BIG DATA: NEW SCIENCE, NEW CHALLENGES, NEW DIALOGICAL OPPORTUNITIES

by Michael Fuller

Abstract. The advent of extremely large data sets, known as "big data," has been heralded as the instantiation of a new science, requiring a new kind of practitioner: the "data scientist." This article explores the concept of big data, drawing attention to a number of new issues—not least ethical concerns, and questions surrounding interpretation—which big data sets present. It is observed that the skills required for data scientists are in some respects closer to those traditionally associated with the arts and humanities than to those associated with the natural sciences; and it is urged that big data presents new opportunities for dialogue, especially concerning hermeneutical issues, for theologians and data scientists.

Keywords: analysis; big data; data scientist; ethics; hermeneutics; interpretation

What Is Big Data?

With the exponential growth in the use of computers in all walks of life, from social media to e-commerce to record-keeping to collecting experimental results in a laboratory, immense quantities of data are now routinely being collected and stored. The term "big data" has been used in recent years to denote the immensely large data sets being generated by these activities. What constitutes "big" in this context is constantly changing, due to the ever-increasing capacity of machines to store and manipulate data, and a rigorous definition of big data is therefore problematic (Mayer-Schönberger and Cukier 2013, 6). However, it is often considered to be characterized by "Three Vs:" volume, variety, and velocity (Laney 2012).[1] There is a huge volume of such data: it is highly varied; and it is accumulated at great speed. To these three Vs a fourth has recently been added: value (Chen et al. 2014, 1). It is in these respects that big data differs from more conventional data sets, which are generally more focused, more limited in their scope, and more easily interrogated by the standard tools developed for this purpose by statisticians. A big data set is likely to be comprised of data relating to a large number of parameters, collected with little "filtering" or

Michael Fuller is a teaching fellow at New College, University of Edinburgh, Mound Place, Edinburgh EH1 2LX, UK; e-mail: Michael.Fuller@ed.ac.uk.

 ISSN 0591-2385

standardization regarding its content or format. The difference between the two may be illustrated by comparing, for example, responses to a survey on a particular issue, which has been distributed to a known number of people and which requires them to give answers to a specific set of questions, with all the terms put into an Internet search engine over a 24-hour period. The latter data set is likely to be a great deal bigger and a great deal more varied; and its interrogation in search of useful information will be a great deal less straightforward.

The aim of this article is to survey briefly some of the issues surrounding big data, and to underline some of the potential concerns which it raises.[2] Accepting the premise that big data presents new kinds of challenges and opportunities which extend beyond the conceptual territory hitherto considered to be the preserve of science, and hence that it represents a new kind of science, this article then urges that the skills which will be required by the "data scientists" of the future do not lie exclusively within the scientific academy, and that hermeneutical skills, such as those developed, utilized, and analyzed within disciplines like theology, will have an important role to play in their work. There is thus the potential to open up important new areas of dialogue between theologians and data scientists.

Applications of Big Data

Big data potentially has a huge variety of applications. By way of illustration, consider the following examples.

Big data in scientific research. Some scientific work requires the accumulation of huge amounts of data. In the field of astronomy, for example, it has been reckoned that the Sloan Digital Sky Survey (SDSS), based in New Mexico, accumulated more data within a few weeks of its commencing operations in 2000 than had previously been acquired in the entire history of astronomy—and that a scheduled future astronomical project will acquire every 5 days the amount of data accrued by the SDSS over the course of 10 years (Mayer-Schönberger and Cukier 2013, 7). Similarly, the huge amounts of information generated through scientific research programs such as gene sequencing, and the experiments being undertaken at the Large Hadron Collider at CERN (Chen et al. 2013, 21–23), have the potential, on analysis, to advance human knowledge and understanding greatly in the areas of research they are addressing.

