BHBIA Data Analytics Guidelines
Big Data and Data Protection

These guidelines are part of a series designed to provide guidance on the legal and ethical issues impacting data analysts

INTRODUCTION

This article highlights some of the legal and ethical considerations arising when processing and analysing Big Data, above and beyond those likely to affect more traditional data sets.

While there is often little inherent difference in the ethics of holding Big Data compared to other data, the special considerations are rather to do with the way it is likely to be analysed and processed.

GENERAL PRINCIPLES

A review of the information collected should be conducted to determine whether it contains any personal data. As with any information, data protection principles should be considered when dealing with Big Data if it could be used to identify an individual.

By definition, Big Data containing personal data is likely to be considered a high risk asset due to the volume of personal data it contains, and the high impact of any breach or misuse of those data. There is an even higher level of responsibility placed on the controllers and processors of these data in law if these data contain special category data (previously referred to as sensitive data), which includes information regarding an identifiable individual's health.

The first consideration should be whether or not it is necessary to be able to identify individuals from the data. If this is not a requirement, for instance if the data is being used for general analysis of populations or cohorts, then anonymising the data by removing identifiable fields will remove the need to consider data protection on the anonymised dataset.

Processing Big Data containing personal data can provide challenges under some of the data protection principles, and these are considered in the following sections.

THE "FAIR AND LAWFUL" PRINCIPLE

Under this principle, data subjects have the right to know what their data is being used for, and for this to be transparent. It is important to ensure that the privacy statements agreed to by the data subjects detail all of the purposes to which the data is put where consent is the legal basis for processing the data.

With increasing use of complex algorithms and machine learning being applied to Big Data sets, it is increasingly difficult to provide simple, understandable descriptions of the processing. It is key therefore to give substantial attention to the effect on individuals of these types of processing, and to consider whether the data subject would reasonably expect their data to be used in the manner it is, given the consent that they have provided.

Under the General Data Protection Regulation (GDPR)/Data Protection Act (DPA) 2018, data subjects have the right not to be subjected to fully automated decision making where the outcome has a material impact on them. There is therefore an onus on the data controller to ensure that appropriate techniques are applied accurately, and that they could not lead to discrimination based on special category data factors (e.g. race, health etc.).

THE PURPOSE OF PROCESSING

Big Data often offers massive diversity in its potential use, with results from one analysis leading to the formulation of new algorithms to gain new insights from the data. However, just because an analysis is possible in the data, this does not mean that there is an automatic right to perform it. Where consent is the basis for processing, it must be established whether any new use of the data is still covered by the original consent given. Where a legitimate interest of the organisation is used as the legal basis, there is a greater responsibility on the organisation to assess the data use and ensure it is respectful of the rights and interests of the data subjects.

Where data is purchased from a third party and integrated with other Big Data, there is still a responsibility to ensure the data is used in accordance with the purposes set out in the third party's privacy statement.

ADEQUACY OF DATA AND RETENTION

There is often a tendency with Big Data to collect as much information as possible, but where personal data is concerned, organisations still have a responsibility to process and store (including in backups) only the data that is necessary to fulfil the specified purpose. If personal data is required in the data, consider whether there are aspects that are not needed (e.g. date of birth might be stored where year of birth could be adequate for the purpose). Reducing the data stored to the minimum required can also help with data volumes and processing speeds and costs.

There is always an argument to hold on to data and to build up a historical catalogue, but under data protection legislation, data subjects need to be told how long their data will be stored for, and the data must be deleted thereafter in most cases, however valuable.

ACCURACY OF DATA

Data controllers are required by law to make reasonable efforts to ensure the personal data they hold are accurate. This can be a challenge with Big Data, not only because of the volumes involved, but also because there is often less validation built into its collection. Data subjects may provide false information as a means of protecting their privacy, or more inaccuracy may be tolerated in the data as it is thought that the size of the dataset will marginalise incorrect data.

Particular attention should be given to data modelling and machine learning techniques based on whole populations, where inaccuracy in the model may be biased towards subsets of the population (e.g. the model is 95% accurate for the population, but only 50% accurate for a particular minority). This can cause "unfairness" in the decisions made as described above. Where possible, make machine learning algorithms auditable and check them for bias.

Similarly, care should be taken when interpreting outcomes derived from complex algorithms on Big Data to determine whether they represent causal insights or merely correlation of factors.

RIGHT OF DATA SUBJECTS

The quantities of data involved and the complexity of processing Big Data make it difficult and often expensive to comply with the rights of data subjects. Data subjects have the right to receive a copy of the information held on them in a simple form, and for their data to be deleted for instance. How this can be achieved is likely to be a consideration at the design stage of processing when using Big Data containing personal data. The difficulty of meeting these rights is not considered an acceptable excuse for non-compliance with them under the law.

However, where Big Data systems bring all data about data subjects together, the use of such an approach could make compliance with requests to provide and delete data regarding an individual easier than in fragmented legacy systems.

SECURITY AND LOCATION OF PROCESSING

Due to the processing power required to manage Big Data processes, Cloud based computing is often used. Risk assessments should be performed to ensure the security standards met by any such third party are adequate. There are various

internationally recognised standards that can be help assess this, such as ISO 27001 (information security), ISO 14644 (data centre cleaning), ISO 9000 (Quality).
The location of the data centre, as well as any backups or failover systems should also be considered, as there are legal complications if the personal data of European citizens is processed outside of the European Economic Area.

PRIVACY BY DESIGN

As Big Data solutions are often complex or expensive to change once in place, it is important to adopt a privacy by design approach. A thought out strategy for which data can be anonymised, how data can be extracted, what purposes need to be in the privacy statement when the data is collected, and how information can be kept securely will make compliance with data protection law simpler and cheaper in the long run.

For further information see:

Data protection legislation: https://ico.org.uk/for-organisations/guide-to-data-protection/

ICO guidance on Big Data: https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

Guide to EU legislation and privacy by design: https://ec.europa.eu/info/law/law-topic/data-protection/reform/rules-business-and-organisations/obligations/what-does-data-protection-design-and-default-mean_en

For a definition of special category data, see https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data/

For information on the legal bases for processing data see https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/

This guidance is provided by the BHBIA for information purposes only and is not intended and should not be construed as regulatory or legal advice. It does not cover all legislative and regulatory requirements pertaining to Members and it is the responsibility of all Members to familiarise themselves with these.

The Guidelines are provided by the Data Analytics Guidelines Team within the BHBIA's Ethics & Compliance Committee,

Jason Bryant, Data Analytics Team Lead
Darren Kottler, Data Analytics Team
Klaas Breukel, Data Analytics Team
Catherine Ayland, BHBIA Ethics Advisor

If you have any queries about these Guidelines, please visit www.bhbia.org.uk and submit your query via 'My BHBIA' dashboard. Please note: this ad hoc advisory service is available to full BHBIA members only.

British Healthcare Business Intelligence Association
Ground Floor, 4 Victoria Square, St. Albans, Herts AL1 3TF
t: 01727 896085 • admin@bhbia.org.uk • www.bhbia.org.uk

A Private Limited Company Registered in England and Wales No: 9244455