International Journal of Knowledge, Innovation and Entrepreneurship
Volume 5 Nos. 1-3, 2017, pp. 5—20

Developing a Framework for Ethical Big Science

DERYN GRAHAM
Data Analytica, United Kingdom
Received 09 August 2017; received in revised form 08 December 2017; approved 30 December 2017

ABSTRACT This paper describes the development of a framework for ethical big science, through a study of the ethical and societal issues that big science brings. The paper begins by defining big data and data analytics, before discussing some of the sources of big data. Ethical issues, such as privacy (and anonymity) and security are looked at, as well as existing legislation and frameworks for ethical analytics, the emergence of the term "data harm" and the impact of big science on the epistemology of knowledge. Several technological and business cases raising ethical concerns for the big data society are considered; examples include smart meters (Hive), Care.data and DeepMind. Finally, the main points are summarised, discussed, and conclusions drawn, thus leading to the identification of Key Performance Indicators (KPIs) and the proposal of a framework for ethical big science.

Keywords: big science, ethics and society, Key Performance Indicators (KPIs), framework for ethical big science.

Introduction

The emergence of the Big Science phenomenon is grounded in the convergence of technologies, such as Cloud Computing and the Internet of Things (IoT), and the digitization of society and the resulting massive increase in (big) data generation. The information pool, or data lake, that is generated worldwide is said to double every twenty months (Heger, 2014), moving from data lakes to data oceans. However, it is the increasing fusion of the data, not just its generation, which poses challenges for ethics, society, and the epistemology of knowledge. Issues regarding the exploitation of big data and analytics, resonate between questions of "how" and "if" data can or should be exploited. The question of "how" revolves around the pragmatic concerns and the issues of People, Processes and Technologies (Graham, 2016), whilst the question of "if" is intrinsically linked with ethical concerns and the issues of privacy and security.

This paper describes the development of a framework for ethical big science, through a study of the ethical and societal issues that big science brings. The paper begins by defining big data and data analytics, before discussing some of the sources of big data. Ethical issues, such as privacy (and anonymity) and security are looked at, as well as existing legislation and frameworks for ethical analytics, the emergence of the term "data harm" and the impact of big science on the epistemology of knowledge. Several technological and business cases raising ethical concerns for the big data society are considered; examples include smart meters (Hive), Care.data and DeepMind. Finally, the main points are summarised, discussed, and conclusions drawn, thus leading to the identification of Key Performance Indicators (KPIs) and the proposal of a framework for ethical big science.

Definitions

Big Data

Whilst there are several definitions of big data, big data are commonly described as massive heterogeneous data (unstructured, semi-structured and structured) sets, not solvable (manageable data analysis) using conventional data models, such as relational databases (Graham, 2016). All definitions refer to very large data sets, with some combination of five characteristics: Volume, Variety, Velocity, Value and Veracity. Volume is where the amount of data to be stored and analysed is sufficiently large to require special consideration. Variety refers to the data being of multiple types and from multiple sources. Data sources can be structured data held in tables or objects for which metadata is well defined, for example, semi-structured data in documents where the metadata is contained internally (XML documents), or unstructured data such as photographs, video, or any other form of binary data. Velocity refers to the data being produced at high rates and operating on "stale" data is not valuable. Value is where the data has perceived or quantifiable benefit to the enterprise or organisation using it. Finally, veracity is where the correctness of the data can be assessed. McKinsey Global Institute (Neaga and Hao, 2013) suggested models for big data characteristics based on the source, with the main key characteristics being those of volume, velocity, variety and value, with an additional characteristic, veracity. Characteristics of Variability and Complexity have also been added (SAS, 2012). Variability describes the variability of the data flow in addition to its speed. Complexity refers to the data "relationships, such as complex hierarchies and data linkages, among all data" (SAS, 2012, p.3).

These characteristics, particularly value, have an implicit temporal element (data at rest, for example), through associations with definitions of data, information and knowledge, and relationships with established models (heuristic, causal and statistical). Big data that is outside the domain specific state-space is not data specific to a given domain, and as data, it is also, as supported by McKinsey's model (Neaga and Hao, 2013), temporally unspecific. The lack of spatial and temporal constraints is in keeping with big data's features of volume and velocity. The term "Big data" is all encompassing as it fits anywhere and everywhere within the domain specific state-space (Graham, 2015), and, more importantly, outside. Unlike information and knowledge, the value of data is absolute (it is not altered by time). The value of big data (using analytics) is obtained through converting the data into temporally relevant information or knowledge.