Commercial exploitation of big data. Information about the purchasing habits of individuals and communities, and also information about the manufacturing and distribution of goods, and about the internal organization of large companies, is of tremendous commercial significance

(Davenport and Dyché 2013). A great many financial transactions now routinely take place on the Internet, enabling data to be collected by those selling goods and services; and the use of store cards and credit cards in retail outlets also allows information about individuals' purchasing habits to be compiled. Companies can use big data of this kind in an "inward-facing" way, to streamline their own internal processes and make them more efficient, and in an "outward-facing" way, using data retrieved from customers to analyze those customers' preferences and hence to advertise products which those customers are likely to want to buy. Users of retail websites such as Amazon will be aware that items for which they search are routinely logged, and that the information thus supplied is used to make recommendations for future purchases. Most users of such websites will probably not find this use of the data they supply problematic, and may indeed find it helpful.

Big data in medical research. Information gleaned from hospital records, and from the results of medical trials, is potentially of great importance in the forecasting, diagnosis, and treatment of disease (Chen et al. 2013, 73–4). This makes the large data sets that may be gathered through the scrutiny of patient records extremely valuable for pursuing scientific research; and it also means that they constitute a valuable resource from the perspective of those who might wish to exploit such research for profit, such as pharmaceutical and insurance companies. There are, of course, standard ethical issues around the collection and storage of such data which pre-date the big data age (cf. British Medical Association 2012, esp. chapters 2 and 5); but big data introduces new issues of its own, some of which will be explored briefly below.

An example of the use of big data. At this point, a much-commented-upon example will perhaps serve to illustrate some of the potential benefits of big data, and some of its potential pitfalls. The Google Flu Trends (GFT) service is an attempt by the Internet search service Google to make accurate predictions about flu outbreaks (Ginsberg et al. 2009). In the United States, these outbreaks can be tracked through the analysis made by the Centers for Disease Control and Prevention (CDC) of the visits made to physicians by flu sufferers: these data are published with a lag time of 1–2 weeks. A number of common search queries were analyzed to discover which best fitted this CDC data, using archived information on Google usage. Attempts were then made to provide a much quicker analysis of the occurrence of flu epidemics, with the aim of "enabl[ing] public health officials and health professionals to better respond to seasonal epidemics" (Ginsberg et al. 2009, 1013). Subsequent work suggested that "Google Flu Trends can provide timely and accurate estimates of the influenza activity

in the United States, especially during peak activity, even in the wake of a novel form of influenza" (Cook et al. 2011, 7).

However, a more recent U.S. flu epidemic in winter 2012–2013 was presaged by "drastically overestimated peak flu levels" predicted by GFT (Butler 2013, 155). Although it is likely that the refinement of the methods used in making GFT predictions in the light of such wayward results will subsequently reduce their occurrence, it has been suggested (Lazer et al. 2014) that two issues have been particularly significant in producing these errors: big data hubris and algorithm dynamics. "'Big data hubris' is the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis. . . . The core challenge is that most big data that have received popular attention are not the output of instruments designed to produce valid and reliable data amenable for scientific analysis" (Lazer et al. 2014: 1203). In addition to such hubris, the more practical point is made that Google constantly modifies the algorithms it is using, in order to try to improve its service to its customers (for example, by making recommendations for further searches) and in response to its business model (for example, by promoting advertising revenue). Such algorithm modification has the potential to introduce instability into predictive systems such as GFT. Lazer et al. (2014) suggest that big data offers significant possibilities, but that there remain issues around transparency (those working on GFT do not give the search terms they identified as most significant), understanding how algorithms change over time, and integrating research on big data and the more traditional "small data"—that is to say, data collected in traditional ways, by repeatable means.

In the work summarized here by those engaged with GFT, and by those offering a critique of their approach, we may perceive the beginnings of a critical interrogation of big data. This is a topic to which we shall return in the section on hermeneutics.

Issues Around Big Data

Big data sets present a number of new challenges, many of which relate to ethical and interpretative issues. These challenges are particularly acute when the data involve information gathered from individual people.

