledged or require further attention. Overall, the quality of the reviewed literature leaves gaps based on a dearth of empirical research and 'deep' conceptual analysis. In particular, the prevalence of 'opinion pieces' and 'editorials' that briefly raise issues but do not discuss them in depth shows the need for further scholarship in this area of emerging ethical import.

As the results were presented as a narrative overview with accompanying commentary, this section will take the next step by drawing attention to issues that have received insufficient attention in the literature. Specifically, the discussion highlights issues that are expected by the authors to be key ethical issues in the near future, and which require further exploration in the context of specific Big Data practices and domains. These issues include group-level ethics, ethical implications of growing epistemological challenges (e.g. Floridi 2012), effects of Big Data on fiduciary relationships, the ethics of academic versus commercial practices, ownership of intellectual property derived from Big Data, and the content of and barriers to meaningful data access rights.

Group-Level Ethics

Technological means to prevent ethical problems through Big Data tend to focus on the individual, ignoring harms which affect groups. Data protection legislation and anonymisation techniques implicitly focus on the individual in seeking an appropriate balance between the value of the anonymised dataset for subsequent analysis and the privacy of individual data subjects. Such technical solutions to avoid the potential ethical harms of Big Data practices are only partially successful and remain fallible. Advances in analytic methods and technologies of re-engineering identity (e.g. Cassa et al. 2008; Hay et al. 2008), or failures in the oversight processes preceding the release of datasets which fail to identify potential means of re-identification guarantee future vulnerability.

In the face of such technological and practical uncertainties (e.g. Mittelstadt et al. 2015), employing punitive measures for attempts to re-identify data, or emphasising professional responsibility (for example through codes of ethics for data custodians; see "Fiduciary Relationships" and Oboler et al. 2012a, p. 11) may prove more effective than increasingly restrictive anonymisation protocols. Alternatively, data may be hosted in 'safe harbours' within which data uses are screened and controlled (Dove et al. 2014). Although these measures do not address group-level effects, they are pragmatically responsive to possibilities of re-identification, while not further restricting movement of anonymised data.

Even where such solutions are implemented, the emphasis on protecting the individual problematically focuses ethical assessment on harms at the individual level (see "Anonymisation"); perfectly anonymised datasets still allow for group-level ethical harms for which the identities of members of the group or profile are irrelevant (van der Sloot 2014). Algorithmic grouping of data points and identification of statistical relationships allows for profiling and grouping of individual data subjects (see "Profiling and Surveillance"). Profiling connects data subjects to one another, meaning the behaviours, preferences and interests of others affect how the individual is treated in ethically relevant ways. Preferential treatment and decision-making in a variety of contexts of variable ethical acceptability can be justified on this basis, such as personalised pricing in e-commerce or genetic discrimination.[15]

To address potential discrimination against particular demographic, genomic or other groups, an 'ethic of care' approach may be required which would set aside particular forms of research or hypotheses as 'off limits' (cf. Lewis et al. 2012). Alternatively, it may be possible to conceive of privacy as a group-level concept and thus speak of 'group privacy rights' that could restrict the flow and acceptable uses of aggregated datasets and profiling. However, the feasibility and practicalities of expanding privacy rights require further investigation, in particular the potential barriers created for desirable research similar to the informed consent

[15] As an example of the latter, if biobanking research utilising genome sequences were to reveal that obesity is linked primarily to behaviour rather than genes, or an ethnic group were shown to have a higher genetic pre-disposition to cancer (cf. Angrist 2009; Mathaiyan et al. 2013), well-meaning research may inadvertently lead to future discrimination against these groups.

debate currently underway in Europe (see "Informed Consent"; Taylor and Floridi 2015).

Epistemological Difficulties

As discussed above, a loss of qualification or contextual aspects of data has been observed in Big Data analytics, which in some cases can be attributed to complex interpretations of data performed by computers or analytical algorithms (Bowker 2013, p. 170). While this position problematically appears to place the responsibility for interpretation (seeing data *as* something) entirely on (learning) algorithms while exonerating designers of algorithms and the ontological categories within which they interpret, it helpfully emphasises the loss of context through quantification and categorisation of diverse datasets to facilitate analysis and connectivity. This loss of context or 'decontextualisation' can be understood as an instance of 'ontic occlusion' (see Bowker 2014; Knobel 2010), or the process by which emphasising particular aspects of a phenomenon in a discourse necessarily occludes or 'downplays' other aspects.

Ontic occlusion, originally developed to describe ontological characteristics of archiving, can be extended to Big Data to describe a qualitative loss or degradation of the data implied by acts of interpretation, classification or categorisation of the data in collection and analysis. Archives or datasets, conceived of as discourses, "cannot in principle contain the world in small…most slices of reality are not represented" (Bowker 2014, p. 1797). If data is seen as describing a particular instance of a phenomenon, for example data describing the case of a particular cancer patient, the instance and data become equivalent; the profile becomes a representation of the profiled (e.g. Floridi 2012). While undoubtedly a problem with any type of data collection and analysis, in Big Data this necessary loss of context is exacerbated by the sheer scale of data being analysed. It is tempting to view the profile, or the data, as representative of the whole phenomenon (Bowker 2014, p. 1797); increasing the scale of data to be considered only increases the difficulty of identifying what is stripped from data to make sense of it. The implications of this problem require further attention in specific Big Data practices; for example, it is likely more ethically problematic to strip context from data used to track the behaviours of individuals than it is to remove identifying information from tissue samples for medical research.