Data Analytics

Analytics have been described by their use; Descriptive, Diagnostic, Prescriptive or Predictive. Analytics have also been categorised by their data format and origin; Text analytics, Speech analytics, Video/image analytics and Combined analytics (Marr, 2015, pp. 105-149). Analytics refer to the analysis of data (usually big) to identify patterns or anomalies, and so to provide descriptions, diagnoses, prescriptions, or to make predictions, using techniques such as machine learning (ML), e.g. Artificial Neural Networks (ANNs) modelled on human brain neurons, regression, etc.

Graham (2014) depicted the "transformations" from data to information and then from information to knowledge, discriminating between data, information and knowledge through the dimension of time for the purpose of learning (competence

achievement). Human learning appears to involve the taking in of raw data with a specific goal, organising the data so that it has meaning, and analysing this information (compare and contrast, applying elements of Bloom's (1956) taxonomy) to a more structured form, namely knowledge. Machine learning aims to emulate human learning through deep learning, employing ANNs, and the realisation of meta-knowledge, etc. Such knowledge or expertise is the basis of knowledge-based systems and heuristic knowledge models.

Big data is exploited (value obtained) by the application of analytics, effectively reducing the state-space, "converting" the data into information (contemporary), or knowledge (future predictive) and/or making it domain specific. Big data analytic techniques may lead to the application of established models, such as mathematical (possibly statistical) models or decision trees (which may be part of a knowledge-based model), post processing (filtering, etc).

Analytics are essentially the application of a set of processes (algorithms) and technologies (systems), plus people (skills), to make sense of data. For example, ML algorithms are a process of learning a model of the world to predict future outcomes. The type of analytics used is based on the outcome, e.g. classification or clustering (if the outcome is discrete) for a numerical regression problem. Not all big data is stored, as it is not normally possible or desirable. It is the transformation of data to information and/or knowledge through analytics and the fusion of data where issues of ethics become prevalent.

Sources of Big Data

In order to develop an ethical framework for big science, it is necessary to understand the origins of big data. The distinguishing feature of big data is its source or mode of origin, often more ad hoc than by design in comparison with the established models, where most, if not all of the knowledge embodied within is methodically sought and structured for a specific domain. Big data originates from multiple sources, and is often a bi-product of other things, for example, data stored in conventional databases, in public and private clouds, gleaned through social media interactions, or sensor data generated as a result of the IoT. It has no restrictive characteristics, and is of multiple formats (variety) and veracity.

Big data analytics can exploit data held in the cloud and cloud storage, adding public cloud data to private cloud data (Gordon, 2013). Many technologies and data sources can be combined to be more pervasive and intrusive, e.g. using Closed-Circuit Television (CCTV) and Global Positioning Satellite (GPS).

The Internet of Things

The Internet of Things has been described as "the idea that any item can be embedded with software, sensors and connectivity to exchange data with one another or a central hub" (Palmer, 2015a, p. 14). The IoT has contributed much to the increase in data as a source of big data. An example of the IoT is the smart meter such as Hive (British Gas, 2015), which permits the remote control of heating in the home via the internet. Hive employs a wireless thermostat, a hub (plugged into the customer's broadband router so that the thermostat can connect to the internet and be controlled remotely), a receiver (so that the thermostat can "communicate" with the boiler and vice-versa) and an app (software). Embedded "things" lead to the generation and usually the recording

of significant amounts of data (much of which is from sensors) that are amenable to analytics.

Cloud Computing

Massive amounts of data require storage. Cloud Computing both adds to, as well as provides a solution for, big data characteristics such as velocity, variety, and (particularly) volume. "The Cloud Computing Model, more commonly referred to simply as Cloud Computing or 'The Cloud', provides access to 'clouds' of shared computing resources, such as storage and applications, over a network, usually the Internet" (Graham, 2013, p. 7). Clouds are commonly classified into Public, Private, Hybrid and Community Clouds (Chang and Wills, 2013, pp. 233-234). Cloud offers a solution for the big data characteristic of volume by providing storage, often in the form of the Hybrid Cloud (a combination of public and private cloud, with the necessary interoperability between the two), allowing organisations and individuals flexibility, with access to variable amounts of data storage, as and when required (for analytics for example), through distributed computing. The existence of data in the Cloud, of course, also adds to the total volume of data per se, e.g. Microsoft's OneDrive.

Social Media

Social media platforms like Facebook, Twitter, etc. are generators of interminable, uncontrollable and vast amounts of data. Often used primarily for social communications between friends and family, it has become much more powerful and insidious, and membership terms and conditions and their implications are often seldom fully read or understood.

Sensor Data

Sensor data is created by a wide range of technologies, such as Radio Frequency Identification (RFID) on goods for tracking and tracing in a supply chain, or wearable technologies, such as fitness apps, constantly streaming data.

Mobile Technologies

An exemplar is Global Positioning Satellite (GPS) technology employed in vehicle SAT NAV (Satellite Navigation), but also incorporated within smart phones. Therefore, GPS can be used to accurately locate the geographical position of the vehicles or smart phones and therefore users, providing additional data.