Ethical challenges (i): Confidentiality and consent. In both the generation and storage of big data sets containing personal information, there are issues raised around the confidentiality of the information involved. "Confidentiality" here involves preserving information entrusted to the generator of the data set from access by an unauthorized third party.

To some extent, it may be possible to preserve confidentiality through processes of anonymization or "deidentification"—that is to say, "stripping information from a data record that might link the record to the public name of the record's subject" (Berman 2013: 28). Such a process would need to be irreversible if confidentiality is to be maintained (which may then limit the uses of the data). However, it has been suggested that given enough data, anonymization is all but impossible, as it is possible to triangulate from multiple "anonymized" data sets and hence re-establish individual identities (Mayer-Schönberger and Cukier 2013, 154ff). Hence the rise of big data raises new issues around the practicability of maintaining individual confidentiality. (We may note in passing that in the case of the GFT research cited above, the researchers state that "None of the queries in the Google database for this project can be associated with any particular individual" (Ginsberg et al. 2009, 1014).)

Similarly, new issues are raised by big data around the obtaining of informed consent for research. This is a routine procedure in the conduct in all kinds of research, medical, psychological, sociological, and so on, involving human subjects (cf. British Medical Association 2012, 59–63). There are straightforward practical problems here, since the generation and storage of such consent forms is itself a big data issue, and so the questions already raised around confidentiality apply to them; also, obtaining consent can be a very expensive process in terms of the time involved (Berman 2013, 190). There is also a more theoretical problem that arises with the storage of data which may then be open to revisiting for future research. To what exactly are people consenting? Suppose I give a blood sample as part of a research program seeking information about susceptibility to a particular infectious disease. If the information gleaned from that sample is retained as part of a large database, and is subsequently used in research relating to cancer, should my consent be re-sought for this new research? In general, if data is being retained for (potentially) multiple uses, some of which may not be known at the time when consent was obtained, to what extent can the consent given be truly "informed?" Similarly, if blanket consent were to be sought from individuals for potentially any application of data derived from their submissions to a research project, to what extent could consent given under such circumstances be said to be "informed?"

It has been suggested that a "privacy framework for the big data age" will be "one focused less on individual consent at the time of collection and more on holding data users accountable for what they do" (Mayer-Schönberger and Cukier 2013, 173). But this itself raises a new set of questions, not least regarding how boundaries for accountability may be set, and how legislation might be enacted on an international basis to ensure that such accountability could be maintained when data from the population of more than one nation-state is involved. Considerably more

work is needed in the areas of confidentiality and consent if the new challenges presented by big data are to be fully addressed.

Ethical challenges (ii): Storage and access. "Over the past half-century, the cost of digital storage has been roughly cut in half every 2 years, while the storage density has increased 50 million-fold" (Mayer-Schönberger and Cukier 2013, 101). It is to be expected that these trends—reducing expense and increasing capacity—will continue, with the result that more and more data will be accumulated and stored. This presents at least two potential problems concerning the secure storage of big data: first, ensuring that data are neither accessed without authorization nor stolen; and second, ensuring that it remains uncorrupted. The widely reported hacking of Sony's computer systems in December 2014 and the consequent revelation of confidential information on the Internet, in a supposed "revenge" attack over its proposed release of a movie depicting the assassination of the North Korean leader Kim Jong-un, illustrates how real the first of these dangers is, and anyone who has been the victim of a computer virus will realize how vulnerable computer systems can be to corruption and degradation. The secure storage of data is an ongoing problem, and is likely to remain one indefinitely as both security systems and the ingenuity of hackers become more sophisticated.

Leaving aside questions arising from real or potential criminal activity, it has been said that "the most important question confronting the future of Big Data [is]: will it be open (to the public) or closed (to all but the data owners)?" (Berman 2013, 224). Given the variety of possible readings that may be generated by big data (see below), it is important that that data be readily available to anyone who might wish to interrogate it, not least in order to confirm or challenge conclusions which have been derived from it. For a variety of reasons, Berman is pessimistic in his appraisal of this issue, concluding that for commercial, practical, and propagandistic reasons there are vested interests involved in restricting access to big data (Berman 2013, 224), to the extent that the kind of free access that is desirable is unlikely to materialize in reality. This in turn creates a raft of questions around the credibility of ideas advanced on the basis of big data, if the data in question are not available for scrutiny. (It is, after all, axiomatic that scientific experiments should be repeatable: to publish results without accounting in detail for their provenance would, in classical scientific contexts, be impermissible.)

