Nicolae Sfetcu

Big Data Ethics in Research

ESSAYS Collection

MultiMedia Publishing

Big Data Ethics in Research

Nicolae Sfetcu

15.09.2019

Sfetcu, Nicolae, " Big Data Ethics in Research ", MultiMedia (September 15, 2019), MultiMedia Publishing (ISBN 978-606-033-307-4) , DOI: 10.58679/MM14975 ,
URL = https://www.telework.ro/en/e-books/big-data-ethics-in-research/

Email: nicolae@sfetcu.com

Abstract

The main problems faced by scientists in working with Big Data sets, highlighting the main ethical issues, taking into account the legislation of the European Union. After a brief *Introduction* to Big Data, the *Technology* section presents specific research applications. There is an approach to the main philosophical issues in *Philosophical Aspects*, and *Legal Aspects* with specific ethical issues in the EU Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (Data Protection Directive - General Data Protection Regulation, "GDPR"). The *Ethics Issues* section details the specific aspects of Big Data. After a brief section of *Big Data Research*, I finalize my work with the presentation of *Conclusions* on research ethics in working with Big Data.

1. Introduction

The term Big Data refers to the extraction, manipulation and analysis of data sets that are too large to be routinely processed. Because of this, special software is used and, in many cases, also dedicated computers and hardware. Generally, for these data the analysis is done statistically. Based on the analysis of the respective data, predictions of certain groups of people or other entities are usually made, based on their behavior in various situations and using advanced analytical techniques. Thus, tendencies, needs and behavioral evolutions of these entities can be identified. Scientists use this data for research in meteorology, genomics, (Nature 2008) connectomics, complex physical simulations, biology, environmental protection, etc. (Reichman, Jones, and Schildhauer 2011)

With the increasing volume of data on the Internet, in social media, cloud computing, mobile devices and government data, Big Data is both a threat and an opportunity for researchers to manage and use this data while maintaining the rights of the involved people.

1.1 Definitions

Big Data usually includes sets of data that exceed the capacity of ordinary software and hardware, using unstructured, semi-structured and structured data, with an emphasis on unstructured data. (Dedić and Stanier 2017) Big Data has grown in size since 2012, from dozens of terabytes to many data exabytes. (Everts 2016) Making data efficient with Big Data involves machine learning to detect patterns, (Mayer-Schönberger and Cukier 2014) but often this data is a by-product of other digital activities.

A 2018 definition states that "Big data is where parallel computing tools are needed to handle data," which represents a turning point in computing, using parallel programming theories and the lack of assurances assumed by previous models. Big Data uses inductive statistics and concepts of identifying nonlinear systems to deduce laws (regressions, nonlinear relationships and causal effects) from large data sets with low information density to obtain relationships and dependencies or to make predictions of results and behaviors. (Billings 2013)

At European Union level there is no mandatory definition but, according to the Opinion 3/2013 of the European Working Group on data protection,

> "Big Data is a term that refers to the enormous increase in access to and automated use of information: It refers to the gigantic amounts of digital data controlled by companies, authorities and other large organizations which are subjected to extensive analysis based on the use of algorithms. Big Data may be used to identify general trends and correlations, but it can also be used such that it affects individuals directly." (European Economic and Social Committee 2017)

The problem with this definition is that it does not consider reusing personal data.

Regulation no. 2016/679 defines personal data (Article 4, paragraph 1) as

> "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person." (European Economic and Social Committee 2017)

The definition applies at EU level also to unidentified persons, but which can be identified by correlating anonymous data with other additional information. Personal data, once anonymized (or pseudo-anonymized), can be processed without the need for authorization, however, taking into account the risk of re-identifying the data subject.

1.2 Big Data dimensions

The data is shared and stored on servers, through the interaction between the entity involved and the storage system. In this context, Big Data can be classified into active systems (synchronous interaction, entity data is sent directly to the storage system), and passive systems (asynchronous interaction, data is collected through an intermediary and then entered into the system).

Also, the data can be transmitted directly, consciously or non-consciously (if the person whose data is transmitted is not notified on time and clearly). The data is then processed to generate statistics.

Depending on the target of the respective statistics analyzes, the data dimensions may be a) individual (only one entity is analyzed); social (there are analyzed discrete groups of entities within a population); and hybrids (when an entity is analyzed from the perspective of its belonging to an already defined group).

The current huge output of user-generated data is expected to grow by 2000% worldwide by 2020 and are often unstructured. (European Economic and Social Committee 2017) In general, Big Data is characterized by:

- Volume (amount of data);
- Variety (products from different sources in different formats);
- Speed (speed of online data analysis);
- Accuracy (data is uncertain and must be verified);
- Value (evaluated by analysis).

The volume of data produced and stored is currently evolving exponentially, over 90% of them being generated in the last four years. (European Economic and Social Committee 2017) Large volumes require high speed of analysis, with a strong impact on veracity. Incorrect data has the potential to cause problems when used in the decision process.

One of the major problems with Big Data is whether the complete data is needed to draw certain conclusions about their properties, or a sample is enough. Big Data contains in its name a term related to size, which is an important feature of Big Data. But (statistical) sampling allows the selection of correct data collection points from a larger set to estimate the characteristics of the entire population. Big Data can be sampled across different categories of data in the process of sample selection with the help of sampling algorithms for Big Data.

2. Technology

Data must be processed with advanced collection and analysis tools, based on predetermined algorithms, in order to obtain relevant information. Algorithms must also take into account invisible aspects for direct perceptions.

In 2004 Google published a paper about a process called MapReduce that offers a parallel processing model. (Dean and Ghemawat 2004) MIKE2.0 is also an open source application for information management. (MIKE2.0 2019) Several studies from 2012 have shown that the optimal architecture for addressing Big Data issues is multi-layered. A distributed parallel architecture distributes data on multiple servers (parallel execution environments) thus dramatically improving data processing speeds.

According to a report from the McKinsey Global Institute in 2011, the main components and ecosystems of Big Data are: (Manyika et al. 2011) data analysis techniques (machine learning, natural language processing, etc.), big data technologies (business intelligence, cloud computing, databases), and visualization (charts, graphs, other data views).

