Data Mining: Ethics Written Work

William H Muter

1 Introduction

Data Mining refers to the extraction of valuable information, generated through the identification of previously unknown patterns and trends within a dataset. It is performed in a wide spectrum of domains, and the collections of data upon which these techniques are usually applied tend to be so large that they render traditional analysis techniques as ineffective; these datasets are commonly referred to as "Big Data" (Lee, 2017).

In recent years, as both companies and government agencies have been accumulating ever-increasing quantities of data, the pursuit of valuable knowledge through Data Mining techniques has raised questions over the ethics of Big Data; especially regarding the handling of private and sensitive information (Oussous et al., 2017). This essay will discus these issues of morality and ethics.

2 Ethical Perspectives of Data Mining

Within the field of ethics there are two key orientations; relativism (that suggests that there is no objective sense of right and wrong) and absolutism (the notion that there are objective moral truths). As relativism requires a definite context, and as big data is applicable within a wide range of domains, this essay focuses on absolutism. When considering the ethics of Big Data in general, three of these perspectives were considered; utilitarianism, Kantian ethics and ethics of rights.

2.1 Utilitarianism

Utilitarianism is a teleological (outcome-focused) ethical perspective that states that an action is morally right if it results in the greatest amount of benefit for the largest number of people (Crane and Matten, 2016). Using this perspective of gauging morality, it can be suggested that if the improvement in the quality of service that results from the implementation of data mining techniques outweighs the cumulative adverse outcomes that may occur, it can be considered as an ethical venture; implying that aggressive data elicitation and mining is an acceptable practice. The issue with this analysis however is that it can be difficult to predict or foresee the improvement to service / adverse effects that can result from data mining and therefore placing values on either of these variables is a very complex task.

2.2 Kantian Ethics

Kant (1785) produced three maxims that were designed to question an individuals categorical imperative (the obligations felt by a person, i.e. duty). The first maxim encourages the decision-maker to consider the longevity of their actions, such as what would happen if everybody behaved this way. His second maxim considers whether the interests of others have been considered, or if they have just been ignored to achieve goals. The final maxim considers whether an act would be considered acceptable by others.

Providing entities collecting the data are not behaving in a reckless manner with data, are not unduly profiteering from the data they collect concerning individuals, and believe that external parties would consider their actions as acceptable; then data mining can be seen as an ethical endeavour from a Kantian perspective.

2.3 Ethics of Rights

Ethics of rights is the concept that all individuals have a set of entitlements that should be protected during any decision-making processes. Two clear examples are Article 8 of the European Convention on Human Rights (1970) and Article 12 of the The Universal Declaration of Human Rights (1948) that both defend a person's right to respect for their private and family life, as well as their correspondence. In addition to these protected rights, laws such as the Data Protection Act 1998 (1998), and the European Union directives, Directive 95/46/EC (1995) and Directive 2002/58/EC (2002), have been implemented to safeguard the protection of privacy and its free movement.

The argument that these laws address any prevalent ethical concerns is explored by van Wel and Royakkers (2004), however they find this unconvincing and state that "the law is not, and never will be fully sufficient with respect to privacy problems". Al-Saggaf and Islam (2015) reinforce this notion by stating that the information generated via data mining "circumvents the normative protection of personal informational privacy".

Using this view, data mining can be seen as an ethical practice, providing individuals involved are aware of what their data is being used for and consents to its collection. It should also be noted that in some cases, business my need to go beyond what is required of them by law to ensure they are taking an ethical approach.

3 Ethical Issues in Big Data

3.1 Mass Surveillance

A key ethical dilemma surrounding Big Data is its use within mass surveillance projects to track the online behaviours of individuals. This notion dates back to Jeremy Bentham's work on Panopticon prisons, and the exertion of control through constant observation. Watt (2017) discussed the idea of peacetime espionage, mentioning that although it has been occurring for many years, the documents leaked by Edward Snowden have only recently revealed the extent of government monitoring of communication technology. Examples of this can be seen in endeavours such as PRISM, the mass surveillance program run by the United States National Security Agency (Sottek and Kopfstein, 2013); and more worryingly, China's Social Credit System, where the findings of their surveillance are being use to directly influence social opportunities (Botsman, 2017).

Although the use of Big Data for these purposes breach human rights, arguments have been made in proposition of these practices, as it is considered by some necessary for terrorism prevention and cybercrime, and has resulted in the implementation of legislation such as the Patriot Act; however infringement on human rights should be taken very seriously and the ethics regarding this topic are still a matter for much debate.

3.2 Use Transparency

Another prominent ethical dilemma surrounding Big Data is the notion that data is being collected, stored, and managed for purposes other than those stated to users; with companies such as Facebook, WhatsApp and Twitter having already been criticised for their unclear and confusing data privacy policies (Alharthi et al., 2017).

Lipworth et al. (2017) discuss this issue further, highlighting its particular prevalence with medical data, and note that several medical sources have found that the purposes which people are consenting for their data to be used for can no longer be relied upon, and highlight the loss of control that individuals are facing regarding what their data is being used for.