In all the areas discussed above, there is an opportunity for specifically Christian ethical engagement with the challenges brought by big data. There are important opportunities here for dialogue between theologians and those involved with the management of big data. Such dialogue may not be straightforward, and where business issues are concerned it may

involve a re-examination of the roots of contemporary economic life (cf. Stackhouse 2001). This is an important area for future research.

Interpretative challenges: Hermeneutics. Perhaps the most significant of all the challenges relating to big data are those that cluster around its interpretation. Very big data sets can offer support for any number of responses to any number of questions used to interrogate it, as Berman (2013, 145) points out: "When the amount of data is sufficiently large, you can find almost anything you seek lurking somewhere within. . . . Also, whenever you select a subset of data from an enormous collection, you may have no way of knowing the relevance of the data you excluded." He adds that "Data scientists walk a thin line. If they start their project with a preconceived theory, then they run the risk of choosing a data set that confirms their bias. If they start their project without a theory, then they run the risk of developing a false hypothesis that happens to fit the data" (Berman 2013, 147). Berman surveys various pitfalls of big data analysis, including the biases that the analytical techniques used can themselves introduce. His conclusion should give us pause for thought: "Technically, Big Data does not produce answers. At best, Big Data points us in the direction of answers and inspires new questions. At worst, Big Data pretends to give us answers when it cannot" (Berman 2013, 226).

We will return to the important issue of interpretation in the sections on the data scientist and hermeneutics below.

The Distinctiveness of Big Data

We have seen that big data raises new ethical and interpretative questions and concerns, above and beyond those raised by traditional means of research. A number of further methodological and conceptual distinctions between the "big data" approach to questions and the approaches which it has been more customary for the sciences to take may be noted, supporting the idea that a new kind of science is emerging with the arrival of big data.

Classically, science proceeds from the asking of questions to the construction of experiments, and hence to the generation of data. Where big data is concerned, this process is reversed: we begin with the data (however it has been accumulated) and then interrogate it, and what is discovered as a result will be critically dependent on the way in which that interrogation is carried out (cf. Mayer-Schönberger and Cukier 2013, 72). This is an interesting new gloss on the idea expressed by Ian Barbour (1998, 108), that "all data are theory-laden": his aphorism may originally have related to the acquisition of data, but it applies equally to their interrogation.

Further, it has been observed that the insights derived from big data relate not to causation, but rather to correlation: as Mayer-Schönberger and Cukier put it (2013, 52), "Knowing *what*, not *why*, is good enough." This

again signals a very different approach to that of much human endeavor, not least in the sciences, which seeks to discern the causes of things, the reasons why they happen. It denotes a new, and very different, mindset from that which has governed much scientific practice in the past. The same authors also suggest that "the biggest impact of big data will be that data-driven decisions are poised to augment or overrule human judgment" (Mayer-Schönberger and Cukier 2013, 141), with "subject-area specialists" becoming less important than data analysts: the human corollary, perhaps, of the privileging of "what" over "why."

Mayer-Schönberger and Cukier (2013, 108) also point out that "With big data, the sum is more valuable than its parts, and when we recombine the sums of multiple datasets together, that sum too is worth more than its individual ingredients." This is suggestive of an holistic perspective of the kind that is increasingly infusing scientific outlooks, such as that of systems biology (cf. Peacocke 1993, 41ff.; Clayton 2004, 89 ff.), in defiance of the reductionism that has been assumed by some to be inseparable from the scientific method (Barbour 1998, 78).