Big Data provides real-time or near real-time information, thus avoiding latency whenever possible.

2.1 Applications

Big data in government processes increases cost efficiency, productivity and innovation. Civil records are a source for Big Data. The processed data helps in critical areas of development, such as health care, employment, economic productivity, crime, security and management of natural disasters and resources. (Kvochko 2012)

Also, Big Data provides an infrastructure that allows for highlighting uncertainties, performance, and availability of components. Trends and predictions in the industry require a large amount of data and advanced prediction tools.

Big Data contributes to the improvement of healthcare by providing personalized medicines and prescriptive analyzes, clinical interventions with risk assessment and predictive analysis, etc. The level of data generated in health systems is very high. But there is a pressing problem with generating "dirty data", which increase with increasing volume of data, especially since most are unstructured and difficult to use. The use of Big Data in healthcare has generated significant ethical challenges, with implications on individual rights, privacy and autonomy, transparency and trust.

In the field of health insurance, data is collected on the "determinants of health", which helps to develop forecasts on health costs and to identify clients' health problems. This use is controversial, due to the discrimination of clients with health problems. (Allen 2018)

In the media and advertising, for Big Data, numerous information points are used about millions of people, to serve or transmit personalized messages or content.

In sports, Big Data can help improve competitors' training and understanding using specific sensors and predict future performance of athletes. Sensors attached to Formula 1 cars collect, inter alia, tire pressure data to make fuel burning more efficient.

Big data and information technology complement each other, helping together to develop the Internet of Things (IoT) for interconnecting smart devices and collecting sensory data used in different fields.

2.1.1 In research

In science, Big Data systems are used extensively in particle accelerators at CERN (150 million sensors transmit data 40 million times per second, for about 600 million collisions per second, of which they are used after filtering only 0.001% of the total data obtained), (Brumfiel 2011) in astrophysical radio telescopes built from thousands of antennas, decoding the human genome (initially it took a few years, with Big Data can be done in less than a day), climate studies, etc. .

Big IT companies use data warehouses of the order of tens of petabytes for search, recommendations and merchandising. Most data is collected by Facebook, with over 2 billion monthly active users (Constine 2017) and Google with over 100 billion searches per month. (Sullivan 2015)

The research uses a lot of encrypted search and cluster formation in Big Data. Developed countries are currently investing heavily in Big Data research. Within the European Union, these researches are included in the Horizon 2020 program. (European Commission 2019)

Often, research programs use API resources from Google and Twitter to gain access to their Big Data systems, for free or at no cost.

Large data sets come with algorithmic challenges that previously did not exist, and it is imperative to fundamentally change the processing methods. To this end, special workshops have been created that bring together scientists, statisticians, mathematicians and practitioners to discuss the algorithmic challenges of Big Data.

3. Philosophical aspects

Big Data can generate, through inferences, new knowledge and perspectives. The paradigm that results from using Big Data creates new opportunities.

One of the major concerns in the Big Data case is that data scientists tend to work with data on topics they do not know and have never been in contact with, being alienated from the final product of their activity (application of analyzes). A recent study (Tanner 2014) states that this may be the reason for a phenomenon known as digital alienation.

Big Data has great influence at the governmental level, positively affecting society. These systems can be made more efficient by applying transparency and open governance policies, such as Open Data.

After developing predictive models for target audience behavior, Big Data can be used to generate early warnings for various situations. There is thus a positive feedback between research and practice, with rapid discoveries taken from practice.

A. Richterich, in "Examining (Big) Data Practices and Ethics", states that the popularization of user activity monitoring was motivated by claims that using (and collecting data with) these devices would improve users' well-being, health and life expectancy, and significantly reduce healthcare costs. (Richterich, 2018) To obtain user consent, many companies offered discounts to those customers who would be willing to provide access to their monitoring data.(Mearian 2015) But there are also concerns about the influence of these technologies on society, especially in issues related to fairness, discrimination, privacy, data abuse and security. (Collins 2016)

Conceptually, Big Data should be understood as an umbrella term for a set of emerging technologies. In their use, we must take into account the cultural, social and technological contexts, networks, infrastructures and interdependencies that may make sense on Big Data. The term "Big Data" refers not only to the data as such, but also to the practices, infrastructures, networks and policies that influence their various manifestations. Understanding big data as a set of emerging technologies seems to be conceptually useful, as it "encompasses digitally enabled developments in data collection, analysis, and utilization." (Richterich, 2018)

In this context, Rip describes the dilemma of technological developments: "For emerging technologies with their indeterminate future, there is the challenge of articulating appropriate values and rules that will carry weight. This happens through the articulation of promises and visions about new technosciences." (Rip 2013, 192) Thus, emerging technologies are places of "pervasive normativity" characterized by articulating promises and fears, conceptualizing it as an approach "in the spirit of pragmatist ethics, where normative positions co-evolve" (Rip 2013, 205)

Pragmatic ethics emphasizes that new technologies are developing in societies in which they are discursively associated/dissociated by certain norms and values. At the same time, pragmatism states that increasing the large number of data and research-related practices is not a simple matter of technological superiority. They form a field of normative justification and contestation.

The *neo-pragmatic* approach to ethics addresses epistemological knowledge through the falsification of (scientific) knowledge, with critical evaluations of social power structures. Keulartz et al. have proposed a pragmatic approach to ethics in a technological culture (Keulartz et al. 2004) "as alternative which combines the strengths of applied ethics and science and technology studies, while avoiding the weaknesses of these fields." (Richterich, 2018) Thus, applied ethics is an effective approach in terms of detecting and expressing the norms involved in (inter-) socio-technical actions or resulting from socio-technical actions, but it has no possibilities to capture the inherent normativity and the agent of technologies. (Keulartz et al. 2004, 5)

Keulartz et al. believes that the lack of normative technological evaluations can thus be overcome: "'impasse that has arisen from this" (i.e. the respective 'blind spots' of applied ethics and STS) "can be overcome by a re-evaluation of pragmatism." (Keulartz et al. 2004, 14) Ethical pragmatism can be characterized by three common principles: anti-foundationalism, anti-dualism and anti-scepticism.