3.3 Third Party Access

Data Mining allows for the generation of very commercially valuable data, and consequently it provides an incentive for businesses to sell some of the insights they discover. Horvitz and Mulligan (2015) identify the role that machine learning plays in the analysis of publicly available data. As an example they highlight a suicide prevention application created by the charity Samaritans was shut down by twitter for not complying with its terms of use policy. The app trawled through twitter feed of individuals and used an algorithm to search for key terms that identified people who may be considered at risk of committing suicide (Samaritans, 2015).

Although Hacker and Petkova (2017) defend the use of data analytics for the generation of targeted advertisement in the pursuit of revenue, labelling it as "conscionable", they state this evaluation changes once third parties are involved; however it could be argued that information placed on public forums (such as social media websites), is "fair game" and that these companies should not be responsible for safeguarding what a person shares online.

3.4 Re-identification of Anonymous Data

Re-identification refers to the phenomenon of processing a previously sanitised dataset in such a way that it allows for connections between data and the individuals that it concerns to be established. Lubarsky (2011) discusses personal data, particularly focussing on the notion of differing levels of identifiability. He suggests that information can be split into five categories ranging from data which is very individual specific to "scrubbed" data which has either minimal or no analytical value at a personal level.

Re-identification is primarily achieved through the combination of data sources which when amalgamated provide an indication of who the data belongs to. Algorithms are used to search for correlations between anonymous identifiers (identifiers that do not rely on personal information); in the hope of associating them with data that has high levels of identification, thus revealing the owners of this data.

Data handlers can take measures to prevent the occurrence of re-identification by anonymising data and subsequently increasing the complexity of identifying individuals in datasets, however, this does not prevent companies collecting and analysing information on their own platforms (Barocas and Nissenbaum, 2014). This essentially means that although effective data sanitisation can help to prevent personal data from being leaked; it means that insights can still be derived and distributed to third parties, so does not necessarily ensure privacy.

Level of Identifiability	Description	Example
High	Direct identifier	Name, address, passport number, national insurance number
	Indirect identifier	Date of birth, postcode, vehicle registration number, IP address, geolocation
	Data that can be linked to multiple individuals	Film or shopping preferences
	Data that cannot be linked to any individual	Aggregated census data, survey results
Low	Data not related to individuals	Weather reports

Table 1: Levels of Identification: Adapted from Lubarsky (2011)

Currently, the act of data re-identification is not an illegal practice in its own right, however, Phillips et al. (2017) state in their paper that is is unclear what the criminalisation of re-identification would add to data protection laws already in place.

4 Example Case Study: The "Netflix Prize" Algorithm Improvement Contest

Despite a paper published in 2001, by Ramakrishnan et al. (2001), which stated the potential risk of identification from recommendation systems, in 2006 the video streaming service Netflix launched a public competition, offering one million dollars to anyone who could improve the algorithm used by their viewing recommendation system by 10% (Netflix, 2006). For the algorithms to be optimised, a sanitised dataset containing over one-hundred-million movie ratings generated by nearly half-a-million users was released to the public, for developers to use for training and testing purposes. Two years after the release of this dataset, researchers from the University of Texas published a paper regarding re-identification which demonstrated that they had successfully identified some of the individuals whose preference were contained within the dataset (Singel, 2009). Within this paper they described how they combined the data released by Netflix with auxiliary information harvested from public profiles on the International Movie Database (IMDB) website to generate the identities of some individuals (Narayanan and Shmatikov, 2008).

This case was brought to prominence in the media a year after the release of this paper, when an anonymous closeted lesbian mother (named as Jane Doe for anonymity), who claimed that she could be outed because of the potential leak of her viewing preferences, led a class action lawsuit against Netflix, seeking $2,500 for each person aggrieved by their actions (Valdez-Marquez et al v. Netflix Inc, 2009). Following this lawsuit, under the terms of the settlement Netflix paid $9,000,000 into a fund with the purpose of benefiting privacy groups and agreed to change their data retention policies to no longer link customers to their viewing histories for more than one year after customer cancellation (Vagle and Klein, 2012).

5 Conclusion

This essay has examined the morality of data mining in general, analysing it from multiple perspectives; and has also highlighted some of the key ethical issues prevalent in both Data Mining and Knowledge Discovery, discussing topics such as surveillance, transparency, data sharing, and re-identification.

Günther et al. (2017) provide the notion, that as most of the entities that are collecting data a large scale are businesses, the free market will drive their decision making. Therefore it can be argued that moral choices will be undertaken as a consequence of consumer pressure; and businesses will use their approach to ethical issues as a method of differentiating themselves from competitors. However, this notion is rebutted by Herschel and Miori (2017), who have no doubt that there is a growing issue with the ethics of Big Data, with predictions that by 2018, 50% of ethical violations within businesses will occur through improper use of Big Data analytics. Additionally, this view assumes that individuals are aware of what their information is being used for, which is not always true; and also ignores the findings of Lee (2017), who suggests that protecting privacy is often counter-productive to firms and customers stifling their ability to use big data to enhance services; therefore, companies will disregard ethical approaches to data handling to maximise efficiency, and customers will be reluctant to question business motives if it results in a higher quality of service.