Data Centres

Data Centres are literally as described; large, centralised centres of data. Data Centres can house Electronic Patent Record (EPR) data, for example, Care.data.

Surveillance Data

Surveillance data can be the product of several technologies, such as CCTV, possibly combined with face recognition software or drones for mobile monitoring. These can provide real-time image and location data.

Conventionally Stored Data

Examples include legacy data in Relational databases, Object-oriented databases, or knowledge-bases, applying heuristic or statistical models, such as for Google DeepMind. The data is structured and so manageable by conventional means including normalization, relational algebra, propositional logic or predicate calculus, and statistical analysis.

Simulation

Simulation data can be produced and used by Automatic Testing Equipment (ATE) for instance, applying causal reasoning for computer hardware fault diagnosis.

Ethical Issues

Now that big data and data analytics have been defined and the sources of big data discussed, the ethical issues can now be considered, such as those pertaining to data anonymity, privacy and security. These ethical and societal impacts of Big Science are then summarised in Table 1.

There are several issues regarding the ethical exploitation of big data and analytics, predictive or otherwise, although it is often prediction that raises most concern, especially in relation to privacy (Marr, 2015, pp.149-154). The European Commission makes reference to a Big Data Value Association Hexagon with Skills, Legal, Technical, Applications, Business and Social components (de Lama, 2016, p. 10).

Vast quantities of data are increasingly being generated, collected, aggregated, linked, analysed, stored and shared. Individuals may generate and reveal some of this data by choice through social media platforms and e-mail, or as a condition of activities such as banking or travel, where disclosure is compulsory. However, much more data is a consequence of the digitization of society, and is collected by an array of sensors, through smart phones, RFID, CCTV, etc. A study by Hewlett-Packard in 2015 (Cate, 2016, p. 4) suggests that "9/10 of the most popular internet-connected devices carry personal data". The generation of data is further magnified with the growth of the IoT; much of this data is simply produced as a by-product of its modus operandi, and some of it is, again, individually instigated, for example, through personal fitness apps.

The question of "if" (ethics) data should be exploited is not only relevant to big data per se, but to Cloud Computing (in terms of the availability of data and its storage) and the Internet of Things (its incessant and infinite generation). The "if" revolves around maintaining privacy and therefore anonymity, which has been found to be much more difficult than anticipated.

Data Anonymity, Privacy and Security

Big data analytics usually require contextual data which is likely to necessitate the use of personal data, e.g. sex, age and postcode, to enable meaningful analysis. In some cases, this contextual data could lead to the identification of individuals and so to a lack of privacy for the individuals concerned. Algorithms exist for making data anonymous through adding layers (which aim to "bury" the data's origin) and the use of fictive data (CloverETL, 2017). However, it is doubtful that such algorithms would

work in all cases where exceptional data attributes (those resulting from genetic testing for example) all but guarantee the identification of a rare individual. Fusion of data from other sources, effectively re-engineering, can also enable the identification of "anonymous" data. Individuals would also be unwilling to contemplate the provision of any of their data to any party (whether or not they are authorised) if there is any security risk, for example, disclosure, data loss or damage, etc.

An example of this ethical dilemma is the NHS Care.data initiative (Palmer, 2015). Care.data is the controversial data harvesting of big data and analytics programme by the NHS in England. Proponents such as Tim Kelsey (Palmer, 2015) argue that it will improve patient care, whilst opponents including MPs and GPs have raised concerns over patient privacy. In order to make causational sense of the data (analytics), demographic data such as age and sex may be necessary, whilst postcodes are needed for epidemiological studies.

Another very recent example is the transfer of 1.6 million medical records to Google from the Royal Free London NHS Trust in the UK (Burton, 2017). The transfer of identifiable patient records took place in September 2015, when patient data was shared with Google DeepMind artificial intelligence subsidiary for the testing of an application called 'Streams' to improve the care of patients with chronic kidney disease. Google claimed it needed the five years of patient records for trend analysis and the detection of historical tests and diagnoses affecting patient care. This transfer of highly confidential records was deemed to be inappropriate, and its purpose was not for the provision of direct care to patients, but for the testing of the Streams application. "Implied consent is only an appropriate legal basis for disclosure of identifiable data for the purposes of direct care if it aligns with the people's reasonable expectations, i.e. in a legitimate relationship, and that there was no legal basis for implied consent for direct care by patients", according to the National Data Guardian, Dame Fiona Caldicott.