Anti-foundationalism refers to the principle of falsifiability, considering that we cannot reach certainty in terms of knowledge or values ("ultimate truth"), but knowledge, as well as values and norms, changes over time. Moral values are not static but can be renegotiated depending on technological developments.

Anti-dualism implies the need to refrain from predetermined dichotomies. Among the dualisms criticized by Keulartz are the essence/appearance, theory/practice, consciousness/reality and facts/value. Applied ethics tends to assume such dualisms as *a priori*, as opposed to pragmatism, which underlines the blurred interrelations and lines between such categories.

Anti-scepticism is closely linked to the need for situated perspectives and explicit normativity, relating to the anti-Cartesian foundation of pragmatism.

In European research, pragmatism was usually dismissed as superficial and opportunistic, being associated with negative stereotypes, (Joas 1993) being accused of "utilitarianism and meliorism." (Keulartz et al. 2004, 15) At the end of the 1990s and 2000s, pragmatism experienced a revival in European research. (Baert and Turner 2004)

European Economic and Social Committee, in "Big Data: Balancing economic benefits and ethical questions of Big Data in the EU policy context", states that Big Data analysis from an ethical point of view involves two main interdependent aspects: a theoretical one (the philosophical description of the elements subject to ethical control) and a pragmatic vision (of the impact on the lives of people and organizations). (European Economic and Social Committee 2017)

There are ethical problems caused by artificial intelligence, and a close link between Big Data and artificial intelligence and its derivatives: machine learning, semantic analysis, data exploitation.

An ethical approach is through the moral agency with at least the three conditions of causation, knowledge and choice. According to Noorman: (Noorman 2012)

- There are causal links between people and the outcome of actions. The person's responsibility derives from the control over the result.
- The subject should be informed, including on possible consequences.
- The subject must give his consent and act in a certain way.

Professor Floridi, in *The Fourth Revolution*, identifies the moral problem of Big Data with the discovery of a simple model: a new frontier of innovation and competition. (Floridi 2014) Another problem associated with Big Data is the risk of discovering these patterns, thus changing the predictions.

The basic rule of Big Data ethics is the protection of privacy, freedom and discretion to decide autonomously. It is worth noting that there is a continuous tension between the individual needs and those of a community.

It is possible to identify several ethical issues arising from the exploitation of Big Data: (European Economic and Social Committee 2017)

- *Privacy* - The extreme limit of confidentiality is the seclusion, defined by Alan F. Westin as "the voluntary withdrawal of a person from the general society through physical [means] in a state of solitude". Moor and Tavani defined a privacy model called Restricted Access Control (RALC) that differentiates between privacy, justification, and privacy management.
- *Tailored reality and the filter bubble* - The application on a server collects information by learning from it, and then uses that information to build a model of our interests. When a system uses these models to filter information, we may be induced to believe that what we see is a complete view of a specific context, when in fact we are limited by the "understanding" of an algorithm that built the model. The ethical effects can be multiple: some information can be hidden, imposing prejudices which we do not know, our vision of the world can become progressively limited, and in the long term could generate a certain point of view.
- *After death data management* - What happens to the data of a deceased user? Do the heirs become their owners? Can data be removed from the digital world? There are legal and technological problems here.
- *Algorithm bias* - Data interpretation almost always involves certain biases. In addition, there is a possibility that an error in an algorithm may introduce bias forms. An ethical issue is our

implicit trust in algorithms, with high risks when risks are not taken into account due to programming or running errors of the algorithms.

- *Privacy vs. growing analysis power* - It refers to the emergent nature of information as a complex system: the result of data from different contexts is more than the simple sum of the parts.
- *Purpose limitation* - It is very difficult or even impossible to limit the use of data. Privacy is not a single block, with subtle forms of privacy being lost.
- *User digital profile inertia and conformism* - This is about the subject of personalized reality. A model that involves a user's interests is usually based on past behavior and past information. Thus, the algorithms are not based on the actual identity of the person, but on an earlier version. This will influence the real behavior of the user, being pushed to maintain their old interests and therefore not be able to discover other opportunities. If the user is not aware of this problem, the influence of inertia will be much greater.
- *User radicalization and sectarism* - Big Data can form opinions using filtering/recommendation algorithms, information, personalized articles and posts, and specific recommendations from friends. Thus, users will be more and more in touch with the people, opinions and facts that will support their original position. This tendency is often hidden from the users of Big Data based systems, with the tendency to develop prejudices, ranging from conformity to radicalization. It is possible to postulate the formation of a kind of technological subconscious with impact on the development of the personality of the users, phenomena evident in the case of social networks, where the distance between the real ("physical") world and the Internet is strongly attenuated.
- *Impact on personal capabilities and freedom*

- *Equal rights between data owner and data exploiter* - Usually the person whose data is used is not their legal owner. Therefore, a minimum requirement is for that person to have access to their own data, allowing them to download them and eventually delete them.

4. Legal aspects

The use of Big Data presents significant legal problems, especially in terms of data protection. The existing legal framework of the European Union based in particular on the Directive no. 46/95/EC and the General Regulation on the Protection of Personal Data provide adequate protection. But for Big Data, a comprehensive and global strategy is needed. The evolution over time was from the right to exclude others to the right to control their own data and, at present, to the rethinking of the right to (digital) identity.

The collection and aggregation of data in Big Data are not subject to data protection regulations, due to new perspectives on confidentiality, with the possibility of specific forms of discrimination.

In 2014, Podesta's report concluded that "big data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education, and the marketplace." (European Economic and Social Committee 2017) It follows that new specific ways of protecting citizens are needed, because the legal framework, although theoretically applicable, does not seem to provide adequate and full protection.

4.1 GDPR

The General Data Protection Regulation, "GDPR" (Regulation EU 2016/679) deals with data protection and privacy of persons in the European Union and the European Economic Area. It specifically addresses the export of personal data outside EU and EEA areas. The GDPR intends to simplify the regulatory environment by unifying the regulation within the EU. (European Parliament 2016)

GDPR applies in two cases for the processing of personal data: (a) access to goods or services for a fee by persons in the EU, or (b) monitoring their behavior within the EU. Thus, the regulation allows it to be extended to all Internet service providers, even if they are not established in the EU. More generally, GDPR applies to all large data aggregators, regardless of geographical or physical connections.