Overall, this essay has found that ensuring ethical behaviour in data mining is a complex task and that laws may not necessarily be enough; therefore there is a strong argument for the democratisation of data collection (as suggested by Hacker and Petkova (2017)), where large companies that collect vast amounts of data should conduct regular surveys among their users to ensure there is a general consensus of understanding regarding what these businesses do with data. This paper also recommends that companies deriving insight from data should consider mandatory legal and ethical consideration training for staff.

References

Al-Saggaf, Y. and Islam, M. Z. (2015). Data mining and privacy of social network sites users: Implications of the data mining problem. *Science and Engineering Ethics*, 21(4):941 – 966.

Alharthi, A., Krotov, V., and Bowman, M. (2017). Addressing barriers to big data. *Business Horizons*, 60:285 – 292.

Barocas, S. and Nissenbaum, H. (2014). *Big Datas End Run around Anonymity and Consent*. Cambridge University Press.

Botsman, R. (2017). Big data meets big brother as china moves to rate its citizens. Available at: `http://www.wired.co.uk/article/chinese-government-social-credit-score-privacy-invasion`.

Crane, A. and Matten, D. (2016). *Business ethics : managing corporate citizenship and sustainability in the age of globalization*. Oxford, United Kingdom : Oxford University Press, [2016].

Data Protection Act 1998 (1998). Legislation.gov.uk. Available at: `http://www.legislation.gov.uk/`.

Directive 2002/58/EC (2002). The european parliament and the council of the european union.

Directive 95/46/EC (1995). The european parliament and the council of the european union.

European Convention on Human Rights (1970). The european parliament and the council of the european union.

Günther, W. A., Mehrizi, M. H. R., Huysman, M., and Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. *Journal of Strategic Information Systems*, 26(3):191–209.

Hacker, P. and Petkova, B. (2017). Reining in the big promise of big data: Transparency, inequality, and new regulatory frontiers. *Northwestern Journal of Technology & Intellectual Property*, 15(1):1.

Herschel, R. and Miori, V. M. (2017). Ethics & big data. *Technology in Society*, 49:31–36.

Horvitz, E. and Mulligan, D. (2015). Data, privacy, and the greater good. *Science*, 349(6245):202–213.

Kant, I. (1785). *Grounding for the Metaphysics of Morals [1993]*. Hackett. Translated by Ellington, James W.

Lee, I. (2017). Big data: Dimensions, evolution, impacts and challenges. *Business Horizons*, 60:293–303.

Lipworth, W., Mason, P. H., and Kerridge, I. (2017). Ethics and epistemology of big data. *Bioethical Inquiry Symposium on Ethics and Epistemology of Big Data*, 14:485–488.

Lubarsky, B. (2011). Re-identification of anonymized data. *Georgetown Law Technology Review*, pages 202–213.

Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *IEEE Symposium on Security and Privacy*, pages 111–125.

Netflix (2006). Netflix prize: Review rules. `https://www.netflixprize.com/rules.html`.

Oussous, A., Benjelloun, F.-Z., Lahcen, A. A., and Belfkih, S. (2017). Big data technologies: A survey. *Journal of King Saud University – Computer and Information Sciences*.

Phillips, M., Dove, E. S., and Knoppers, B. M. (2017). Criminal prohibition of wrongful reidentification: Legal solution or minefield for big data? *Bioethical Inquiry Symposium on Ethics and Epistemology of Big Data*, 14:527–539.

Ramakrishnan, N., Keller, B. J., Mirza, B. J., Grama, A. Y., and Karypis, G. (2001). Privacy risks in recommender systems. *IEEE Internet Computing*, 5(6):54–62.

Samaritans (2015). Samaritans radar. `https://www.samaritans.org/how-we-can-help-you/supporting-someone-online/samaritans-radar`.

Singel, R. (2009). Netflix spilled your brokeback mountain secret, lawsuit claims. `https://www.wired.com/2009/12/netflix-privacy-lawsuit/`.

Sottek, T. and Kopfstein, J. (2013). Everything you need to know about prism. Available at: `https://www.theverge.com/2013/7/17/4517480/nsa-spying-prism-surveillance-cheat-sheet`.

The Universal Declaration of Human Rights (1948). United nations general assembly in paris.

Vagle, J. L. and Klein, S. R. (2012). Class actions adding to the cost of data breaches. `http://www.pepperlaw.com/publications/class-actions-adding-to-the-cost-of-data-breaches-2012-10-24/1`.

Valdez-Marquez et al v. Netflix Inc (2009). Jane Doe, individually; Nelly Valdez-Marquez, Anthony Sinopoli, Paul Navarro, individually and on behalf of a class of similarly situated individuals v. Netfilx Inc. a Delaware Corporation, and DOES 1 THROUGH 50, inclusive. Available at: `https://www.wired.com/images_blogs/threatlevel/2009/12/doe-v-netflix.pdf`.

van Wel, L. and Royakkers, L. (2004). Ethical issues in web data mining. *Ethics and Information Technology*, 6(2):129–140.

Watt, E. (2017). the right to privacy and the future of mass surveillance. *International Journal of Human Rights*, 21(7):773 – 799.