Fundamental ethical concerns of big data, analytics, the Cloud and the IoT are the why, the what, the how and who of the collection, storage and access to large quantities of (combined) data. In the case of: Who should collect, store, access (control) and extract value from the data? Hive (British Gas, 2015), for instance; smart meters may enable the benefits of remote control, but are still vulnerable to physical and logical security failures, for example, the loss of internet service in the case of the former, or an attack by hackers, in the case of the latter. Analysis of Hive data for domestic heating control could indicate the occupancy and sleeping periods of customers and indirect information such as employment hours (related to occupancy) or health conditions (where room temperature is constantly high). Such technologies may not be as "innocent" as they initially appear, and combining the IoT, Cloud and other big data through analytics may have serious consequences for privacy.

Ethical and Societal Impacts of Big Science

In Table 1, exemplars of common data sources/stores of data are given, as well as some of the ethical considerations and the implications for society. Ethical considerations tend to be those concerning the individual, whilst societal implications relate more to the population.

Common data source/ store	Exemplar	Ethical Considerations	Societal Implications
IoT	Smart meters (Hive)	Can identify periods of home occupancy; infer employment from usage hours and health from average temperature.	Remote control of devices via the internet. Devices and data could be controlled by parties other than legitimate users, such as hackers or hostile governments.
Cloud	Microsoft OneDrive	External collation of personal artefacts (photographs, movies, music and data).	Ownership of data may not be clear. Photographs, etc. could be exploited by marketing companies or used by paedophiles. Data can be fused for profiling of individuals by organisations and governments. Data fusion could also be used for identity theft, fraud, and other e-crimes.
Social Media	Facebook	Intimate knowledge of individuals by commercial organisations. Additional data accrued through likes and dislikes.	Big Brother*. Cyberbullying. Profiling. Control (often by suggestion, e.g. upselling). Massive security issues if data is lost or hacked. Propagation of "fake news", propaganda, interference by foreign governments.
Sensors	Radio Frequency Identification (RFID)	Abundant sources of RFID, from clothing and other goods, to runners' race tags/numbers. Tracking and tracing of individuals as well as goods.	Issues of mass surveillance and privacy.
Mobile technologies	Global Positioning Satellite (GPS)	GPS can be combined with other technologies, like smart phones, for marketing; nearest coffee shop (promotion) for instance.	Real time mass surveillance and privacy issues.

Table 1. Ethical and Societal Impacts of Big Science

Data Centres	Care.data	Storage and analysis of private personal medical records, suggested as being for the potential improvement of health care.	Privacy is a significant issue, as true anonymity cannot be guaranteed. Government control and exploitation of private, confidential data, especially due to the policy of patient opt out, not opt in. Risk of exploitation by third parties, such as insurance companies, if data are leaked or hacked.
Surveillance (Active)	Closed Circuit Television (CCTV)	CCTV when combined with Drones can now also be mobile. CCTV can be combined with face recognition software to enable movements of individuals to be monitored (tracked/traced).	Big Brother in every sense. The location and activities of the population continually observed, recorded and monitored.
Conventional Storage: Database, Data- warehouse, etc.	Google DeepMind	Often data already exists in the form of legacy systems/ data. The mining of such data applying knowledge-based systems (heuristic) or statistical models for genomics, etc.	Access to, and exploitation of, personal information and medical records without consent by commercial organisations.
Simulation (Causal reasoning)	Automatic Test Equipment (ATE)	ATE for computer hardware fault diagnosis has no obvious or immediate ethical issues because of the domain.	Simulation would perhaps only be an issue if the subject of a simulation were to lead to unethical decisions. E.g. a cost-benefit study related to simulation data where the study instigator chooses not to take an ethical course of action because a simulation shows that they "won't get caught".

Key Performance Indicators for Ethical Analytics and a Framework for Ethical Big Science

Legislation and Frameworks for Ethical Analytics

Legislation seldom, if ever, maintains pace with technology; analytics and big data are no exceptions. Laws use informed consent as the main method for protecting privacy on individual data. The Organization for Economic Cooperation and Development (OECD) has guidelines which are the basis for much of the privacy legislation. The *Guidelines on the Protection and Privacy and Transborder Flows of Personal Data* (OCED 1980, para. 7) collection limitation principle requires that personal information be collected "where appropriate, with the knowledge and consent of the data subject". The reuse of personal information is limited to the purposes originally specified at the time of the data collection, with exceptions only allowed through the informed consent of the individual. Presently, the main legislative protection in the UK is provided by the 1998 Data Protection Act (DPA, 1998). The impending (comes into effect in May 2018) European Union's General Data Protection Regulation (GDPR) provides a significant role for individual choice, referring to "consent" more than one hundred times. Important GDPR principles (Veitch, 2017) include:

- demands for clear consent from EU residents for the collection and use of their data;
- the scope of what constitutes personal data will include social media data, photos, e-mail addresses and computer IP addresses;
- the 'right to be forgotten' where an individual's data is deleted/erased on demand; and
- 'privacy by design' for processes and workflows.

Consent remains the primary mechanism for data protection throughout all current legislation.