Stages of processing of personal data

The processing of personal data is defined in Article 4, paragraph 2, as "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."

Big Data includes several personal data processing activities, each with its own specific rules:

1. data collection
2. data storage
3. data aggregation
4. data analysis and use of analysis results

Principles of data processing

Data processing is based on the following principles set out in Article 5 of the GDPR:

1. Legality, fairness and transparency: Users must be fully and properly informed regarding the privacy policy and be able to easily access their own data.

2. Purpose limitation: Data collectors must inform the data subject about the purposes of data collection, which can be further processed for those purposes only.
3. Data minimization: Only personal data relevant to the stated purposes will be collected.
4. Accuracy and updating: The data will be updated and rectified whenever required by the stated purpose. In the case of Big Data, the right of users to cancel or delete personal data is very important.
5. Limitation of storage: Data will be stored only during processing and subsequently destroyed. The duration of storage may be extended to the extent that the data are archived for public interest, scientific or historical research or statistical purposes.
6. Integrity and confidentiality: the data operator: Ensure adequate security for personal data through technical and organizational measures.

Privacy policy and transparency

In the case of data collection in order to complete a form, the principle of data minimization will be respected, only the relevant and strictly necessary data being requested. In the case of automatic data collection, such as cookies, web monitoring or geolocation, the privacy policy must inform the user about this aspect.

Purposes of data processing

Anonymous and aggregate data can be processed to identify the behavior of certain categories of consumers. For this purpose, the data operator performs anonymization and then transfers them to a third party using them.

Design and implicit confidentiality

The concepts of privacy by design and implicit confidentiality were not explicitly included in EU regulations. But, according to art. 78 of the GDPR,

> "In order to be able to demonstrate compliance with this Regulation, the controller should adopt internal policies and implement measures which meet in particular the principles of data protection by design and data protection by default. Such measures could consist, inter alia, of minimizing the processing of personal data, pseudo-anonymizing personal data as soon as possible, transparency with regard to the functions and processing of personal data, enabling the data subject to monitor the data processing, enabling the controller to create and improve security features. When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations."

The (legal) paradox of Big Data

The use of Big Data implies at least one paradox: on the one hand, Big Data ensures maximum transparency but at the same time, there is no adequate transparency regarding the use of Big Data. Transparency is a fundamental issue because it influences the ability of a user to allow the disclosure of his information.

5. Ethical issues

Big Data ethics involves adherence to the concepts of right and wrong behavior regarding data, especially personal data. Big Data ethics focuses on structured or unstructured data collectors and disseminators.

Big Data ethics is supported, at EU level, by extensive documentation, which seeks to find concrete solutions to maximize the value of Big Data without sacrificing fundamental human rights. The European Data Protection Supervisor (EDPS) supports the right to privacy and the right to the protection of personal data in the respect of human dignity. According to these documents, the conceptual conflict between privacy and Big Data, and between intimacy and innovation, must be overcome. It is essential to identify the ways of including the ethical dimension in the development of innovations. (European Economic and Social Committee 2017)

According to the new EU Regulation 2016/679, data operators must implement the confidentiality measures and technologies to improve the confidentiality when determining the processing modalities and the processing itself. Through ENISA75 many privacy strategies have been identified by design (data minimization, hiding personal data and their interconnections, separate processing of personal data, choosing the highest level of aggregation, transparency, monitoring, privacy policy, legal issues).

A basic way for peaceful coexistence between Big Data exploitation and data protection is user *control* of personal data, which leads to transparency and trust between users and digital service providers. As outlined in the GDPR impact assessment,

> "Building trust in the online environment is key to economic development. Lack of trust makes consumers hesitate to buy online and adopt new services, including public e-government services. If not addressed, this lack of confidence will continue to slow down the development of innovative uses of new technologies, to act as an obstacle to economic growth and to block the public sector from reaping the potential benefits of digitization of its services." (European Data Protection Supervisor, Opinion 7/2015 *Meeting the challenges of Big Data A call for transparency, user control, data protection by design and accountability*.)

In the case of Big Data, traditional *consent* models are insufficient and outdated. The "consent should be granular enough to cover all the different processing and purposes of processing and reuse of personal data." (European Economic and Social Committee 2017)

A special problem is data *portability*, supported at EU level by the EDPS in Opinion 7/2015, (MORO 2016) where it is necessary to guarantee the right of citizens to access and correct personal data through an expanded control. Data portability can help increase consumer awareness and control by transferring online services.

The EDPS considers that personal data should be treated just like other important resources, such as oil, where the trading takes place between equally well-informed parties (informational symmetry). In fact, the market for personal information has a character of informational asymmetry, being neither transparent nor fair, customers are not compensated for the personal information they provide. Thus, the portability of the data would encourage a more competitive environment among the beneficiaries of this data, the users having the possibility to choose who offers the personal data.

Another approach involves the *storage* of personal data, with the possibility for the user to grant or withdraw consent for his personal data. (MORO 2016) (DG Connect 2015) The storage of personal data involves a "concept, framework, and architectural implementation that shifts data acquisition and control from a distributed data model to a *user-centric model*." (European Economic and Social Committee 2017) Data portability could ensure this.

The EDPS supports promoting responsible beneficiaries and reducing bureaucracy in data protection, through codes of conduct, audits, certifications, and a new generation of contractual clauses and mandatory corporate rules. The responsibility of Big Data beneficiaries involves the establishment of internal policies and control systems in accordance with the legislation in force, through intelligent and dynamic solutions that guarantee the respect of fundamental principles (data minimization, purpose limitation, data quality, correct and transparent data processing, design, storage limitation, integrity and confidentiality).

Data ethics is based on the following principles: *ownership* (individuals own their data), *transparency* of transactions (users must have transparent access to the algorithm design), *consent* (the user must be informed and expressly consent to the use of personal data), *privacy* (user privacy must be protected), *financial* (the user should know the financial transactions resulting from the use of his personal data), and *openness* (aggregated data sets should be freely available).

Ethics in research

The term critical data studies (CDS) implies that researchers are investigating Big Data from critical perspectives. The study of data in this context involves, in addition to their analysis, the incorporation of data into practices (knowledge), political and economic institutions and systems, through the complex interaction between data and the entities that produce, own and use them.

