Background to BCS Position Statement on the Ethical use of Big Data

Louise Bennett

> "It is a very common saying that figures never lie. But it is very certain that in the hands of the ignorant, the careless, the undiscriminating, they may become most potent instruments of falsehood." – Isaac Ray 1849.

Historical context

Darrell Huff introduced his book How to Lie with Statistics[1] in 1954 with a series of quotes. One from H G Wells: "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write", was truly prescient. We have now entered the era of "Big Data" and it is vital for citizens to understand these new data sets, how they are collected, aggregated, analysed and used. While data used by early statisticians, scientists and accountants was principally numbers and observations set out in a structured form, big data is largely unstructured data, including free form text and videos, that is not in a database per se, it is simply accumulated, often "because we can". Data is all valuable for something, provided it is understood and accurate. However, one of the characteristics of big data is that it is often: not accurate, the provenance may be unclear or unknown, it is seldom peer reviewed and audited, often written off the cuff for one purpose and analysed and used for another. Not only that, but the big data used by advertisers, analysed by governments, used to build business models and government policy includes much personal data, sometimes associated with us as citizens and sometimes anonymised.

When the Statistical Society of London was formed in 1834, their corn sheath logo included the phrase translated as: "to be thrashed out by others". Facts are derived from data. Data forms the basis of evidence. Traditionally data has been numbers – observations of the natural world and accounting figures. Ideally data and numbers are neutral, as science is. Data is needed for evidential purposes and to formulate policy. However, over time, data and numbers have become incentives. Incentives to: stay in power, earn more money, win arguments, build a business. When data become an incentive they will be manipulated[2] and require ethical scrutiny.

The opening up of government, corporate and private data as massive online data sources is of great potential value, from apps that show us crime hot spots to data bases that enable diseases to be understood and cured and victims of disasters to be located. However, alongside the benefits, there is also the threat of abuses, such as: the creation of an infrastructure of permanent surveillance, the discovery of the whereabouts of victims, disclosure of socially stigmatising diseases or conditions or the targeting of fraudsters.

If data sets contain personal data they are subject to the provisions of the Data Protection Act 1998 (DPA) in the UK. The Act is based on eight principles which state that personal data must be:

1. Processed fairly and lawfully;
2. Obtained and only used for specified and lawful purposes;
3. Adequate, relevant and not excessive;
4. Accurate, and where necessary, kept up to date;
5. Kept for no longer than necessary;
6. Processed in accordance with the individual's rights;
7. Kept secure;
8. Only transferred to countries that offer adequate data protection.

---

[1] How to lie with statistics, Darrell Huff 1954

[2] Facts from figures , M. J. Moroney 1951,

As already stated, the distinguishing feature of "Big Data" is that it is largely unstructured, including free form text and video and is also frequently: neither peer reviewed, nor audited, nor necessarily accurate, nor kept in a single global jurisdiction. It is often collected and kept "in case it is needed in the future" rather than collected for a specific purpose to address an identified problem or test a hypothesis. This means that if Big Data includes personal data, it is likely to fall foul of principles 2, 3, 4, 5, 6 and 8 of the DPA. Even if the original data sets do not contain personally identifiable data, it is likely that when several of these sets are aggregated and "mashed up" identity discovery will be possible.
As a society we need to debate the collection, aggregation and analysis of big data and decide where the boundaries of individual and societal good lie. The ethics of big data aggregation and analytics must be transparent and open. Each individual must be able to opt in or out of being part of it in each context of use.

Big Data and the Internet

The internet has continued to expand in 2013 and is becoming increasingly important globally. Growth related to the Internet economy is forecast at almost 11% in the EU, with a contribution to GDP expected to rise from 3.8% in 2010 to 5.7% in 2016[3]. Small and medium-sized enterprises that use the Internet intensively grow almost twice as fast as others. This economic potential is closely bound up with the exploitation of "Big Data". If this is to be exploited to the full society needs to ensure that individuals can access the content, goods and services they want, while controlling what personal data they want to share or not. For this to happen, the exploitation of the internet - by business for economic benefit, by individuals for social ends, and by governments both for the delivery of services to their citizens and to enable them to protect their citizen's from harm needs open discussion. The requirement is to maintain transparency and consensus on proportionality, so that we proceed in a way that maximises benefits and minimises the opportunity for harm.