Chessell (2014) describes an "ethical awareness framework", developed by the UK and Ireland Technical Consultancy Group (TCG) to aid in the development of ethical policies for analytics and big data. There are nine facets to this framework, namely; Context, Consent & Choice, Reasonable, Substantiated, Owned, Fair, Considered, Access and Accountable.

Data Benefits, Data Harm and the Epistemology of Knowledge

Much of this paper appears to be a criticism of big data and big data analytics. This inferred viewpoint is in part a counter balance to much of the available literature to-date which espouses only the benefits of big data analytics. The benefits of big data and analytics at the moment seem to be more hype than reality. Big science has the potential to offer a great deal, but so far there is no obvious paradigm shift from its application. There are smaller scaled, lower profile benefits of big science; pattern recognition has aided fraud detection and damage mechanics, analysis of surveillance data has let to the prevention of terrorism, etc. However, the benefits of big science regarding the prevention of terrorism in the society may conflict ethically with individual civil rights and freedoms.

Data Harm, broadly speaking refers to the dangerous or undesirable use of data, with the emphasis on the word "use". There are many perceived harms, for example,

the use of data in a way that might cause an individual injury or embarrassment, rather than the focus on notice and consent in existing data protection law (Cate 2016).

There are two perspectives on data; data as research objects and data as scientific methodology. Pale et. al. (Cate, 2016) argue that big data is a new approach to scientific inquiry in which data collection and mining alone (without theories) is a legitimate form of scientific enquiry. The data driven approach of big data analytics has led to the epistemology of knowledge itself being challenged, from theory based hypothesis and experiment driven, to data synthesis and mining.

Key Performance Indicators for Ethical Analytics

In Table 1, all of the societal impacts may come under *George Orwell's (1948) notion of a Big Brother society in his book entitled "Nineteen Eighty-Four".

The majority of the issues listed is negative and can lead to data harm, but some are contradictory in nature. For instance, the use of big data and analytics can be both positive and negative for security. It can be negative for security through the risks associated with the storage of the data in the cloud, or through data generation by the IoT, and consequently there are greater risks to data privacy and control. Equally, big data analytics and data mining can be used to identify patterns of behaviour (trend analysis) and so be used positively in the prevention of fraud or terrorism. A major problem remains in that computer and data science have evolved faster than the realisation of consequences or the legislation needed to address them.

From Table 1, the main data use issues or the Key Performance Indicators (KPIs) for the ethical uses of individual data are:

- Why
- What
- How
- When
- Data Harm
- Unforeseen (Future) Consequences

Ethics need to begin with the individual use, rather than consent, as reflected by newer legislation. The starting point should always be why the data are being used. If there is no justification (benefit) for the use of the data for the individual, then it should simply not be used. The exception is where the use benefits society and it is here where the explicit, informed consent of individuals must be obtained. The questions of: What are the data for; How will the data be used; When will the data be used; Data Harm (risk of injury or embarrassment); Unforeseen (Future) Consequences, must then be considered.

The KPIs for the ethical use of societal data, or individual data by or for the use of society are:

- Control
- Exploitation/Prediction
- Privacy
- Security
- Surveillance
- Cyber-Crime

In the case of the use of data created as a society (societal data), or individual data used by a society, the KPIs are wide reaching, so it is harder to decide on an order of precedence. Perhaps the ultimate use or misuse of data would be for control of the individual, organisation or state, by another individual, organisation or state. Such control could be for political or financial gain, or merely power. All other societal KPIs orbit around control. Ethical issues surrounding exploitation/prediction, privacy, security, surveillance and cyber-crime can emerge from big data (analytics) and through the use of smart devices.

A Framework for Ethical Big Science

Figure 1 depicts a Framework for Ethical Big Science. The inner hexagon relates to KPIs for the individual, the outer shows KPIs for society.