An OECD report (2013) underlines that, unlike the ethical norms applied to common research data, in the case of Big Data: (OECD 2013)

- Data collection was not subject to a formal ethical review process.
- Common ethical rules will not be implemented in the case of Big Data
- The use of research data may differ from the initial purpose.
- Data is no longer held as discrete sets.

The relationship between those who provide the data and those who use it is often indirect and variable. A more recent OECD report (2016) argues that this relationship is weaker or non-existent, with Big Data limiting common capabilities. (OECD 2016)

Data storage is important for research integrity. The data must have a clear provenance, with known, identified and documented sources and processing.

Many data that are not specifically collected for research have different standards in data research.

For some data, often of commercial value (e.g., data collected on Twitter), there are legal restrictions on their reproduction. (UK Data Service 2017)

Data storage must comply with standards of transparency and reproducibility.

Awareness

Awareness of the type of data that is provided during an online registration (for creating an account, or a subscription, for example) is a rare fact, especially since there is the possibility of using an existing digital identity (Facebook profile, for example) instead of a separate registration for faster access. Such situations create an opacity regarding the data shared between the identity provider and the service used.

Consent

In order to use the personal data of a person, his or her informed and explicit consent is required regarding who, when, how and for what purpose they are used. When data needs to be shared, these uses must be made known to the person. It should always be possible to withdraw consent for future use.

In Big Data analytics, very little can be known about the intended future uses of data, and about the benefits and involved risks. Here, there are procedures for "broad" and "generic" consent to share genomic data, for example, and for different purposes. Even when done correctly, there are some specific practical challenges: obtaining informed consent can be impossible or very costly, and the validity of consent is disputed when the agreement is required to access a service.

Control

In today's world, personal data can be traded just like any currency in Big Data implementation. There are different opinions to what extent this situation is ethical, including who to participate in the profit obtained from these transactions.

In the trading model of personal data, the transmission of personal data is a framework that offers people the opportunity to control their digital identity and create granular agreements of data sharing.

The idea of open data, centered around the argument that data should be freely available, is now emerging. Willingness to share data varies by person.

In the case of children, parents or tutors have responsibility for their data, which cannot be traded for financial benefits.

At national level, a government is sovereign over the generated and collected data. On October 26, 2001, the Patriotic Act entered into force in the US, and on May 25, 2018, the General Data

Protection Regulation 2016/679 (GDPR) at the European Union level, for the issues related to the protection of personal data.

In Big Data, the human-data relationship is asymmetrical, based on data control. The "right to be forgotten", adopted at EU level, is one of the basic elements of an individual's control over his personal data.

Transparency

Anticipatory governance involves Big Data-based predictive analytics to evaluate potential behaviors, with ethical implications that can encourage prejudice and discrimination.

A person who accepts the inclusion of his personal data in Big Data has the right to know why the data is collected, how it will be used, how long it will be stored, and how it can be modified.

Trust

Confidence in Big Data systems is linked to interdependence with confidentiality and awareness. So far, trust has been considered from a strictly technological perspective. It is hoped that hardware and software architectures will be developed that could increase trust between human beings and objects, and thus a greater acceptance of the use of personal data.

Ownership

A fundamental question in the ethics of Big Data research is, who owns the data? This involves the subject of property rights and obligations. In European law, the GDPR indicates that people have own their own personal data.

The sum of an individual's personal data forms a digital identity.

The protection of the moral rights (the right to be identified as a source of data, and to control them) of an individual is based on the opinion that personal data are a direct expression of his personality, and can only be transferred to another person, possibly, by succession when the individual dies.

The property implies exclusivity, i.e. the implicit restriction of others regarding access to the property. An efficient ownership of personal data involves portability, the ability to use alternatives without losing data. Standardization would also help to clean up your personal data.

At present, the data is owned by the owner of the sensors, the one who makes the recording or the entity that owns the sensor.

In the EU, the possibility of EU citizens' data being stored outside the so-called "Euro cloud" has been progressively reduced, but the problem of data already stored and processed elsewhere has not been resolved, and "does not resolve the ethical dilemma of how data ownership is defined philosophically, before passing to a more down-to-earth approach of law and policy making." (European Economic and Social Committee 2017)

Surveillance and security

More and more data sources are available with the help of advanced technologies such as CCTV, GPS, mobile devices, credit cards, ATMs. Also, active surveillance is a method of collecting data, but at the same time limiting the freedoms of citizens. Such permanent surveillance determines the increase of people's stress and creates their tendency to behave in a certain way that conforms to the expected norms.

Digital identity

Digital identity has the advantage of quick access to online content and related services. The use of digital identity has the potential to generate discrimination based on the representation of a person according to their online data, which may often not correspond to the real situation, in a process called "data dictatorship" in which "we are no longer judged on the basis of our actions, but on the basis of what all the data about us indicates our probable actions may be", (Norwegian Data Protection Authority 2013) personal interaction not being placed in a secondary plan.

Tailored reality

Any interaction we have with the Internet implies the possibility of storing our personal data. The processing and analysis of this data determines the personalized results that appear later on the Internet, through our search results, the display of products in online stores, the display of advertisements, etc. This generates a narrower and more personalized version of a user's previous online experience (the so-called "filter bubble." (Pariser 2011) An advantage is that the user will quickly find what he or she usually looks for, but excluding certain aspects, perspectives and ideas can lead to a restriction of creativity and the development of a tolerant attitude through the political and social isolation of the other aspects, by the lack of pluralistic views. (Crawford, Gray, and Miltner 2014)

De-identification

De-identification involves deleting or hiding elements that could immediately identify a person or organization. Legislation in different countries on data protection defines different treatments for identifiable data. Identifiability is increasingly seen as a continuum, not a binary aspect. Disclosure risks increase simultaneously with the number of variables, data sources and the power of data analysis. Disclosure risks may be mitigated but not eliminated. De-identification remains a vital tool for ensuring the safe use of data. (UK Data Service 2017)

Perfectly anonymous information taken separately can be combined with other data to uniquely identify a person with varying degrees of certainty. Profiling can become a powerful tool, raising concerns about the degree to which intrusion into an individual's life is allowed, the possibility of ensuring security, and surveillance.