More negative corollaries of the increased use of big data have also been advanced. Mayer-Schönberger and Cukier (2013, 173) have suggested that, should big data be used in the future to make predictions regarding the likelihood of individuals behaving in particular ways, this might lead to the erosion of our understanding of free will. Their doomsday scenario, involving people being arrested for crimes in advance of those crimes actually being committed, might appear rather fanciful; but their conclusion—that "as society assigns individual responsibility (and metes out punishment), human volition must be considered inviolable" (Mayer-Schönberger and Cukier 2013, 193)—can only be applauded. It serves, perhaps, as an interesting contribution to the arguments surrounding free will, not least those which cast doubt on its existence (cf. Murphy and Brown 2007).

Most disturbing of all, perhaps, is the risk of a collapse of "values" into "value." It is noteworthy that Chen et al. (2014, 5f.) appear on occasion to use these words synonymously: this may simply be an error of translation, but the implication—that the sole (ethical) "value" to be derived from big data lies in the (fiscal) "value" generated by it—may be inferred from much of the literature on big data, focusing as it does on the financial benefits that may be derived from the exploitation of big data. This is, to say the least, grounds for concern.

The Data Scientist

A data set, however huge, means nothing until it is analyzed. Data analysis is a commonplace in the sciences, but hitherto has generally involved small, discrete sets of data. Because a new skills set is emerging for those tasked with dealing with big data, the term "data scientist" has been suggested for

such individuals (Mohanty, Jagadeesh, and Srivasta 2013, 251ff.). What is it that characterizes the role of the data scientist, differentiating it from that of the "classical" scientist?

A data scientist is defined by Mohanty et al. (2013, 253) as "a person who takes raw materials (in this case data) and uses skill, knowledge, and vision to craft it into something of unique value." They further note that "the data scientist must have the ability to bring scenarios to life by using data and visualization techniques: this is nothing but storytelling" (Mohanty et al. 2013, 255). The profession of "data scientist" is said by Mayer-Schönberger and Cukier (2013, 125) to combine "the skills of the statistician, software programmer, infographics designer, and storyteller." Davenport and Dyché (whose research was conducted among big companies) similarly observe that "[a] key skill involves being able to explain big data to executives . . . several interviewees commented that their quantitative people need to 'tell a story with data'" (Davenport and Dyché 2013, 14). These authors all clearly recognize that the interpretation of data is a *skill*, and that it involves the ability to *tell a story*. This brings us back to the topic of *interpreting* big data, and of presenting interpretations of big data in ways that are comprehensible to others. These are attributes of the data scientist which do not necessarily obtain in more traditional sciences, since the data derived from a single experiment might be expected to be relatively unambiguous, and the presentation of experimental results is more likely to take place in contexts where most of those present will be specialists who are familiar with their provenance and significance.

Data sets contain information (of many kinds), which, when the data are interpreted, must be expressed in language. It has been pointed out that even if that language is of a technical or mathematical kind, this leaves the information open to hermeneutical study (Diamante 2014, 187). Indeed, Hans-Georg Gadamer observed that "What is established by statistics seems to be the language of facts, but which questions these facts answer and which facts would begin to speak if other questions were asked are hermeneutical questions" (Gadamer 1976, 11). Hermeneutics is, of course, familiar territory to the theologian, who is heir to a long tradition of interpretation and reflection upon interpretation relating to the texts of scripture. Are there particular skills which the theologian, versed in hermeneutics, might advance as being of benefit to the data scientist? And does this lead to a fresh new way in which the dialogue between scientists and theologians might be explored?