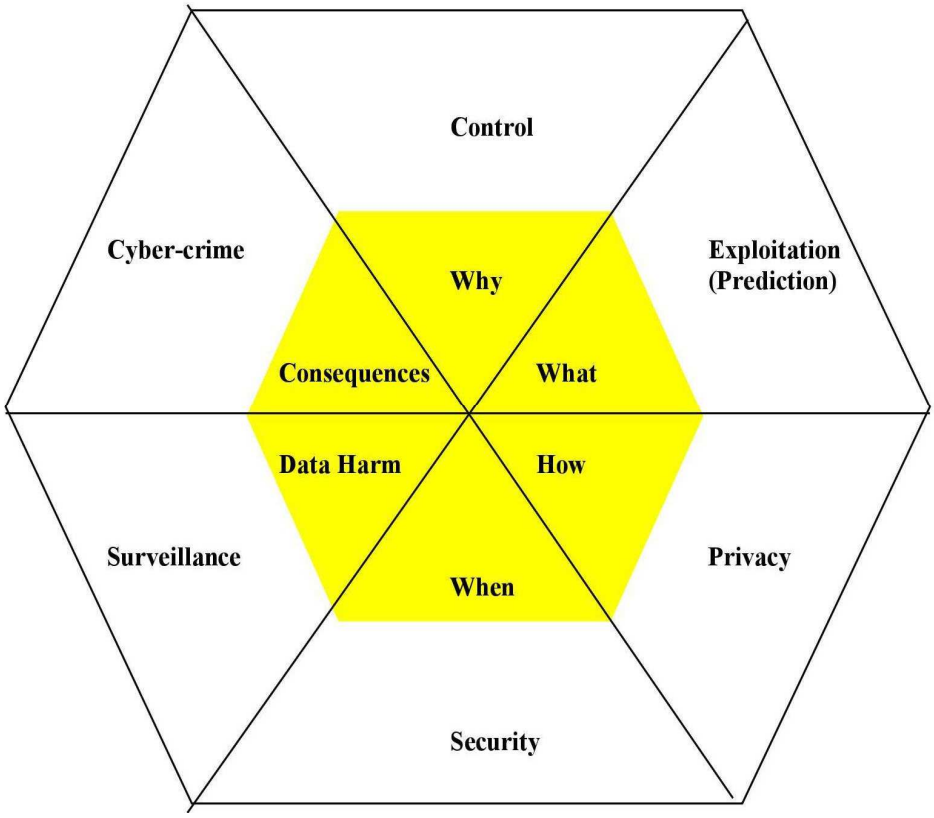

Figure 1. A Framework for Ethical Big Science

Discussion

To summarise; the emergence of the Big Science phenomenon is grounded in the convergence of technologies, such as Cloud Computing, the IoT, and the digitization of society and the resulting massive increase in (big) data generation. The issues affecting big science are pragmatic and ethical. Pragmatic issues relate to people, processes

and technologies. Ethical issues are fundamentally those of privacy and security. The opportunities for the exploitation of big data may be unlimited; however, unless the issues are met ethically, they are unlikely to be realised.

Big data itself is absolute, temporally and ethically; it just is. Ethics and the implications for individuals and society revolve around the collection, storage and use of big data and the application of analytics.

The essential nature of big data: volume, velocity, value, variety, veracity, variability and complexity, generated by a multitude of sources, immediately raises problems. What and how do you collect, store and manage such huge volumes of data at such velocities and of such variety, veracity, variability and complexity? The use of the Hybrid Cloud is both a part of the solution (the pragmatics) and the problem itself (the ethics). Quantum computing potentially offers another solution to the how, but it is still in its infancy in terms of any large-scale implementation. DNA storage may allow the collection, storage and management of big data. It was reported that "just 1 gram of DNA is theoretically capable of holding 455 exabytes – enough for all the data held by Google, Facebook and every other major tech company, with room to spare" (Aron, 2015). More recently (Shipman et. al., 2017), E.coli bacteria DNA has been used for the storage of horse images. The use of E.coli bacteria intuitively raises safety fears as a potential bio-hazard.

Big data analytics can be used positively or negatively (data harm); analytics can be used to identify an individual's shopping habits to enable up-selling. The same analysed data identifying purchase patterns can be used to determine identity theft and fraud.

The opportunities for the application of big data and analytics or "Big Science" are suggested by the hype to be endless. Current applications include medical applications, such as the Human Genome Project, with Genomics becoming increasingly important (Palmer, 2015). There is a temptation to use big data simply because it is there. A significant proportion of big data is likely to be spurious to any specific application or domain, and it may be unethical to use it. One domain source of big data has apparently been utilised successfully for another unrelated domain; the use of an earthquake aftershocks mathematical prediction model applied to crime prediction in Los Angeles (MIT, 2013). This had been suggested to be the possible identification of a natural generic pattern (akin to fractals) for seemingly disparate phenomena, or a unique feature of crime data models, a question that required further cases and research. However, the paradigm shifts envisaged for the application of big data analytics have yet to be seen.

Another vital consideration is the, as yet, unforeseen consequences of big data analytics. A proposed scenario (Graham, 2016) was that, one day in the near future, genetic data would be collected at birth and past big data analytics would enable the determination of all inherited medical conditions, intellect and other attributes, and even compute the likely date of death of an individual. This future was proven to be not so far fetched or simple paranoia, with the recent proposal by the UK Health Minister suggesting that all newborns be DNA tested (Smith, 2017). This would be deemed undesirable by many people. In the paper (Graham, 2016), the justification for the collection of genetic data was suggested to be security and privacy, as a person's DNA is unique and is, therefore, the ideal "password" for secure access to private information. However, in the case of the UK Health Minister's proposal, the justification was the supposed improved health care (genomics). There are likely to be suspicions about such altruism after the (Google) DeepMind fiasco, where private, confidential patient records were used for the development of commercial software (Streams). In

addition, rather conveniently, newborns are not capable of giving informed consent and their parents (and society) may not conceive any future data harm. There is a genuine, tangible fear that any database containing the nation's DNA data will be an infringement of privacy. This is the ultimate ethical dilemma for big science not equating to Big Brother, balancing the positive possibilities against the potential for data harm. Information is power. Once the genie is out of the bottle, there is no going back; you cannot "un-know" something, and the digitization of society ensures that such things are recorded in perpetuity.