Digital inequality

The advantages of Big Data size are clear, but there are also opinions that the accumulation of data on a huge scale presents specific risks. Because of this, there are few entities that have access, through infrastructure and skills, to Big Data systems. In this context, the costs and skills needed for access lead to certain specific digital inequalities addressed by ethics.

Privacy

In data transactions it is very important to ensure confidentiality:

> "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks." - United Nations Declaration of Human Rights Article 12.

In many countries, public monitoring of the data by the government to observe citizens requires explicit authorization through an appropriate judicial process. Privacy is not about keeping secrets, but about choice, human rights, and freedom.

Often privacy is wrongly viewed as a binary choice between isolation and scientific progress. Identity protection in data is technologically possible, for example using homomorphic encryption and algorithmic design.

Privacy as a limitation of the use of data can also be considered unethical, (Kostkova et al. 2016) especially in healthcare, but it should be kept in mind that it is possible to extract the value of the data without compromising privacy.

Privacy is recognized as a human right by numerous national and international regulations. Privacy in research is achieved through a combination of approaches: limiting the collected data, anonymizing them; and regulating access to data. In the case of Big Data research, specific problems arise: the ambiguity between the terms "privacy" and "confidentiality; the declaration of social spaces as public or private; the ignorance of the risks of privacy by users; the blurred distinction between public and private users. Currently there are disputes whether data science it should be classified as a research of human subjects, and therefore not subject to the usual rules of privacy.

6. Big Data research

Through the new concepts of "algorithmic damage", "predictive analysis", etc., the algorithms currently used in Big Data operations go beyond the traditional view of privacy. According to the US National Science and Technology Council,

> "“Analytical algorithms” as algorithms for prioritizing, classifying, filtering, and predicting. Their use can create privacy issues when the information used by algorithms is inappropriate or inaccurate, when incorrect decisions occur, when there is no reasonable means of redress, when an individual’s autonomy is directly related to algorithmic scoring, or when the use of predictive algorithms chills desirable behavior or encourages other privacy harms.” (NSTC (National Science and Technology Council) 2016, 18)

Big Data research is what the ethicist James Moor would call a "conceptual muddles" due to the "inability to properly conceptualize the ethical values and dilemmas at play in a new technological context." (Buchanan and Zimmer 2018) In this situation privacy is ensured through a combination of different tactics and practices (controlled or anonymous environments, limitation of personal information, anonymization of data, access restrictions, data security, etc.). In general, all related concepts become confusing in the case of Big Data. Thus, social posts are considered public on social networks in case of an appropriate setting. But social networks are complex environments of socio-technical interactions where users do not always understand the functionality of the settings and terms of use. Thus, there is uncertainty about users' intentions and expectations, and these conceptual deficiencies in the context of Big Data research lead to uncertainties regarding the need for informed consent.

Conclusions

Critical data studies in Big Data reflect specific practices, cultures, policies and economies. (Dalton, Taylor, and Thatcher 2016) Issues can range from the intimacy and autonomy of individuals to the ethics of data science and institutional change due to Big Data research. It follows the need to analyze Big Data practices aware of power relations, prejudices and inequalities.

A definition that would restrict critical research to the field of normative and critical theory would be counterproductive.

The common principles of critical data studies highlight the interdependencies between emerging technologies and (human) actors in increasingly presented societies. Big Data are also a product of contemporary socio-technical conditions, because they are producing such conditions. (Richterich, 2018)

The field of science and technology studies (STS) has a rather ambiguous relationship with the normative evaluations of technology.

In STS, some components are more concerned with descriptive approaches than normative ones.

In contrast to the common STS ideal of "worthless" relativism, (Pels 1996, 277) Pels calls for the recognition of "third positions" in evaluations of scientific knowledge production that "‘[...] are not external to the field of controversy studied but are included and implicated in it. [...] They are not value-free or dispassionate but situated, partial and committed in a knowledge-political sense." (Pels 1996)

A major problem in Big Data is that the empirical micro-processes that underlie the appearance of their typical network characteristics are not well known. (Snijders, Matzat, and Reips 2012) Big Data should always be contextualized in their social, economic and political contexts. (Graham 2012)

Supporters of privacy are concerned about the threat to privacy due to the increased volume of storage and integration of personally identifiable information. In this regard, there are different policy recommendations to comply with the practice and privacy. (Ohm 2012) The misuse of Big Data by the media, companies and even the government has led to the loss of trust in social institutions. In order to protect individual freedoms, Nayef Al-Rodhan believes that a new type of social contract is needed, with the closer monitoring and regulation of Big Data. (Al-Rodhan 2018)

Scientific experiments tend to analyze data using specialized clusters and high-performance computers, rather than cloud, thus differentiating culturally and technologically from the rest of society.

The use of Big Data, due to the manipulation of large amounts of data, has led to the neglect of the principles of science, such as choosing representative samples, causing biases in the analysis of results. This analysis is often superficial compared to the analysis of smaller data sets. (Piatetsky 2014) Some data sources, such as Twitter, are not representative of the total population. Ioannidis argued that in using Big Data, "most published research findings are false" (Ioannidis 2005) as the probability of a "significant" result being false increases rapidly with the volume of data, but only positive results are published.

In using Big Data, the UK Data Service highlights several specific ethical issues: (UK Data Service 2017)

- Alternatives to informed individual consent, such as "social consent", have emerged and are more permissive.
- The need to respect the data source and, in general, "contextual integrity" in the case of data reuse has increased.
- Research ethics is mainly based on the idea that the researched entity is an individual person, so it would be possible to de-identify for protection.
- In the case of considering a group as a whole, social protection decreases. In this case it was proposed that the data be considered as "public benefits" or "public interest", but this does not solve the responsibility of the data users.

Matthew Zook et al. proposes "ten simple rules" for using Big Data in research. (Zook et al. 2017) The first five rules concern how to reduce the chances of injury resulting from research practices, and the other rules refer to best practices.