Fiduciary Relationships

Further research may also be required into the effects of Big Data on the 'internal goods' (cf. MacIntyre 2007) of relationships and interactions between data custodians (e.g. researchers, commercial organisations, repositories) and data subjects. The background disciplines and sentiments informing the conceptualisation of 'Big Data' in ongoing discussion is important in defining the obligations that can be attributed to data custodians. When Big Data is thought of as a form of business based around the selling and processing of data for commercial advantage, it is perhaps inappropriate to expect a relationship based on 'trust' or professionalism to exist between subjects and custodians (cf. Terry 2012).

The mediating role of data in these relationships, by which data subjects are 'represented' or revealed to custodians through their data, may be of ethical importance in certain contexts. In medicine for example, greater reliance on data representations of patients brought about by adoption of Big Data practices may create new gaps in care or doctor-patient relationships (cf. Beauchamp and Childress 2009; MacIntyre 2007; Pellegrino and Thomasma 1993). Traditional fiduciary 'healing relationships' do not scale well to Big Data or even institutional care (Terry 2014, p. 838), meaning that, as data representations and models are increasingly used to understand the patient's condition, the 'virtues' or internal goods of traditional medical relationships may be subtly undermined or realised less frequently.

Harm can occur to the data subject through misinterpretation or overreliance on data representing the subject's state(s) of being. The 'goods' provided by such relationships, which extend beyond issues of efficiency or effectiveness of interventions and are derived from the character of the individual providing care, may be undermined; for instance, care providers may be less able to demonstrate understanding, compassion and other desirable traits found within 'good' medical interactions *in addition to* applying their knowledge of medicine to the patient's case (cf. Beauchamp and Childress 2009; MacIntyre 2007; Pellegrino and Thomasma 1993). Put another way, the patient's body and voice may increasingly be replaced or supplemented by data representations of state of being if Big Data practices are adopted in medicine (Barry et al. 2001). Further research is required into the effects of these representations on the quality of relationships through which care is provided. Medical relationships are of particular concern due to the patient being in a vulnerable (and trusting) state (Pellegrino and Thomasma 1993).

Academic Versus Commercial Practices

In terms of the likelihood of future problematic uses, a distinction should be drawn between 'academic' and 'commercial' Big Data practices in order to allow data subjects to retain realistic expectations over potential uses and implications of authoring data (cf. Lupton 2014). The need for such a distinction can be seen for example in the deficiencies of existing patient experience websites, many of which fail to inform users whether collected data will be used for research or commercial purposes (Lupton 2014), or in ethically controversial research being permitted in commercial contexts which would not pass the scrutiny of an academic ethical review board (Schroeder 2014). While 'research' and 'commercialisation' are not mutually exclusive, meaningful ethical distinctions can be drawn. The purpose here is not to distinguish between types of Big Data practices, but rather the motivations behind them. For example, commercial and academic research may be qualitatively similar, in terms of the experiences of the data subject and methods of research, but differ substantially in motives, e.g. basic research to advance scientific knowledge versus product development. Furthermore, data subjects may be interested in the degree of oversight for particular practices. In general, research-based practices will require some form of ethical review and monitoring, whereas commercial practices will not. Clearly, this distinction requires further specification to distinguish between 'types' of Big Data practices in terms of their ethical dimensions.

Ownership of Intellectual Property

In the reviewed literature, ownership was discussed as a mechanism to control data. While undoubtedly important, ownership can also refer to owning products and intellectual property produced through Big Data practices. This issue was only discussed in one article which called for benefit sharing with data subjects to allow for innovation led by data subjects in developing products and services from Big Data (Tene and Polonetsky 2013). Despite this relative paucity of attention, this topic deserves further debate due to the potential to develop commercially valuable material through analysis of data collected from or volunteered by members of the public. Currently, data subjects tend not to benefit from analysis of data collected about them—users of Facebook, for instance, do not share in the revenue derived from targeted advertisements. As similar products and services become increasingly common and commercially viable in the future, the ownership of personal data will attain renewed importance. In the future, Big Data will likely raise questions over ownership structures in which data subjects forfeit all rights to personal data generated through usage of networked products and services. It could alternatively become the norm for data subjects to share in (financial) benefits derived from their data, or at least to be guaranteed access to it for personal uses and development. At the very least, ownership structures for personal data require further attention due to the apparent potential of Big Data, to encourage and exploit exponential growth of personal data.

Data Access Rights

Following on from ownership, access mechanisms and rights for data subjects require further attention. As discussed in the context of ownership (see "Ownership"), data subject rights to access and modify data are reliant upon the subject being aware of what data exist about her, who holds them, what they (potentially) mean and how they are being used. Assuming such rights are sought (as specified in data protection legislation, for example), significant technical and practical barriers to their realisation exist which may be insurmountable, thus precluding the possibility of meaningful data access rights in the era of Big Data.

For access rights to be meaningful, data subjects must be able to exercise them with reasonable effort. For instance, being provided with thousands of printed pages of digital data would require unreasonable effort on the part of the data subject to compile and understand the data, and would therefore fail to preserve a meaningful right to access. As discussed in the context of the 'Big Data divide', resource, skillset and comprehension barriers exist which would prevent a 'lay' data subject from being able to exercise the aforementioned access rights. Big Data requires significant computational power and storage, and advanced scientific know-how. As with any data science, analysis will require discipline-specific skills and knowledge, often only accessible through extensive training and education. Even for willing subjects, the amount of time and effort required to attain the background knowledge and skills to understand the totality of data held about oneself may easily be overwhelming. Ascertaining the extent and uses of data held about an individual is also difficult, given the often ‘hidden’ and seemingly ubiquitous nature of personal data processing (see “Background”).

Considered together, the emerging picture is of data subjects in a disempowered state, faced with seemingly insurmountable barriers to understanding *who* holds *what* data about them, being used for *which* purposes. Further, in relation to modification and correction of personal data, it is unclear how subjects can possibly propose changes to data without first understanding the contents and inferences drawn from them, or the perhaps inaccurate or incomplete ways in which the data represent the subject and her behaviours. For a meaningful right to modification and correction it may therefore be necessary for data custodians to provide oversight and explanations of categories, profiles or other criteria used in sorting the data to, at a minimum, allow subjects to understand the ‘silos’ into which they have been placed (see “Profiling and Surveillance”).