Big Data collection, aggregation and analysis, particularly where parts of the data sets contain personally identifiable data, is a major ethical issue. The fact that it has recently been emotively dominated by NSA surveillance has detracted attention from other aspects and an understanding that it is a much broader issue than this. For example, people tend to overlook what non-government organisations and commercial organisations are doing with big data, and instead focus their attention solely on governments spying. Yet the privacy issues surrounding data collection and analytics are enormous and require a rational, unemotional debate about when societal good outweighs personal privacy.

The analysis of large volumes of data for specific data types, known as data aggregation, can lead to identity discovery even where data has been "anonymised" and is a relatively easy task where the data has only been "pseudo-anonymised".

Ethical scrutiny of both the collection and analysis of massive open data online is needed so that it can be used for social good. Since the collection and processing of large volumes of data is both cheap and possible, people and organisations are going to do it. It is important to control the analytical risks and not forgo the opportunities to use such data for humanitarian reasons and individual benefits.

In societal terms, the potential benefits include such things as: using mobile phone geo-location to track victims of disasters, as was done in Haiti and after Hurricane Sandy, or using internet performance monitoring data to identify patterns of censorship in repressive regimes, like Iran. Discussion is needed on the ethical issues of the collection,

[3] Boston Consulting Group, 'The $4.2 Trillion Opportunity – the Internet Economy in the G-20', 3/2012

“anonymisation” and “pseudo-anonymisation”, and analysis of big data online. Such analysis can be benign, as in tracking disaster victims or plotting the spread of a new disease as it crosses the globe, or oppressive as in tracking a country’s citizens through cyberspace and using that information for “control” purposes.

The emergence of crowdsourced large meta-data data sets such as the MetaPhone data set[4] adds more questions to the ethical discussion[5]. For example, how do you control such crowdsourced initiatives; who is behind such initiatives? Are their motives benign? Could the initiative be subverted to support another more nefarious cause? Indeed the emergence of these crowdsourced data sets shows that the internet itself can be viewed as one large repository of data just waiting to be harvested and mined for useful nuggets of information.

Treating the internet as a big data source and using tools such as web crawlers and other automated and manual tools such as Paterva’s Maltego to interrogate available resources on the internet opens the door to the darker side of big data analysis: that of actively seeking out specific individuals and targets for crime. As these tools become more powerful and ubiquitous, it will become increasingly difficult to guarantee an individual’s anonymity in the internet even when they are using anonymising mechanisms[6].

While search engines are a source of data that can be mined, social media sites offer a far richer and more interesting source of data. However, the one thing both have in common is the monetisation of information obtained from analysis of data held. In this sense, the user of these services is not a consumer of the services, but the product or part of the product sold by these services. Because of the revenue gained from selling the information obtained from analysing the held data, many of these services are free to the end user. One could ask, “is data the next snake oil?”

In describing the future of omnipresent computing where augmented reality is an ever-present convenience, and where the population expects to have information on products, ideas, and other people appear just by looking at them, theoretical physicist Dr Michio Kaku says that privacy will be a problem.

"People will demand to live in a world where they know everything about a product, and the producer will demand to know everything about the consumer [thanks to big data]," said Kaku. "And this is just how we are going to live."

Despite being able to have biographies appear about other people involved in meetings, Kaku believes that banks, corporations and governments will turn to quantum cryptography to ensure that their communications are secure. He also thinks they will gradually leave the Internet and retreat to their own Intranet so that they are confident nobody is listening in. This is already being mooted in the Brazilian Government response to the Snowden NSA leaks.

---

[4] http://antiwar.com/blog/2013/12/26/stanford-study-it-is-trivially-easy-to-identify-people-with-metadata/

[5] http://webpolicy.org/2013/12/23/metaphone-the-nsas-got-your-number/

http://webpolicy.org/2013/11/13/whats-in-your-metadata/

http://www.commondreams.org/view/2013/12/27#.Ur2o6ykcPr0.twitter

http://www.theatlantic.com/technology/archive/2013/12/stanford-researchers-it-is-trivially-easy-to-match-metadata-to-real-people/282642/

[6] See separate anonymity papers

As with other aspects of internet security and privacy, what is required is not Balkanisation, but better education and user awareness to help protect the naïve from themselves. Key to this is not putting personal information in to websites you do not trust or using computers you do not trust, such as those in a cyber café.