1. *Data is people and can harm*: most data represent or influence people. Start with the assumption that the data is personal (until proven otherwise) and guide your analysis on this basis.
2. *Privacy is more than a binary value*: privacy depends on the nature of the data, the context in which it was created and obtained, and on the expectations and norms of those affected. It extends to groups. Contextualize the data to anticipate a breach of privacy and to minimize harm.
3. *Avoid re-identifying your data*: it often fails to effectively anonymize your data. The data considered to be anonymous are combined with other variables that can lead to re-identification. Identify the possible vectors of re-identification and minimize them in published results.
4. *Practice ethical data exchange*: for some projects, such as genetics, data sharing is a social necessity, but informed consent and the right of withdrawal remain valid. Share the data in accordance with the research protocols but take into account the potential damage generated by the data collected informally.
5. *Consider the strengths and limitations of your data*: bigger does not automatically mean better: datasets must be grounded in their proper context, including taking into account conflicts of interest. In data acquisition, it is important to understand the source of the data, and to comply with the regulations. In poorly regulated environments, ethical rules can be used. Researchers need to be sensitive to the multiple potential meanings of the data. Document the provenance and evolution of the data.
6. *Debate tough, ethical choices*: the lack of clear solutions and protocols should be avoided. Such debates can produce very useful peer reviews. Consultation services can be used in the field

of research ethics in universities. Involve your colleagues and students in ethical practice for a large-scale Big Data research.

7. *Develop a code of conduct for your organization, research community or industry*: "false ethics", as well as falsifying data or results, are unacceptable. It is necessary to develop codes of conduct, which can provide guidance in the mutual evaluation of publications and in the examination of funding. Establish appropriate codes of ethical conduct, along with representatives of affected communities.
8. *Design your data and systems for auditing*: audit provides a mechanism for verifying work, increasing understanding and replicability. Plan and initiate audits of Big Data practices.
9. *Get involved with smaller consequences in data practices and analysis*: it is important for researchers to think beyond traditional values. Providers may be required to store in the cloud, and data processing centers may switch to sustainable and renewable energy sources. Carrying out large-scale research has effects at the society level.
10. *Know when to break these rules*: you must know what to expect when you move away from these rules, such as in natural disaster or emergency situations. Responsible Big Data research depends on several checklists.

Regardless of ethical or legal norms, scientists must be rigorous in the use of techniques and methodologies, and very careful in ethical issues. The idea that "data is already public" (Zimmer 2016) is an unjustified simplification. Data are not abstract; they are actually real people.

Responsible Big Data research does not aim at restricting research, but at ensuring confidence, fairness and maximizing positive aspects while reducing harm. Big Data offers fantastic opportunities for a better understanding of society and world, but ethical responsibility in the choices, practices and actions of research must also be taken into account.

Bibliography

Allen, Marshall. 2018. "Health Insurers Are Vacuuming Up Details About You — And It Could Raise Your Rates." Text/html. ProPublica. July 17, 2018. https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates.

Al-Rodhan, Nayef. 2018. "The Social Contract 2.0: Big Data and the Need to Guarantee Privacy." OpenMind. June 11, 2018. https://www.bbvaopenmind.com/en/humanities/beliefs/the-social-contract-2-0-big-data-and-the-need-to-guarantee-privacy-and-civil-liberties/.

Baert, Patrick, and Bryan Turner. 2004. "New Pragmatism and Old Europe: Introduction to the Debate between Pragmatist Philosophy and European Social and Political Theory." *European Journal of Social Theory* 7 (3): 267–74. https://doi.org/10.1177/1368431004044193.

Billings, Stephen A. 2013. "Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains." Wiley.Com. 2013. https://www.wiley.com/en-ro/Nonlinear+System+Identification%3A+NARMAX+Methods+in+the+Time%2C+Frequency%2C+and+Spatio+Temporal+Domains-p-9781119943594.

Brumfiel, Geoff. 2011. "High-Energy Physics: Down the Petabyte Highway." *Nature* 469 (7330): 282–83. https://doi.org/10.1038/469282a.

Buchanan, Elizabeth A., and Michael Zimmer. 2018. "Internet Research Ethics." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2018. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/ethics-internet-research/.

Collins, Tom. 2016. "Security Fears Sparked over Wearable Technology." Mail Online. December 19, 2016. http://www.dailymail.co.uk/~/article-4049154/index.html.

Constine, Josh. 2017. "Facebook Now Has 2 Billion Monthly Users… and Responsibility." *TechCrunch* (blog). 2017. http://social.techcrunch.com/2017/06/27/facebook-2-billion-users/.

Crawford, Kate, Mary L. Gray, and Kate Miltner. 2014. "Big Data| Critiquing Big Data: Politics, Ethics, Epistemology | Special Section Introduction." *International Journal of Communication* 8 (0): 10. https://ijoc.org/index.php/ijoc/article/view/2167.

Dalton, Craig M., Linnet Taylor, and Jim Thatcher. 2016. "Critical Data Studies: A Dialog on Data and Space." In . https://doi.org/10.1177/2053951716648346.

Dean, Jeffrey, and Sanjay Ghemawat. 2004. "MapReduce: Simplified Data Processing on Large Clusters." http://static.googleusercontent.com/media/research.google.com/en//archive/mapreduce-osdi04.pdf.

Dedić, Nedim, and Clare Stanier. 2017. "Towards Differentiating Business Intelligence, Big Data, Data Analytics and Knowledge Discovery." In *Innovations in Enterprise Information Systems Management and Engineering*, edited by Felix Piazolo, Verena Geist, Lars Brehm, and Rainer Schmidt, 114–22. Lecture Notes in Business Information Processing. Springer International Publishing.

DG Connect. 2015. "Study on Personal Data Stores Conducted at the Cambridge University Judge Business School." Text. Digital Single Market - European Commission. August 7, 2015. https://ec.europa.eu/digital-single-market/en/news/study-personal-data-stores-conducted-cambridge-university-judge-business-school.

European Commission. 2019. "Horizon 2020." Text. Horizon 2020 - European Commission. 2019. https://ec.europa.eu/programmes/horizon2020/en.