Considered together, these barriers may preclude the exercise of meaningful data access rights within current Big Data practices. However, further research is required to justify this assertion. Specifically, specifications are required of reasonable access rights, domain-specific barriers to access, and alterations to practices or data protection legislation which will ensure data custodians assist data subjects in gaining meaningful access as far as possible.

A small number of mechanisms to address issues of data sharing and irresponsible usage of data have been proposed in the reviewed literature. For instance, McNeely and Hahm (2014, p. 1654) have proposed a set of ‘core principles of expanded data sharing’ to be followed by “any system that is ultimately adopted for expanded access to participant-level data.” These principles emphasise responsibility, privacy, equal treatment of all data requesters/trial sponsors, accountability of data custodians and requesters, and the practicality of the system in terms of transparent and timely responses to data requests and a lack of other such unnecessary barriers to access. Other suggestions include granting data subjects a ‘right to be forgotten’, a ‘right to data expiry’, and the ‘ownership of a social graph’. The first refers to the ability of data subjects to request that links to information about them be deleted. The second refers to the automatic deletion of unstructured data after a set period of time if they no longer have any commercial or research value. The third will detail what data exist about an individual, when and how they were collected, and where they are stored (Nunan and Di Domenico 2013).

While each of these concepts faces theoretical and practical difficulties, such as defining ‘commercial’ or ‘research’ value, they nevertheless represent an attempt to realise meaningful data rights in the era of Big Data. Modifications appear to be required given the existing inaccessibility and incomprehensibility of Big Data algorithms and practices to ‘lay’ data subjects—some form of assistance or ‘hand holding’ is required by data custodians given the increasing prevalence of data in mediating human interactions. Going forward, competitive interests and desires for commercial secrecy need to be balanced against meaningful access rights for data subjects.

Conclusion

As is often the case with emerging technologies and sciences, a tendency has been recognised to overemphasise the potential benefits of Big Data as a means of explaining ‘everything’, perhaps without the need for theories or frameworks of understanding (Callebaut 2012; Crawford 2013). “Data fundamentalism,” or the idea that “correlation always indicates causation, and that massive data sets and predictive analytics always reflect objective truth” (Crawford 2013), problematically influences the public, mass media and researchers where a tendency exists to view the advancement of Big Data into all information-based disciplines as inevitable. In such cases, beneficial outcomes of this shift are often similarly ‘inevitable’ (e.g. Costa 2014, p. 436), with practitioners more concerned with communicating how ‘good’ or ‘responsible’ they are rather than investigating what these concepts mean in the context of specific Big Data practices. Such broad brush attitudes towards Big Data should be avoided if its ethical implications are to be given serious consideration throughout the life of emerging Big Data practices, products and applications.

The analysis offered in this article is intended to contribute to transforming such general and perhaps overly optimistic attitudes by providing a starting point and comprehensive reference for future discussions of the ethics of Big Data, especially in the very sensitive context of biomedical research. An overview of key ethical issues of Big Data has been offered, against which areas requiring further research in the near term have been identified. In particular, biomedical applications of Big Data have been identified as particularly ethically challenging due to the sensitivity of health data and fiduciary nature of healthcare. It is our hope that the analysis will contribute to ethically responsibility development, deployment and maintenance of novel datasets and practices in biomedicine and beyond in the era of Big Data.

Acknowledgments The research leading to this work has been funded by a <removed for anonymity> major research grant. An initial version of this paper was discussed at a workshop organised at the <removed for anonymity> on <removed for anonymity>. We wish to acknowledge the extremely valuable feedback received during that meeting and from the two anonymous reviewers. This study was funded by the University of Oxford’s John Fell Fund.

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

See Table 3.

Table 3 Ethical themes

Reference	In-depth	Biomedical	General	Informed consent	Privacy	Anony-misation	Data protection	Owner-ship	Episte-mology	Power	Digital divide	Article type	Empirical
Andrejevic (2014)	X		X		X			X		X	X	Research	X
Bail (2014)	X		X									Research	
van den Berg and van der Hof (2012)			X		X		X					Research	
Berry (2011)	X		X								X	Research	
Bonilla (2014)	X		X		X		X			X		Research	
Booch (2014)			X		X			X				Commentary	
Bowker (2013)			X						X			Commentary	
Bowker (2014)			X						X			Commentary	
Boyd and Crawford (2012)	X	X	X	X	X	X	X		X	X	X	Research	
Busch (2014)	X				X				X			Research	
Callebaut (2012)		X							X			Research	
Cheng et al. (2013)		X		X	X							Research	
Choudhury et al. (2014)	X	X		X	X	X		X		X		Research	
Clayton (2005)		X		X	X					X		Research	
Coll (2014)	X		X		X		X	X		X		Research	
Costa (2014)		X		X	X					X		Research	
Craig (2011)			X	X	X		X	X				Commentary	
Crawford et al. (2014)	X		X						X	X	X	Research	
Crawford (2013)			X						X			Commentary	
Currie (2013)		X		X	X							Research	
Davis (2012)			X	X	X		X					Commentary	