Hermeneutics: "How We Read, Understand and Handle Texts"[3]

Theologians in the Judeo-Christian tradition have wrestled with the texts of their scriptures for centuries. From this engagement in biblical

hermeneutics, the following seven points emerge: (1) hermeneutics is an *interdisciplinary* exercise, which involves combining the insights of theological, philosophical, literary, linguistic, and other areas of academic study; (2) hermeneutics concerns issues of *meaning*, and the generation of meaning through the interaction of reader and text; (3) hermeneutics is a *practice*, requiring the exercise of skill and the insights of wisdom in engaging with a text; (4) hermeneutics requires the recognition of the historical *context* of the text under consideration; (5) in parallel with this recognition, hermeneutics requires the acknowledgment of the reader's own prejudices and preconceptions, which are themselves the product of the reader's social, intellectual, and temporal location (the basis of Ricoeur's "hermeneutics of suspicion," cf. Thiselton 2009, 233); (6) hermeneutics stresses the role of the community in providing a common understanding in the forming and framing interpretations; (7) hermeneutics rejects the possibility of unique, objective meanings. As Thiselton (2009, 226) puts it, "Everything is hermeneutical; everything requires interpretation."

How might these principles map onto the interpretation of big data?

An interdisciplinary exercise. In order to maximize the benefit that may be obtained from the analysis of big data, it is important to have a number of different perspectives on it. In particular, data originating in particular scientific, clinical, or commercial contexts is very likely to require the involvement of appropriate subject specialists in its interpretation. It may well be that a simple consideration of the statistics throws up odd correlations, the interpretation of which requires an expert eye. (In an amusing case reported by Mayer-Schönberger and Cukier [2013, 67], statisticians commissioned by a used-car trader found that cars painted orange had fewer engine defects than those painted other colors. It is not immediately obvious what interpretation might be placed on this correlation; in generating any such interpretation, expertise from the manufacturers and distributors of cars is likely to be necessary.)

Issues of meaning. Just as meaning emerges from an interaction of reader and text, likewise meaning emerges from the interaction of data and an analyst. In both cases, it is important to pay close attention to the meaning of that meaning. Texts tell us stories: so too do data. Any meanings derived from the presentation of those stories—the "messages" or "morals" they are presenting—require careful and scrupulous consideration. This is particularly the case when it comes to the telling of stories derived from big data, which are likely to be accompanied by graphics—charts, graphs, and diagrams—which may themselves embody, in an impressive or even apparently irrefutable form, biases or misreadings of big data which are consequent on the method of its analysis.

The practice of data analysis. As already noted, the treatment of data requires skill. Understanding, interpreting, and presenting data to others will require future data scientists to be trained, and to practice their craft in the expectation that their skills, and the wisdom they bring to bear in their work, will increase with experience. ("Wisdom" here may usefully be understood as "an interpretation of knowledge that is not separated from the ethical claims of truth and goodness" [Deane-Drummond 2000, 153]). Not only this: as Michael Polanyi commented, "Making sense of experience is a skillful act, which impresses the personal participation of the scientist on the resulting knowledge" (Polanyi 1958, 60). It is important to recognize that as in the conduct of experiments, so in the analysis of data, the individual scientist brings a unique perspective to bear, and leaves in consequence a unique personal imprint on the results he or she generates.

The context of data. Just as it is important to have some understanding of the original historical, cultural, and social context of a text if any interpretation of it is to be of value, so it is important to understand as fully as possible the provenance of any data that are being used if maximum benefit is to be obtained from them. Many phenomena have a complex etiology and cannot be fully understood unless as wide a picture of them as possible is obtained. Contributory factors to a particular disease, for example, may include genetic, familial, environmental, economic, and other factors, not all of which may be revealed through an analysis based solely on hospital records.

The bias of the analyst. "Interpretation has some specific subjective connotations, such as the reciprocity between interpretation of the text and self-interpretation" (Ricoeur 2013, 26). Interpreters of a text need to interrogate their own historical, cultural, and social contexts, and the expectations that these place on them, in addressing their task. So, too, the interpreters of data need to interrogate the expectations and demands that their contexts place on them. This does not include just their social and cultural contexts, but may well involve such questions as: do future promotions, or salary rises, hinge on my deriving the "right" results from this data? Am I using the data I am analyzing in order to seek some kind of competitive edge in the field in which I am operating, either personally or on behalf of my employer?