The value of Big Data – transparency and the perceptions of organisations and individuals
The concept at the heart of big data, of sucking up all of the information and data that you can and keeping it as long as you can in case you might need it in the future, is highly dangerous. The very idea of getting informed consent from every individual whose data is included in a big data set is unrealistic, since the data have been hoarded in order to answer questions you never thought of asking when you collected it. This means the organisations that collect and use big data must be transparent, responsible for its use and ethical in their dealings. If they are not they will forgo citizen's trust.

When organisations are transparent in their intention to collect and use big data this can be a win-win situation. The collection of data in free form from individuals can also result in valid scientific work. An example of this is www.patientslikeme.com. This is unambiguously a web site set up by big pharma to collect data on diseases and the effects of drugs on those diseases as a longitudinal study tracking treatments, their efficacy and side effects from a patient perspective. It offers patients something in return for their data. Patients can join fora and chat groups to discuss their conditions and their reactions to different treatments for those conditions in return for inputting their data in a semi-structured form that assists epidemiological analysis, through being linked to extant background coding and related to expert classification systems. The patients can also take the personal logs that they create to their clinicians when they go for appointments as an aide memoire for their personal treatment plan. This provides a valuable research resource that might eventually benefit the patients who input their data.

On the more negative side, in many online situations the citizen does not realise their data is being collected, or fully understand that their data has value and may be passed on to others or aggregated in a big data set. The problem is that once one organisation has a business model that relies on scooping up every scrap of data they can find about potential customers and making money out of it, then, if they are successful, their competitors have to do the same. The value of companies like Facebook and Google is based on the current perceptions of the value of big data, so big data could become a race to the bottom as far as the privacy of individuals is concerned. This is covered below in the data protection and privacy sections.

Information and data today is very cheap to acquire and has little or no variable cost associated with it – only the fixed cost of setting up the system to collect and analyse. This may lead organisations to attach little value to it (which is strange when their company may be valued on the data). When data is not valued there may also be little incentive to secure it properly (security of data is not covered here: see PDGC, Security Top Tips and SCoE papers). However, while an organisation might not value an individual's data set, the individual to whom it pertains usually will.