Table 3 continued

Reference	In-depth	Biomedical	General	Informed consent	Privacy	Anonymisation	Data protection	Ownership	Epistemology	Power	Digital divide	Article type	Empirical
Dereli et al. (2014)		X										Commentary	
Docherty (2014)		X		X		X						Commentary	
Fairfield and Shtein (2014)	X		X	X	X	X			X	X	X	Research	
Fan and Bifet (2013)			X								X	Commentary	
Floridi (2012)			X						X			Commentary	
Goodman (2014)			X		X							Commentary	
Hansson (2009)	X	X		X	X	X						Research	
Hayden (2012)		X		X		X						Commentary	
Higuchi (2013)		X			X		X					Commentary	
Hoffman and Podgurski (2013)		X							X			Research	
Ioannidis (2013)		X		X	X							Commentary	
Joly et al. (2012)		X		X	X							Research	
Krotoski (2012)		X		X		X						Commentary	
Larson (2013)		X		X	X							Commentary	
Lazer et al. (2009)			X		X	X						Commentary	
Lewis et al. (2012)		X										Research	
Liyanage et al. (2014)		X			X	X		X				Research	
Lomborg and Bechmann (2014)	X		X	X	X	X					X	Research	
Lupton (2014)	X	X			X			X	x		X	Research	X
Lynch (2008)			X								X	Commentary	
Mahajan et al. (2012)			X							X	X	Research	

Table 3 continued

Reference	In-depth	Biomedical	General	Informed consent	Privacy	Anony-misation	Data protection	Owner-ship	Episte-mology	Power	Digital divide	Article type	Empirical
Majumder (2005)	X	X		X	X						X	Research	
Markowetz et al. (2014)			X		X						X	Research	
Master et al. (2014)		X		X								Research	X
Mathaiyan et al. (2013)		X		X							X	Research	
McGuire et al. (2012)		X		X	X	X				X		Research	X
McGuire et al. (2008)		X		X	X		X					Commentary	
McNeely and Hahm (2014)	X		X		X				X	X	X	Research	
Mello et al. (2013)	X	X			X	X						Research	
Nunan and Di Domenico (2013)			X	X	X	X		X			X	Research	
Oboler et al. (2012b)			X		X					X		Research	
Prainsack and Buyx (2013)	X	X		X	X		X			X		Research	
Puschmann and Burgess (2014)	X		X						X	X	X	Research	
Richards and King (2013)	X		X		X					X	X	Research	
Rothstein and Shoben (2013)		X		X	X							Commentary	
Safran et al. (2006)		X		X	X	X		X		X		Research	X
Schadt (2012)		X		X	X	X		X				Commentary	
Schroeder (2014)	X		X	X	X		X		X		X	Research	
Schroeder and Cowls (2014)			X		X		X		X		X	Commentary	

Table 3 continued

Reference	In-depth	Biomedical	General	Informed consent	Privacy	Anony-misation	Data protection	Owner-ship	Episte-mology	Power	Digital divide	Article type	Empirical
Shilton (2012)	X		X	X	X	X				X	X	Research	
Slashinski et al. (2012)		X						X				Research	X
Steinsbekk et al. (2013)		X		X		X						Research	X
Tene and Polonetsky (2013)	X		X	X	X	X	X			X	X	Research	
Terry (2012)	X	X		X	X	X	X			X	X	Research	
Terry (2014)		X			X		X			X		Commentary	
The NIH HMP Working Group et al. (2009)		X		X								Research	
Watson et al. (2010)		X		X		X		X		X		Commentary	
Total	23	36	31	34	44	20	14	12	14	22	22	N/A	7

References

Advisory Council to Google on the Right to be Forgotten. (2015). Report of the advisory council to google on the right to be forgotten. *Google Docs*. https://drive.google.com/file/d/0B1UgZshetMd4cEI3SjlvV0hNbDA/view?pli=1&usp=embed_facebook. Accessed 19 Mar 2015.

Andrejevic, M. (2014). Big data, big questions the big data divide. *International Journal of Communication*, *8*(0), 17. Accessed 7 Oct 2014.

Angrist, M. (2009). Eyes wide open: The personal genome project, citizen science and veracity in informed consent. *Personalized Medicine, 6*, 691–699.

Apple. (2014). iBeacon for developers: Apple developer. https://developer.apple.com/ibeacon/. Accessed 17 Nov 2014.

Bail, C. A. (2014). The cultural environment: Measuring culture with big data. *Theory and Society, 43*(3–4), 465–482. doi:10.1007/s11186-014-9216-5.

Barry, C. A., Stevenson, F. A., Britten, N., Barber, N., & Bradley, C. P. (2001). Giving voice to the lifeworld. More humane, more effective medical care? A qualitative study of doctor-patient communication in general practice. *Social Science and Medicine, 53*, 487–505. doi:10.1016/s0277-9536(00)00351-8.

Beauchamp, T. L., & Childress, J. F. (2009). *Principles of biomedical ethics*. New York: Oxford University Press.

Berry, D. M. (2011). The computational turn: Thinking about the digital humanities. *Culture Machine*, *12*(0). ftp://121.171.90.140/big.data/%EB%B9%85%EB%8D%B0%EC%9D%B4%ED%84%B02_20131024_sunup/THE%20COMPUTATIONAL%20TURN%20Digital-Humanities.pdf. Accessed 7 Oct 2014.

Bonilla, D. N. (2014). Information Management professionals working for intelligence organizations: Ethics and deontology implications. *Security and Human Rights, 24*(3–4), 264–279. doi:10.1163/18750230-02404005.

Booch, G. (2014). The human and ethical aspects of big data. *IEEE Software*, *31*(1), 20–22. Accessed 30 Sept 2014.

Bowker, G. C. (2013). *Data flakes: An afterword to "Raw Data"is an oxymoron*. Raw data" is an oxymoron. Cambridge: MIT Press. http://www.ics.uci.edu/~vid/Readings/bowker_data_flakes.pdf. Accessed 14 Oct 2014.

Bowker, G. C. (2014). Big data, big questions the theory/data thing. *International Journal of Communication*, *8*(0), 5. Accessed 7 Oct 2014.

Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data: Provocations for a cultural, technological, and scholarly phenomenon. *Information Communication & Society, 15*(5), 662–679. doi:10.1080/1369118X.2012.678878.

Boye, N. (2012). Co-production of Health enabled by next generation personal health systems. *Studies in health technology and informatics, 177*, 52–58.

Busch, L. (2014). Big data, big questions a dozen ways to get lost in translation: Inherent challenges in large scale data sets. *International Journal of Communication*, *8*(0), 18. Accessed 7 Oct 2014.

Butler, D. (2013). When Google got flu wrong. *Nature, 494*(7436), 155–156. doi:10.1038/494155a.

Callebaut, W. (2012). Scientific perspectivism: A philosopher of science's response to the challenge of big data biology. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 43*(1), 69–80. doi:10.1016/j.shpsc.2011.10.007.

Cassa, C. A., Wieland, S. C., & Mandl, K. D. (2008). Re-identification of home addresses from spatial locations anonymized by Gaussian skew. *International Journal of Health Geographics, 7*(1), 45. doi:10.1186/1476-072X-7-45.

Cheng, L., Shi, C., Wang, X., Li, Q., Wan, Q., Yan, Z., et al. (2013). Chinese biobanks: Present and future. *Genetics Research, 95*(6), 157–164. doi:10.1017/S0016672313000190.

Choudhury, S., Fishman, J. R., McGowan, M. L., & Juengst, E. T. (2014). Big data, open science and the brain: Lessons learned from genomics. *Frontiers in Human Neuroscience, 8*, 239. doi:10.3389/fnhum.2014.00239.

Clayton, E. W. (2005). Informed consent and biobanks. *Journal of Law, Medicine & Ethics, 33*(1), 15–21. doi:10.1111/j.1748-720X.2005.tb00206.x.

Coll, S. (2014). Power, knowledge, and the subjects of privacy: Understanding privacy as the ally of surveillance. *Information Communication & Society, 17*(10), 1250–1263. doi:10.1080/1369118X.2014.918636.

Collingridge, D. (1980). *The social control of technology*. Palgrave Macmillan.

Costa, F. F. (2014). Big data in biomedicine. *Drug Discovery Today, 19*(4), 433–440. doi:10.1016/j.drudis.2013.10.012.

Craig, T. (2011). *Privacy and big data*. Sebastopol; Cambridge: O'Reilly.

Crawford, K. (2013). The hidden biases in big data. *Harvard Business Review*. http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/. Accessed 10 Oct 2014.

Crawford, K., Gray, M. L., & Miltner, K. (2014). Critiquing big data: Politics, ethics, epistemology special section introduction. *International Journal of Communication*, *8*, 10. Accessed 2 Oct 2014.

Currie, J. (2013). "Big Data" Versus "Big Brother": On the appropriate use of large-scale data collections in pediatrics. *Pediatrics, 131*(Supplement), S127–S132. doi:10.1542/peds.2013-0252c.

Davis, K. (2012). *Ethics of big data*. O'Reilly Media, Inc.

Dereli, T., Coskun, Y., Kolker, E., Guner, O., Agirbasli, M., & Ozdemir, V. (2014). Big data and ethics review for health systems research in LMICs: Understanding risk, uncertainty and ignorance-and catching the black swans? *American Journal of Bioethics, 14*(2), 48–50. doi:10.1080/15265161.2013.868955.

Devos, Y., Maeseele, P., Reheul, D., Van Speybroeck, L., & De Waele, D. (2008). Ethics in the societal debate on genetically modified organisms: A (Re)Quest for sense and sensibility. *Journal of Agricultural and Environmental Ethics, 21*(1), 29–61. doi:10.1007/s10806-007-9057-6.

Docherty, A. (2014). Big data: Ethical perspectives. *Anaesthesia, 69*(4), 390–391. doi:10.1111/anae.12656.

Dove, E. S., Knoppers, B. M., & Zawati, M. H. (2014). Towards an ethics safe harbor for global biomedical research. *Journal of Law and the Biosciences, 1*(1), 3–51. doi:10.1093/jlb/lst002.

Enjolras, B. (2014). Big Data and social research: New possibilities and ethical challenges. *Tidsskrift for Samfunnsforskning, 55*(1), 80–89.

EURORDIS. (2013). Statement on the EP Report on the Protection of Personal Data. http://www.publichealth.ox.ac.uk/helex/Statement%20Data%20Prot%20FINAL.pdf. Accessed 22 Oct 2014.

Fairfield, J., & Shtein, H. (2014). Big data, big problems: Emerging issues in the ethics of data science and journalism. *Journal of Mass Media Ethics, 29*(1), 38–51. doi:10.1080/08900523.2014.863126.

Fan, W., & Bifet, A. (2013). Mining big data: Current status, and forecast to the future. *ACM SIGKDD Explorations Newsletter*, *14*(2), 1–5. Accessed 2 Oct 2014.

Floridi, L. (2008). The method of levels of abstraction. *Minds and Machines, 18*(3), 303–329. doi:10.1007/s11023-008-9113-7.

Floridi, L. (2012). Big data and their epistemological challenge. *Philosophy & Technology, 25*(4), 435–437. doi:10.1007/s13347-012-0093-4.

Floridi, L. (2013). *The philosophy of information* (Reprint ed.). Oxford: OUP Oxford.

Floridi, L. (Ed.). (2014a). *The onlife manifesto*. New York: Springer. http://www.springer.com/philosophy/epistemology+and+philosophy+of+science/book/978-3-319-04092-9. Accessed 2 Dec 2014.

Floridi, L. (2014b). Open data, data protection, and group privacy. *Philosophy & Technology, 27*(1), 1–3. doi:10.1007/s13347-014-0157-8.