The role of community. The meanings of texts are shaped by communities: so too are the meanings of big data. The attitudes, practices, and beliefs of the "data science community," and of any corporate community which data scientists are serving, will inevitably affect the way in which those data scientists operate. This may have a positive role, in reinforcing good practice; or it may have a more negative role, in discouraging

original thinking. Much data is currently treated as "junk," and discarded; for example, it is estimated that less than 0.1% of the data generated by the Large Hadron Collider at CERN is currently collected and stored (Mayer-Schönberger and Cukier 2013, 197). This information is presumably, by a general consensus, reckoned to be the most useful; but who knows what potentially valuable information, currently seen as worthless, might turn out to have been lost as a consequence of this consensus among the community of practitioners working at CERN?

The rejection of unique, objective meanings. To some extent, one might hope to be preaching to the choir on this issue, as far as data scientists are concerned: as we have seen, there is already a realization that any number of meanings may be derived from big data sets. However, given the particular status which is generally accorded to mathematized conclusions, which may well appear to be "objective" in the ways in which they are presented, it is perhaps also incumbent on us to remember at all times that big data can be fallible: that not all the correlations it presents to us will mean anything.

In the case cited earlier in this article—the use of big data by GFT—and in the subsequent analyses of that project's shortcomings, it may be seen that many of these hermeneutical principles are starting to be applied. The work of GFT is interdisciplinary, involving computer scientists and epidemiologists. There is a recognition in the critiques of GFT analyses that any "meaning" derived from the data used requires scrutiny and checking; and a realization, too, that big data originates from a particular context, and is shaped by the algorithms used to collect it. Concerning potential analyst bias, it is noteworthy that Cook et al. (2011) explicitly acknowledge the "competing interests" in their research, in that three of the authors are Google employees. Finally, we may perhaps also see here some evolution of the "big data community" as those engaged in big data research initiate and respond to a particular instance of its use, seeing it as "a case study where we can learn critical lessons as we move forward in the age of big data analysis" (Lazer et al. 2014, 1205).

Conclusion

There is no doubt that the potential of big data is huge; and there is also no doubt that that potential is in part beneficial, in terms of the good that it may bring to humankind. However, big data also gives rise to serious concerns. Some of these concerns relate to privacy and consent, long acknowledged as problematic issues before the arrival of big data but now presenting new questions which need to be addressed: others, not least those around the interpretation of big data, are new. This article has

identified a number of areas where more work is required if these concerns are to be satisfactorily addressed.

The coining of the term "data scientist" indicates that a new profession is emerging, and this is therefore a good time to consider the expectations and responsibilities of its practitioners. It might be urged that particular standards of competence, training, professional validation, and so on, such as obtain in other professions, should apply to them (cf. Resnik 1998, 34 ff.) Additionally, along with the suggestion that data users be held accountable for the exploitation of the data at their disposal, consideration ought perhaps to be given to the idea of a new kind of "Hippocratic Oath" for data scientists. (The original Hippocratic Oath, governing professional behavior for those in the medical profession, may be found in British Medical Association 2012, 887.) This might be a valuable way of foregrounding the ethical issues which are raised by the use of big data, and establishing a set of ethical standards which should be upheld by those charged with its interrogation, interpretation, and presentation.

Given the significant and acknowledged role of *interpretation* in the work of data scientists, there is also considerable scope for important dialogue with theologians and philosophers over this issue, since hermeneutics is a skill which has been developed within these disciplines over several centuries. Here, then, is a dialogical opening for scientists and theologians in which the former have something valuable to offer the latter. Given that the flow of ideas between science and religion can often appear to be "one-way traffic" (Southgate and Poole 2011, 29) from the former to the latter, this perhaps might also be a valuable way of redressing the balance in this dialogue.

It is striking that, despite the origins of their discipline in mathematics and statistics, the particular skills set required of data scientists apparently makes what is sought from them, at least in part, a way of thinking which is traditionally encountered more in faculties of arts than of sciences. If this is so, then the possibility for important, sustained future dialogue between theologians and data scientists should not be underestimated.