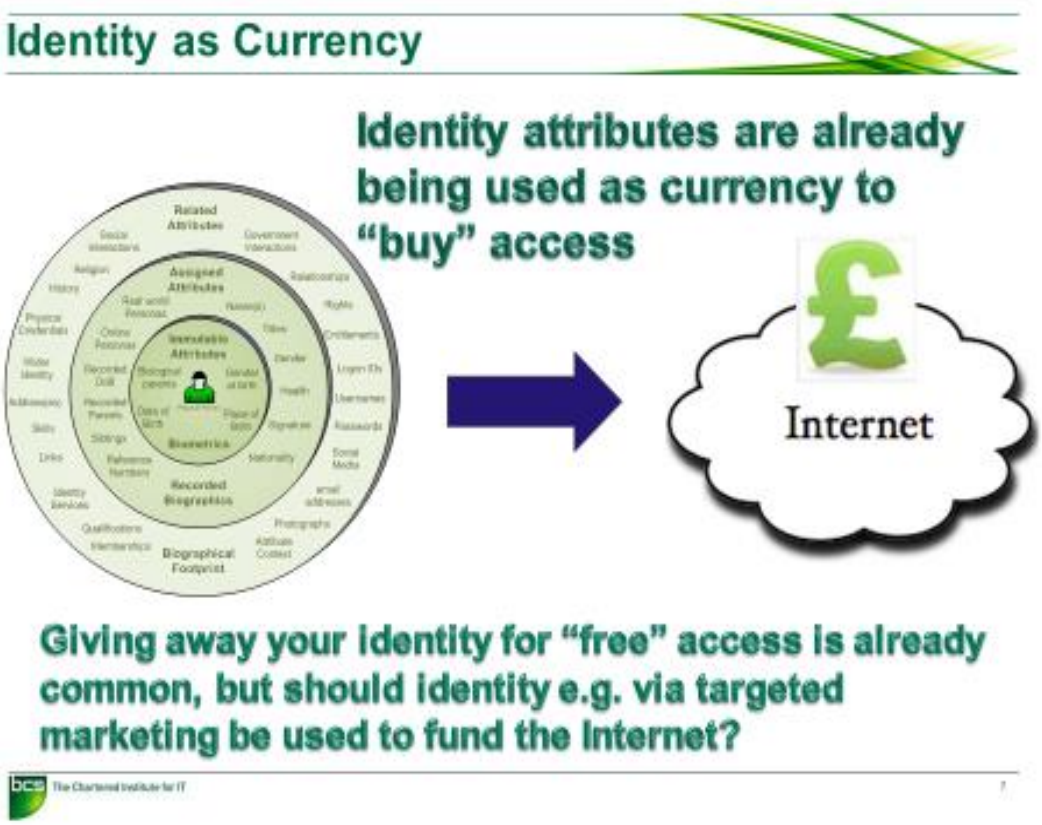
The value of big data – monetisation

There are many models operating in the digital economy. One of the most contentious is the monetising of personal data attributes on the Internet.

We all know that some services are free or below cost. There are many reasons for this. A common one is because a company decides to build market share based on attracting customers to useful or enjoyable free or low cost services. There is value in the data that you, as a customer, give up when you use those sites or services. The quid pro quo is often

targeted advertising, or it may be to sell you “add-ons” to an on-line gaming experience. This is step one in monetisation.

Since organisations are collecting and aggregating our data in this way, it is important that we recognise and accept that truth. However, on the cautionary side, some say: “If you are using a free service you’re not a customer, you’re a product.” So we could call this aggregation of personal data the productisation of people. Personal identity data becomes currency.



On the left hand side of the diagram above, we have the onion rings of data and information that may be associated with an individual. Starting from the inside we have:
“what you are” – your biological attributes – your fingerprints, your face, your voice and so on – these are fairly immutably bound to you. Using these biometrics is pretty good proof that you are linked to your biological identity. Such data is frequently collected for Government IDs and their associated entitlements. Governments need to know you are a citizen of their country.

Then in the next ring we have “what you have” this includes things like your passport and its number. In the on-line context for secure activities, you may use your credit card, an ID card, or a company card. You may be asked to present this token that you have, via a device, to pick up a one-time code for a financial transaction over the internet.

In the outer ring we have less concrete things like “what you know**” and more importantly what is known about you** – your school and social history and your biographical footprint. Knowing these attributes, you can answer questions like: “What is your favourite sport?” to “verify” your ID.

However, lots of people and organisations can find out all about your biographical footprint, maybe they are your friends and family, but it is increasingly easy for strangers,

governments and organisations to find out your biographical data from social networks and by tracing your online history. All the things you have told companies in order to get those freebies, read their magazines, in online searches, or in your tweets will add to your electronic biographical footprint.

This means that strangers, corporations, governments and criminals can discover your identity through data attributes. You can easily and legally observe where someone who does not turn off geo-location tweets from and where a person's mobile phone has been. Then you can deduce: their workplace, their home, their favourite football club, shops, restaurants, their children's schools and so on.

Most of us give this identity attribute information away when we interact on the Internet without even realising it. Companies are storing the data and monetising it to give us superficially free services, be they access to social networks or search company algorithms, or money off vouchers. By and large we want these apparently free or subsidised services and are prepared to put up with the sometimes invasive advertising that may be associated with them.

The Internet is not a free resource, it costs money to build and maintain it. The question is are you happy to help fund the internet and businesses who monetise your data with your personal data and identity attributes?

Data protection and censorship

Data protection and censorship are also inextricably bound up with our views on the ethics of big data analytics and concerns about privacy.

At the start of 2014, the European Commission published a paper on Internet Policy and Governance[7]. This reiterates the EU approach to internet governance summarised in the COMPACT acronym first put forward to the OECD in 2011[8], which builds on the Tunis agenda of 2005. This states that the internet is a space of:

Civic responsibilities,
One unfragmented resource governed via a
Multistakeholder approach to
Promote democracy and Human Rights, based on a sound technological
Architecture that engenders
Confidence and facilitates a
Transparent governance both of the underlying Internet infrastructure and of the services which run on top of it.

BCS endorses the principles given in COMPACT. However, the EU agenda remains one where data protection, privacy and human rights (as enshrined in the European Convention on Human Rights) are of paramount importance. This emphasis on freedom from harm can and does bring the EU in conflict with the cultural norms of other countries and societies. It is at the heart of a long-running difference of approach between the EU and the USA where, as Isiah Berlin said in an essay, freedom "to" trumps freedom "from". This manifests itself most

[7] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions COM(2014)72/4 – Internet Policy and Governance – Europe's role in shaping the future of Internet Governance

[8] Presented at the occasion of the OECD's High-Level Meeting on the Internet Economy, 28.06.2011,

http://ec.europa.eu/commission_2010-2014/kroes/en/blog/i-propose-a-compact-for-the-internet

obviously in the American view that, on the internet is more important to protect freedom of speech than privacy.