Gadamer, H. G. (1976). *The historicity of understanding*. Harmondsworth: Penguin Books Ltd.

Gadamer, H. G. (2004). *Truth and method*. London: Continuum International Publishing Group.

General Medical Council. (2008). Consent guidance. http://www.gmc-uk.org/guidance/ethical_guidance/consent_guidance_index.asp.

Gilligan, C. (1982). *In a different voice*. Cambridge: Harvard University Press.

Goodman, E. (2014). Design and ethics in the era of big data. *Interactions*, *21*(3), 22–24. Accessed 1 Oct 2014.

Habermas, J. (1984). *The theory of communicative action. Volume 1: Reason and the rationalization of society*. Boston: Beacon.

Habermas, J. (1985). *The theory of communicative action. Volume 2: Lifeworld and system: A critique of functionalist reason*. Boston: Beacon.

Hansson, M. G. (2009). Ethics and biobanks. *British Journal of Cancer, 100*(1), 8–12. doi:10.1038/sj.bjc.6604795.

Harris, J. (2005). Scientific research is a moral duty. *Journal of Medical Ethics, 31*(4), 242–248. doi:10.1136/jme.2005.011973.

Hay, M., Miklau, G., Jensen, D., Towsley, D., & Weis, P. (2008). Resisting structural re-identification in anonymized social networks. *Proceedings of the VLDB Endowment, 1*(1), 102–114. doi:10.14778/1453856.1453873.

Hayden, E. C. (2012). *A broken contract*. NATURE PUBLISHING GROUP MACMILLAN BUILDING, 4 CRINAN ST, LONDON N1 9XW, ENGLAND. http://environmentportal.in/files/file/informed%20consent.pdf. Accessed 7 Oct 2014.

Heidegger, M. (1967). *Being and time*. Oxford: Blackwell.

Helbing, D., & Balietti, S. (2011). From social data mining to forecasting socio-economic crises. *European Physical Journal-Special Topics, 195*(1), 3–68. doi:10.1140/epjst/e2011-01401-8.

Higuchi, N. (2013). Three challenges in advanced medicine. *Japan Medical Association Journal, 56*(6), 437–447.

Hoffman, S. (2014). *Citizen science: The law and ethics of public access to medical big data* (SSRN Scholarly Paper No. ID 2491054). Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=2491054. Accessed 13 Oct 2014.

Hoffman, S., & Podgurski, A. (2013). Big bad data: Law, public health, and biomedical databases. *Journal of Law, Medicine and Ethics, 41*(Suppl. 1), 56–60. doi:10.1111/jlme.12040.

IBM. (2014). The four V's of big data. http://www.ibmbigdatahub.com/infographic/four-vs-big-data. Accessed 23 Oct 2014.

Ioannidis, J. P. A. (2013). Informed consent, big data, and the oxymoron of research that is not research. *American Journal of Bioethics, 13*(4), 40–42. doi:10.1080/15265161.2013.768864.

Joly, Y., Dove, E. S., Knoppers, B. M., Bobrow, M., & Chalmers, D. (2012). Data sharing in the post-genomic world: The experience of the international cancer genome consortium (ICGC) data access compliance office (DACO). *PLoS Computational Biology, 8*(7), e1002549. doi:10.1371/journal.pcbi.1002549.

Kass, N. E. (2001). An ethics framework for public health. *American Journal of Public Health, 91*(11), 1776–1782. doi:10.2105/AJPH.91.11.1776.

Kaye, J., Curren, L., Anderson, N., Edwards, K., Fullerton, S. M., Kanellopoulou, N., et al. (2012). From patients to partners: Participant-centric initiatives in biomedical research. *Nature Reviews Genetics, 13*(5), 371–376. doi:10.1038/nrg3218.

Knobel, C. P. (2010). *Ontic occlusion and exposure in sociotechnical systems*. University of Pittsburgh. Retrieved from http://deepblue.lib.umich.edu/handle/2027.42/78763.

Krotoski, A. K. (2012). Data-driven research: Open data opportunities for growing knowledge, and ethical issues that arise. *Insights: The UKSG Journal, 25*(1), 28–32. doi:10.1629/2048-7754.25.1.28.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, *6*.

Larson, E. B. (2013). Building trust in the power of "big data" research to serve the public good. *JAMA, 309*(23), 2443–2444. doi:10.1001/jama.2013.5914.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., et al. (2009). Computational social science. *Science, 323*(5915), 721–723. doi:10.1126/science.1167742.

Lewis, C. M., Obregón-Tito, A., Tito, R. Y., Foster, M. W., & Spicer, P. G. (2012). The Human Microbiome Project: Lessons from human genomics. *Trends in Microbiology, 20*(1), 1–4. doi:10.1016/j.tim.2011.10.004.

Liyanage, H., de Lusignan, S., Liaw, S.-T., Kuziemsky, C. E., Mold, F., Krause, P., et al. (2014). Big data usage patterns in the health care domain: A use case driven approach applied to the assessment of vaccination benefits and risks. Contribution of the IMIA Primary Healthcare Working Group. *Yearbook of medical informatics*, *9*(1), 27–35. doi:10.15265/IY-2014-0016.

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *Information Society, 30*(4), 256–265. doi:10.1080/01972243.2014.915276.

Lupton, D. (2014). The commodification of patient opinion: The digital patient experience economy in the age of big data. *Sociology of Health & Illness, 36*(6), 856–869. doi:10.1111/1467-9566.12109.

Lynch, C. (2008). Big data: How do your data grow? *Nature, 455*(7209), 28–29. doi:10.1038/455028a.

Lyon, D. (2003). *Surveillance as social sorting: Privacy, risk, and digital discrimination*. London: Routledge.

MacIntyre, A. (2007). *After virtue: A study in moral theory* (3rd ed.). London: Gerald Duckworth & Co Ltd.