Conclusions

A framework for ethical big science is required to address who, the what, the when, and the why of big science. There is an important temporal element with regard to these KPIs, as the ethics of collecting and storing data should be considered prior to analytics. Big science can relate to the past (data already in existence in legacy systems), the present (data generated now through new technologies) and the future (what data will be generated next, e.g. DNA profiling). For example, the suggestion by the UK Minister for Health, that people should be DNA tested from birth (Smith, 2017); this is future data with significant outcomes. Societal KPIs for big science are those of control, exploitation (prediction), privacy, security, surveillance and cyber-crime.

The challenges for big science are to solve the "how" (pragmatics), as well as to address the question of the "if" (ethics); the "if" should perhaps always be done on a case by case basis, and the temporal element has a bearing.

Big science has a political element which cannot be ignored. It is through the political system that most societies decide on the morals and ethics they are governed by, but the rate of change continues to provide a dilemma for society and for legislation. GDPR represents a step in the right direction for the reassertion of the data rights of individuals.

Curran (Sumner, 2013, p. 16) argues that "data centres will be the engine rooms driving the 'Fourth Industrial Revolution', which will see the internet of things, and big data transform the way modern businesses operate and societies function". The challenges to the epistemology of knowledge making scientific inquiry data driven, rather than hypothesis or theory led, could also have severe ramifications. The arguments against a data driven approach to scientific inquiry could be the lack of any substantial tangible successes. State-space search needs to be knowledge-based (hypothesis led or at least goal directed) as implied by the four stage PIPAE (Problem Identification, Preparation, Analysis, Evaluation) methodology (Graham, 2016) with the first stage being Problem Identification (PI). Blind mining of data may, therefore, not be a valid approach to achieving the potential offered by big data.

Big data, the IoT and robotics could lead to the automation of intellect and the loss of human ingenuity. Superior interface design offered by the IoT, intelligent automatons and smart devices, coupled with revolutionist scientific inquiry methods, could allow humanity to sleep walk into its own intellectual demise. ML must be an addition to, not a replacement for, methods of scientific inquiry. Society is not threatened by big science or Artificial Intelligence (AI, which doesn't really exist, no truly independent artificial or sentient intelligence), but is threatened by the ill-considered automation and digitization of every aspect of human life.

Ultimately, how ethical issues are addressed depends on what sort of society we want to live in. The argument for big science is often for the betterment of humanity. However, is a society where all is known, without individual privacy or freedom of

action or movement, better or truly desirable? The data that already exists exists, there's no reverse gear, but the consequences for individuals and society must be considered now, as we are presently hurtling towards an unknown or ironically, a predictable future.

Correspondence
Eur. Ing. Dr. Deryn Graham
Data Analytica, UK
Email: D.Graham@mlanalytica.com

Acknowledgments
Care.data, Google DeepMind and Streams, Facebook, British Gas Hive and Microsoft OneDrive are registered trade names.

This paper is based on and extended from the KIE conference paper: Graham D. (2017). The Ethical and Societal Impacts of Big Science: A Framework for Ethical Big Science. *Research Papers on Knowledge, Innovation and Enterprise, Vol. V 2017, KIE Conference Book Series,* J. Ogunleye (Ed.), 2017 International Conference on Knowledge, Innovation and Enterprise, pp. xx-yy.

References

Aron J. (2015). NewScientist magazine, Issue number 3008, 15th February 2015.
Bloom B. (1956). The taxonomy of educational objectives: the classification of educational goals, handbook one: the cognitive domain. New York: Mc Kay, 1956.

BritishGas, "How it works – Hive Active Heating – BritishGas". Available from: http://www.britishgas.co.uk/products-and-services/hive-active-heating/how-it-works. Accessed on: 21 May 2015.

Burton G. (2017), "Transfer of 1.6 million medical records to Google was legally flawed Royal Free Hospital warned". Available from: http://www.computing.co.uk/ctg/news/301068/transfer-of-16-million-medical-records. Accessed on: 18 May 2017.