While those living in the UK and other parts of Europe enjoy strong data protection legislation, this is not the case in many other parts of the world. This is important to keep in mind, for example when visiting websites, since sites that appear to be in the UK can often be operated from other locations, which have cheaper resources, but much weaker legislative protection and security. Until people start reading the privacy statements and refusing to use services operated from countries that do not have effective data protection they will continue to put themselves at risk.

Data protection is fundamental to the success of commerce on the internet. If people lose trust in online services they will stop using them. There have already been a number of online companies that have failed as a result of a data breach. Thus data protection is not just a right, it is also a business driver and something that can make or break an online business.

However, the EU view has also resulted at times in accusations from much of the developing world that EU data protection legislation is simply a means of putting up a trade barrier rather than protecting a fundamental right to privacy. The Western view of privacy and data protection is also at variance with the norms of censorship in totalitarian regimes.

It is also possible to take data protection too far. In some instances this has resulted in inadvertent censorship. In South Korea, the government instituted the requirement that people logging on to certain services had to have an online ID, which was only given to people living in South Korea. This effectively prevented all Koreans living in other countries from accessing these sites. As some of these were discussion groups and news sites, this basically resulted in censorship of those sites from Korean speakers outside Korea. Happily the Korean Constitutional Court ruled this unconstitutional in 2012.

There are other examples where implementing secure identity has – possibly unintendedly - resulted in censorship. Wherever a government or organisation implements access controls to services, they need to ensure that all of those who could benefit from the service or should have access are able to obtain credentials. However, the related issue of remote registration of identity remains an unresolved topic.

Privacy/ surveillance

In order to understand the potential impact of big data on privacy, it is necessary to give historical, legal and cultural context to privacy and surveillance as it pertains in the UK today.

Privacy is a fundamental human right. It underpins human dignity and other values such as freedom of association and freedom of speech. It has arguably become one of the most important human rights of the modern age. However, of all the human rights in the international catalogue, privacy is perhaps the most difficult to define. Definitions vary widely according to culture, context and environment, but the most common aspects to be enshrined in law are bodily, territorial, information and communication privacy. In many countries, the concept has been fused with data protection, which interprets privacy in terms of management of personal information.

The modern concept of a right to privacy originated in the United States of America and is an implicit right guaranteed by the Constitution as interpreted by the US Supreme Court. A

prominent US attorney Dean Prosser argued that 'privacy' was composed of four separate torts[9], the only unifying element of which is a 'right to be left alone.' These torts were:

- Appropriating the plaintiff's identity for the defendant's benefit.
- Placing the plaintiff in a false light in the public eye.
- Publicly disclosing private facts about the plaintiff.
- Unreasonably intruding upon the seclusion or solitude of the plaintiff.

In English law there is no independent tort law doctrine that recognises a right to privacy. Privacy in English law focuses on the right of an individual to protect personal information from misuse or unauthorised disclosure. This area of law has been transformed by the Human Rights Act 1998, which incorporated the European Convention on Human Rights (ECHR)[10] into English law. Article 8 of the Convention guarantees the right to respect for private and family life:

- Everyone has the right to respect for his private and family life, his home and his correspondence (8.1).
- There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others (8.2).

As a signatory to the ECHR, the UK must adhere to Article 8 of the ECHR, subject to restrictions as prescribed by law and necessary in a democratic society. The Convention also requires the judiciary to 'have regard' to the Convention in developing the common law. However, there is no common law right to privacy in English law. This may be compared to the jurisprudence of the Supreme Courts of Europe, the USA and Australia, which, in modern times, have adopted varying and somewhat broader interpretations of the right to privacy.