Mahajan, R. L., Reed, J., Ramakrishnan, N., Mueller, R., Williams, C. B., & Campbell, T. A. (2012). Cultivating emerging and black swan technologies (Vol. 6, pp. 549–557). Presented at the ASME international mechanical engineering congress and exposition, proceedings (IMECE). doi:10.1115/IMECE2012-89339

Majumder, M. A. (2005). Cyberbanks and other virtual research repositories. *Journal of Law, Medicine & Ethics, 33*(1), 31–39. doi:10.1111/j.1748-720X.2005.tb00208.x.

Markowetz, A., Błaszkiewicz, K., Montag, C., Switala, C., & Schlaepfer, T. E. (2014). Psycho-Informatics: Big Data shaping modern psychometrics. *Medical Hypotheses, 82*(4), 405–411. doi:10.1016/j.mehy.2013.11.030.

Master, Z., Campo-Engelstein, L., & Caulfield, T. (2014). Scientists' perspectives on consent in the context of biobanking research. *European Journal of Human Genetics*. doi:10.1038/ejhg.2014.143.

Mathaiyan, J., Chandrasekaran, A., & Davis, S. (2013). Ethics of genomic research. *Perspectives in Clinical Research, 4*(1), 100. doi:10.4103/2229-3485.106405.

McGuire, A. L., Achenbaum, L. S., Whitney, S. N., Slashinski, M. J., Versalovic, J., Keitel, W. A., et al. (2012). Perspectives on human microbiome research ethics. *Journal of Empirical Research on Human Research Ethics: An International Journal, 7*(3), 1–14. doi:10.1525/jer.2012.7.3.1.

McGuire, A. L., Colgrove, J., Whitney, S. N., Diaz, C. M., Bustillos, D., & Versalovic, J. (2008). Ethical, legal, and social considerations in conducting the Human Microbiome Project. *Genome Research, 18*(12), 1861–1864. doi:10.1101/gr.081653.108.

McNeely, C. L., & Hahm, J. (2014). The Big (Data) Bang: Policy, prospects, and challenges. *Review of Policy Research, 31*(4), 304–310. doi:10.1111/ropr.12082.

Mello, M. M., Francer, J. K., Wilenzick, M., Teden, P., Bierer, B. E., & Barnes, M. (2013). Preparing for responsible sharing of clinical trial data. *New England Journal of Medicine, 369*(17), 1651–1658. doi:10.1056/NEJMhle1309073.

Mittelstadt, B. D., Fairweather, N. B., McBride, N., & Shaw, M. (2011). Ethical issues of personal health monitoring: A literature review. In *ETHICOMP 2011 conference proceedings* (pp. 313–321). Presented at the ETHICOMP 2011, Sheffield, UK.

Mittelstadt, B. D., Fairweather, N. B., McBride, N., & Shaw, M. (2013). Privacy, risk and personal health monitoring. In *ETHICOMP 2013 conference proceedings* (pp. 340–351). Presented at the ETHICOMP 2013, Kolding, Denmark.

Mittelstadt, B. D., Fairweather, N. B., Shaw, M., & McBride, N. (2014). The ethical implications of personal health monitoring. *International Journal of Technoethics, 5*(2), 37–60.

Mittelstadt, B. D., Stahl, B. C., & Fairweather, N. B. (2015). How to shape a better future? Epistemic difficulties for ethical assessment and anticipatory governance of emerging technologies. *Ethical Theory and Moral Practice*, 1–21. doi:10.1007/s10677-015-9582-8.

Moor, J. (1985). What is computer ethics?*. *Metaphilosophy, 16*(4), 266–275. doi:10.1111/j.1467-9973.1985.tb00173.x.

Moore, P., Xhafa, F., Barolli, L., & Thomas, A. (2013). Monitoring and detection of agitation in dementia towards real-time and big-data solutions. *2013 Eighth international conference on P2p, parallel, grid, cloud and internet computing (3pgcic 2013)*, pp 128–135. doi:10.1109/3PGCIC.2013.26

Mora, F. (2012). The demise of google health and the future of personal health records. *International Journal of Healthcare Technology and Management*, *13*(5), 363–377. Accessed 11 Nov 2014.

National Science Foundation. (2014). Critical techniques and technologies for advancing big data science & engineer (BIGDATA): Program Solicitation NSF 14-543. http://www.nsf.gov/pubs/2014/nsf14543/nsf14543.pdf. Accessed 17 Oct 2014.

NHS England (2014). NHS England. The care.data programme: better information means better care. http://www.england.nhs.uk/ourwork/tsd/care-data/. Accessed 11 Nov 2014.

Niemeijer, A. R., Frederiks, B. J., Riphagen, I. I., Legemaate, J., Eefsting, J. A., & Hertogh, C. M. (2010). Ethical and practical concerns of surveillance technologies in residential care for people with dementia or intellectual disabilities: An overview of the literature. *International Psychogeriatrics, 22*, 1129–1142.

Nissenbaum, H. (2004). *Privacy as contextual integrity* (SSRN Scholarly Paper No. ID 534622). Rochester, NY: Social Science Research Network. http://papers.ssrn.com/abstract=534622. Accessed 12 Mar 2013.

Noddings, N. (2013). *Caring: A relational approach to ethics and moral education*. Berkeley: University of California Press.

Nuffield Council on Bioethics. (2015). *The collection, linking and use of data in biomedical research and health care: Ethical issues* (p. 198). Nuffield Council on Bioethics. http://nuffieldbioethics.org/wp-content/uploads/Biological_and_health_data_web.pdf.

Nunan, D., & Di Domenico, M. (2013). Market research and the ethics of big data. *International Journal of Market Research, 55*(4), 505. doi:10.2501/IJMR-2013-015.