Notes

1. This reference is taken from a Wikipedia entry on big data (see bibliography). The website cited therein can only be accessed through registering with the company concerned, and hence submitting personal details to it. This is a noteworthy example of several of the issues raised in this article, for example around privacy, and around the ownership and accessibility of big data.
2. Although the word "data" is plural, the term "big data" is generally treated in the literature as singular, and that custom is followed in this article.
3. The quotation is from Thiselton 2009 (p. 1), to which the following paragraph is greatly indebted.

References

Barbour, Ian G. 1998. *Religion and Science: Historical and Contemporary Issues*. London: SCM Press.

Berman, Jules J. 2013. *Principles of Big Data*. Amsterdam: Elsevier.

British Medical Association. 2012. *Medical Ethics Today: The BMA's Handbook of Ethics and Law*. Oxford: Wiley-Blackwell.

Butler, Declan. 2013. "When Google Got Flu Wrong." *Nature* 494:155–56.

Chen, Min, Shiwen Mao, Yin Zhang, and Victor C. M. Leung. 2014. *Big Data: Related Technologies, Challenges and Future Prospects*. Heidelberg, Germany: Springer.

Clayton, Philip. 2004. *Mind and Emergence: From Quantum to Consciousness*. Oxford: Oxford University Press.

Cook, Samantha, Corrie Conrad, Ashley L. Fowlkes, and Matthew H. Mohebbi. 2011. "Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic." *PLOS ONE* 6. Available at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3158788/pdf/pone.0023610.pdf (accessed 21 February 2015).

Davenport, Thomas H., and Jill Dyché. 2013. "Big Data in Big Companies". SAS Institute. Available at http://www.sas.com/reg/gen/corp/2266746 (accessed 10 July 2015).

Deane-Drummond, Celia. 2000. *Creation through Wisdom*. Edinburgh: T&T Clark.

Diamante, Oscar R. 2014. "The Hermeneutics of Information in the Context of Information Technology." *Kritike* 8:168–89.

Gadamer, Hans-Georg. 1976. *Philosophical Hermeneutics.* Translated by David Linge. Berkeley: University of California Press.

Ginsberg, Jeremy, Matthew M. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature* 457:1012–14.

Laney, Doug. 2012. "The Importance of 'Big Data': A Definition." Cited at http://en.wikipedia.org/wiki/Big_data. Accessed 1 December 2014 (see footnote 1).

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." *Science* 353:1203–05.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2013. *Big Data: A Revolution that Will Transform how We Live, Work and Think*. London: John Murray.

Mohanty, Soumendra, Madhu Jagadeesh and Harsha Srivatsa. 2013. *Big Data Imperatives*. Berkeley, CA: Apress.

Murphy, Nancey, and Warren S. Brown. 2007. *Did My Neurons Make Me Do It? Philosophical and Neurobiological Perspectives on Moral Responsibility and Free Will*. Oxford: Oxford University Press.

Peacocke, Arthur. 1993. *Theology for a Scientific Age*, enlarged edition. London: SCM Press.

Polanyi, Michael. 1958. *Personal Knowledge*. London: Routledge and Kegan Paul.

Resnik, David B. 1998. *The Ethics of Science*. London: Routledge.

Ricoeur, Paul. 2013. *Hermeneutics.* Translated by David Pellauer. Cambridge: Polity Press.

Southgate, Christopher, and Michael Poole. 2011. "Introduction." In *God, Humanity and the Cosmos*, edited by Christopher Southgate, 3rd edition, 3–43. London: T&T Clark.

Stackhouse, Max L. 2001. "Business, Economics and Christian Ethics." In *The Cambridge Companion to Christian Ethics,* edited by Robin Gill, 228–42. Cambridge: Cambridge University Press.

Thiselton, Anthony C. 2009. *Hermeneutics: An Introduction*. Grand Rapids, MI: William B. Eerdmans.