Cate F.H. (2016). "Big Data, Consent, and the Future of Data Protection", in C. R. Sugimoto, H. R. Ekbia, M. Mattioli (Editors), 2016. Big Data is not a Monolith, Chapter 1, Massachusetts Institute of Technology ,2016, pp. 3-20.

Chang V., Wills G. (2013). "A University of Greenwich Case Study in Cloud Computing", in D. Graham, I. Manikas and D. Folinas (Editors), 2013. E-Logistics and E-Supply Chain Management: Applications for Evolving Business. Chapter 13, IGI Global Publishers, 2013, pp. 232-231.

Chessell M. (2014). Ethics for big data and anlytics. IBM corporation 2014. Available from: http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD%26A.pdf. Accessed on: 18 March 2016.

Clover ETL (2017). Addressing Data Anonymitization Challenges, White Paper. Available from: www.cloveretl.com. Accessed on: 18 May 2017.

Data Protection Act (DPA). 1998. Available from: https://www.gov.uk/data-protection/the-data-protection-act. Accessed on: 7 August 2017.

de Lama N. (2016). BDV Big Data Value. Available from: https://ec.europa.eu/digital-single-market/en/news/information-and-networking-day-h2020-ict-15-big-data-ppp-lighthouse-projects. Accessed on: 24 March 2016, p. 10.
Gordon K. (2013). "What is Big Data?", ITNOW September 2013, 2013, pp. 12-13.

Graham D. (2016). Big Data Science Education and the PIPEA Methodology for Big Data Analytics. Research Papers on Knowledge, Innovation and Enterprise, Vol. IV 2016, KIE Conference Book Series, J. Ogunleye (Ed.), 2016 International Conference on Knowledge, Innovation and Enterprise, pp. 144-159.

Graham D. (2015). "Big Data and Established Models of Knowledge", International Journal of Developments in Big Data and Analytics, KIE Publications, vol. 1, no. 1 (December 2014), 2015, pp. 6-17.

Graham D. (2014). "The temporal impact and implications of E-Learning". Accepted for Special Issue: The Temporal Dimensions of E-Learning, in Journal of E-Learning and Digital Media (ELEA), vol. 11, no. 4, October 2014, 2014, pp. 323-332.

Graham D. (2013). "Introduction to E-Logistics and E-Supply Chain Management", in D. Graham, I. Manikas and D. Folinas (Editors), 2013. E-Logistics and E-Supply Chain Management: Applications for Evolving Business. Chapter 1, IGI Global Publishers, 2013, pp. 1-8.

Heger D. A. (2014). Big Data & Predictive Analytics: Applications, Algorithms and Cluster Systems. DHTechnologies, 2014.

Marr B. (2015). Big data: using smart big data, analytics and metrics to make better decisions and improve performance. UK: John Wiley & Sons Ltd.

McKinsey Global Institute (2015) "Big Data: The next frontier for innovation, competition and productivity", 2011, in Neaga I., Hao Y. (2013).

MIT Technology Review (2013). Available from: www.technologyreview.com/news/428354/la-cops-embrace-crime-predicting-algorithm. Accessed on: 6 February 2014.

Neaga I., Hao Y. (2013). "Towards Big Data Mining and Discovery", Short Research Papers on Knowledge, Innovation and Enterprise, Part 2 - Innovation, KIE Conference Book Series, J. Ogunleye, D. Heger and U. H. Richter (Eds), 2013 International Conference on Knowledge, Innovation & Enterprise, 2013, pp. 35-43.

Organization for Economic Cooperation and Development (OECD). 1980. Guidelines on the Protection of Privacy and Transborder Flaws of Personal Data (C58 Final), in Cate F.H. (2016).

Orwell G. (1948). Nineteen Eighty-four.

Palmer D. (2015a). "Is 2015 the year IoT comes of age?", in Computing, March 2015, pp. 14-15.

Palmer D. (2015) "Is big data the key to a healthier NHS?", in Computing, February 2015, pp. 8 -9.

SAS (2012). Big Data Meets Big Data Analytics, SAS White Paper. Available from: http:// www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/big-data-meets-big-data-analytics-105777.pdf. Accessed on: 18 March 2016.

Shipman S.L., Nirala, J., Macklis J. D., Church G. M. (2015). CRISPR – case encoding of a digital movie into the genones of a population of living bacteria. Nature International Weekly Journal of Science (online). Nature 547, 345-349, 20 July 2017.

Smith C. (2013). DNA testing at birth, the UK strides ahead. Available from: https:// dnatestingchoice.com/news/2013-12-09-dna-testing-at-birth-the-uk-strides-ahead. Accessed on: 3 August 2017.

Sumner S. (2013). "Data Centre Summit: insight and foresight", in Computing, 3 October 2013, pp. 14-16.

Veitch M. (2017). Why GDPR is important and what you need to do to get to grips with it. BlackBerry, CIO Strategic Marketing Services.