Oboler, A., Welsh, K., & Cruz, L. (2012a). The danger of big data: Social media as computational social science. *First Monday*, *17*(7). https://www.scopus.com/inward/record.url?eid=2-s2.0-84867308941&partnerID=40&md5=0e4cb2f657154c7f82a76c2a657259ab.

Oboler, A., Welsh, K., & Cruz, L. (2012b). The danger of big data: Social media as computational social science. *First Monday*, *17*(7). http://journals.uic.edu/ojs/index.php/fm/article/view/3993. Accessed 1 Oct 2014.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. London: Viking.

Patterson, M. E., & Williams, D. R. (2002). *Collecting and analyzing qualitative data: Hermeneutic principles, methods and case examples* (Vol. 9). Champaign, IL: Sagamore Publishing, Inc. http://www.treesearch.fs.fed.us/pubs/29421. Accessed 7 Nov 2012.

Pellegrino, E. D., & Thomasma, D. C. (1993). *The virtues in medical practice*. New York: Oxford University Press.

Prainsack, B., & Buyx, A. (2013). A solidarity-based approach to the governance of research biobanks. *Medical Law Review, 21*(1), 71–91. doi:10.1093/medlaw/fws040.

Puschmann, C., & Burgess, J. (2014). Big data, big questions metaphors of big data. *International Journal of Communication*, *8*(0), 20. Accessed 7 Oct 2014.

Reuters. (2014, October 3). Facebook plots first steps into healthcare. http://www.telegraph.co.uk/technology/facebook/11139606/Facebook-plots-first-steps-into-healthcare.html. Accessed 15 Nov 2014.

Richards, N. M., & King, J. H. (2013). Three paradoxes of big data. *Stanford Law Review Online*, *66*, 41. Accessed 18 Feb 2015.

Rothstein, M. A., & Shoben, A. B. (2013). An unbiased response to the open peer commentaries on "Does Consent Bias Research?". *The American Journal of Bioethics, 13*(4), W1–W4. doi:10.1080/15265161.2013.769824.

Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C., et al. (2006). Toward a national framework for the secondary use of health data: An American medical informatics association white paper. *Journal of the American Medical Informatics Association, 14*(1), 1–9. doi:10.1197/jamia.M2273.

Schadt, E. E. (2012). The changing privacy landscape in the era of big data. *Molecular Systems Biology*, *8*. doi:10.1038/msb.2012.47

Schaefer, G. O., Emanuel, E. J., & Wertheimer, A. (2009). The obligation to participate in biomedical research. *JAMA*, *302*(1), 67–72. Accessed 19 Mar 2015.

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

Schroeder, R., & Cowls, J. (2014). Big data, ethics, and the social implications of knowledge production. http://dataethics.github.io/proceedings/BigDataEthicsandtheSocialImplicationsofKnowledgeProduction.pdf. Accessed 2 Oct 2014.

Schwandt, T. A. (2000). Three epistemological stances for qualitative inquiry: Interpretivism, hermeneutics, and social constructionism. *Handbook of qualitative research* (pp. 189–214). Thousand Oaks, CA: Sage.

Shilton, K. (2012). Participatory personal data: An emerging research challenge for the information sciences. *Journal of the American Society for Information Science and Technology, 63*(10), 1905–1915. doi:10.1002/asi.22655.

Slashinski, M. J., McCurdy, S. A., Achenbaum, L. S., Whitney, S. N., & McGuire, A. L. (2012). "Snake-oil," "quack medicine," and "industrially cultured organisms:" biovalue and the commercialization of human microbiome research. *BMC medical ethics*, *13*(1), 28. Accessed 13 Oct 2014.

Slote, M. (2007). *The ethics of care and empathy* (New Ed edition.). London, New York: Routledge.

Steinsbekk, K. S., Ursin, L. Ø., Skolbekken, J.-A., & Solberg, B. (2013). We're not in it for the money—lay people's moral intuitions on commercial use of "their" biobank. *Medicine, Health Care and Philosophy, 16*(2), 151–162. doi:10.1007/s11019-011-9353-9.

Taylor, L., & Floridi, L. (Eds.). (2015). *Group privacy: New challenges of data technologies*. New York: Springer (forthcoming).

Tene, O., & Polonetsky, J. (2013). Big data for all: Privacy and user control in the age of analytics. http://heinonlinebackup.com/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/nwteintp11&section=20. Accessed 2 Oct 2014.

Terry, N. (2012). Protecting patient privacy in the age of big data. *UMKC L. Rev.*, *81*, 385. Accessed 2 Oct 2014.

Terry, N. (2014). Health privacy is difficult but not impossible in a post-hipaa data-driven world. *Chest, 146*(3), 835–840. doi:10.1378/chest.13-2909.

The NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., et al. (2009). The NIH human microbiome project. *Genome Research, 19*(12), 2317–2323. doi:10.1101/gr.096651.109.

van den Berg, B., & van der Hof, S.. (2012). What happens to my data? A novel approach to informing users of data processing practices. *First Monday*, *17*(7). doi:10.5210/fm.v17i7.4010

van der Sloot, B. 2014). Privacy in the Post-NSA Era: Time for a fundamental revision? http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2432104. Accessed 17 Feb 2015.

Watson, R. W. G., Kay, E. W., & Smith, D. (2010). Integrating biobanks: Addressing the practical and ethical issues to deliver a valuable tool for cancer research. *Nature Reviews Cancer, 10*(9), 646–651. doi:10.1038/nrc2913.

Wellcome Trust. (2013). *Impact of the draft European data protection regulation and proposed amendments from the rapporteur of the LIBE committee on scientific research*. Wellcome Trust. http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/WTP055584.pdf. Accessed 22 Oct 2014.