European Economic and Social Committee. 2017. "The Ethics of Big Data: Balancing Economic Benefits and Ethical Questions of Big Data in the EU Policy Context." European Economic and Social Committee. February 22, 2017. https://www.eesc.europa.eu/en/our-work/publications-other-work/publications/ethics-big-data.

European Parliament. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA Relevance)*. *OJ L*. Vol. 119. http://data.europa.eu/eli/reg/2016/679/oj/eng.

Everts, Sarah. 2016. "Information Overload." Science History Institute. July 18, 2016. https://www.sciencehistory.org/distillations/magazine/information-overload.

Floridi, Luciano. 2014. *The Fourth Revolution: How the Infosphere Is Reshaping Human Reality*. OUP Oxford.

Graham, Mark. 2012. "Big Data and the End of Theory?" *The Guardian*, March 9, 2012, sec. News. https://www.theguardian.com/news/datablog/2012/mar/09/big-data-theory.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Joas, Hans. 1993. *Pragmatism and Social Theory*. University of Chicago Press.

Keulartz, Jozef, Maartje Schermer, Michiel Korthals, and Tsjalling Swierstra. 2004. "Ethics in Technological Culture: A Programmatic Proposal for a Pragmatist Approach." *Science, Technology, & Human Values* 29 (1): 3–29. https://doi.org/10.1177/0162243903259188.

Kostkova, Patty, Helen Brewer, Simon de Lusignan, Edward Fottrell, Ben Goldacre, Graham Hart, Phil Koczan, et al. 2016. "Who Owns the Data? Open Data for Healthcare." *Frontiers in Public Health* 4. https://doi.org/10.3389/fpubh.2016.00007.

Kvochko, Elena. 2012. "Four Ways to Talk About Big Data." Text. Information and Communications for Development. December 4, 2012. http://blogs.worldbank.org/ic4d/four-ways-to-talk-about-big-data.

Manyika, James, Michael Chui, Jaques Bughin, and Brad Brown. 2011. "Big Data: The next Frontier for Innovation, Competition, and Productivity." 2011. https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Reprint edition. Boston: Eamon Dolan/Mariner Books.

Mearian, Lucas. 2015. "Insurance Company Now Offers Discounts -- If You Let It Track Your Fitbit." Computerworld. April 17, 2015. https://www.computerworld.com/article/2911594/insurance-company-now-offers-discounts-if-you-let-it-track-your-fitbit.html.

MIKE2.0. 2019. "Big Data Solution Offering - MIKE2.0, the Open Source Methodology for Information Development." 2019. http://mike2.openmethodology.org/wiki/Big_Data_Solution_Offering.

MORO, Veronica. 2016. "Meeting the Challenges of Big Data." Text. European Data Protection Supervisor - European Data Protection Supervisor. November 16, 2016. https://edps.europa.eu/data-protection/our-work/publications/opinions/meeting-challenges-big-data_en.

Nature. 2008. "Community Cleverness Required." *Nature* 455 (7209): 1. https://doi.org/10.1038/455001a.

Noorman, Merel. 2012. "Computing and Moral Responsibility." *Stanford Encyclopedia of Philosophy*.

Norwegian Data Protection Authority. 2013. "Big Data – Privacy Principles under Pressure." https://www.datatilsynet.no/globalassets/global/english/big-data-engelsk-web.pdf.

NSTC (National Science and Technology Council). 2016. "National Privacy Research Strategy." https://obamawhitehouse.archives.gov/sites/default/files/nprs_nstc_review_final.pdf.

OECD. 2013. "New Data for Understanding the Human Condition: International Perspectives." http://www.oecd.org/sti/inno/new-data-for-understanding-the-human-condition.pdf.

———. 2016. "Research Ethics and New Forms of Data for Social and Economic Research," November. https://doi.org/10.1787/5jln7vnpxs32-en.

Ohm, Paul. 2012. "Don't Build a Database of Ruin." *Harvard Business Review*, August 23, 2012. https://hbr.org/2012/08/dont-build-a-database-of-ruin.

Pariser, Eli. 2011. *The Filter Bubble: What The Internet Is Hiding From You*. Penguin Books Limited.

Pels, Dick. 1996. "The Politics of Symmetry." *Social Studies of Science* 26 (2): 277–304. https://doi.org/10.1177/030631296026002004.

Piatetsky, Gregory. 2014. "Interview: Michael Berthold, KNIME Founder, on Research, Creativity, Big Data, and Privacy, Part 2." 2014. https://www.kdnuggets.com/2014/08/interview-michael-berthold-knime-research-big-data-privacy-part2.html, https://www.kdnuggets.com/2014/08/interview-michael-berthold-knime-research-big-data-privacy-part2.html.

Reichman, O. J., Matthew B. Jones, and Mark P. Schildhauer. 2011. "Challenges and Opportunities of Open Data in Ecology." *Science* 331 (February): 703. https://doi.org/10.1126/science.1197962.

Richterich, A. 2018. "The Big Data Agenda: Data Ethics and Critical Data Studies." https://doi.org/10.16997/book14.b.

Rip, Arie. 2013. "Pervasive Normativity and Emerging Technologies." In *Ethics on the Laboratory Floor*, edited by Simone van der Burg and Tsjalling Swierstra, 191–212. London: Palgrave Macmillan UK. https://doi.org/10.1057/9781137002938_11.

Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips. 2012. "'Big Data': Big Gaps of Knowledge in the Field of Internet Science." http://www.ijis.net/ijis7_1/ijis7_1_editorial.pdf.

Sullivan, Danny. 2015. "Google Still Doing At Least 1 Trillion Searches Per Year." Search Engine Land. January 16, 2015. https://searchengineland.com/google-1-trillion-searches-per-year-212940.

Tanner, Adam. 2014. "Different Customers, Different Prices, Thanks To Big Data." Forbes. 2014. https://www.forbes.com/sites/adamtanner/2014/03/26/different-customers-different-prices-thanks-to-big-data/.

UK Data Service. 2017. "Big Data and Data Sharing: Ethical Issues." https://www.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethical-issues.pdf.

Zimmer, Michael. 2016. "OkCupid Study Reveals the Perils of Big-Data Science." *Wired*, May 14, 2016. https://www.wired.com/2016/05/okcupid-study-reveals-perils-big-data-science/.