In most countries, privacy has been developed primarily around data protection rather than the broader concept of privacy. In 1980, in an effort to create a comprehensive data protection system, the Organisation for Economic Cooperation and Development (OECD) issued its 'Recommendations of the Council Concerning Guidelines Governing the Protection of Privacy and Trans-Border Flows of Personal Data.' The seven principles governing the OECD's recommendations for protection of personal data were:

1. Notice - data subjects should be given notice when their data is being collected;
2. Purpose - data should only be used for the purpose stated and not for any other purposes;
3. Consent - data should not be disclosed without the data subject's consent;
4. Security - collected data should be kept secure from any potential abuses;
5. Disclosure - data subjects should be informed as to who is collecting their data;
6. Access - data subjects should be allowed to access their data and make corrections to any inaccurate data; and

[9] Tort law is a body of law that addresses, and provides remedies for, civil wrongs not arising out of contractual obligations. A person who suffers legal damages may be able to use tort law to receive compensation from someone who is legally responsible, or liable, for those injuries. Generally speaking, tort law defines what constitutes a legal injury and establishes the circumstances under which one person may be held liable for another's injury. Torts cover intentional acts and accidents. In contrast to criminal law (in which the offense is against the State and the State is the plaintiff), in tort law, the offense is against a person and that person is the plaintiff.

[10] The European Convention on Human Rights and Fundamental Freedoms was adopted under the auspices of the Council of Europe on 4th November 1950.

7. Accountability - data subjects should have a method available to them to hold data collectors accountable for following the above principles.

The OECD Guidelines were non-binding and despite the fact that European countries developed their data protection laws in accordance with the Council of Europe Convention, the laws in force in individual countries vary significantly in scope. The European Commission realised that diverging data protection legislation in the EU member states would impede the free flow of data within the EU zone. Therefore the European Commission decided to harmonise data protection regulation and promulgated the Data Protection Directive 95/46/EC on the protection of personal data.

The Data Protection Act 1998 (DPA)[11] is the principal legislation that governs the protection of personal data in the UK. Although the Act does not mention personal privacy, in practice it provides a way in which individuals can control information about themselves. Anyone holding personal data is legally obliged to comply with this Act, subject to some exemptions.

While the UK adheres to the provisions of Article 8 of the ECHR, since 1999 it has been subject to twice the average number of judgements by the European Court on Human Rights on issues relating to privacy. The UK, like other signatories of the Convention, has progressively introduced enabling legislation that, while supporting public safety or the economic well-being of the country, may be considered to challenge the underlying principle of Article 8(1).

The challenge often arises from differences in interpretation of the need for proportionality in the application of the legislation. Proportionality is a fundamental principle of EU law, and is mentioned explicitly in the Treaty on the European Union. It is also a basic principle of human rights law under the European Convention on Human Rights, and the European Court of Human Rights routinely uses proportionality as a criterion in determining whether data processing is legal. The potential undervaluing of proportionality needs to be balanced by giving explicit guidance to those making decisions on privacy versus use of personal data.

Much of this legislation has been introduced directly or indirectly in response to incidents of high profile crime and/or terrorism, e.g. New York bombings of 2001, Soham murders in 2002, London bombings of 2005, etc. This has resulted in an increased focus on identity, sharing of personal information and surveillance. Examples of the accumulation of legislation could be seen in relation to Acts such as the Regulation of Investigatory Powers (RIPA) and databases produced to assist the protection of children (e.g. ContactPoint). The anomalies this produced have subsequently been altered and reduced by the Protection of Freedoms Bill.

The widespread use of surveillance technologies, CCTV, ANPR[12], Mobile Phone Tracking etc., provides substantial forensic capabilities in fighting crime and terrorism. However this generally involves the collection and storage of all related data to enable authorised forensic investigation. The UK is frequently said to be the most observed society in the world with arguably more CCTV cameras per head of population than any other country. Yet, the UK citizen generally accepts public and private CCTV surveillance as a source of reassurance. This is not so in many other countries where the default position is that CCTV monitoring is perceived as an intrusion on civil liberties and an invasion of privacy. Is the UK's attitude to surveillance based on enlightenment or apathy? Other countries frequently cite retention of so much personal data to aid law enforcement as excessive. It is important to reach the right

[11] The 1998 Act replaced the Data Protection Act 1984, and was intended to bring UK law into line with the EC Data Protection Directive 95/46/EC.
[12] Automatic Number Plate Recognition

balance between our willingness to support the common good and the freedoms, including our right to privacy, on which our society is founded.

The quest for greater effectiveness and efficiency, increased citizen engagement and the economic well-being of the country means that the perceived need to share personal information through the rise of big data has increased throughout the public and private sectors. While the collection and sharing of an individual's personal data has been widespread for generations, technology has enabled that collection, analysis and sharing to occur on a massive scale. The business advantages offered by big data are breaching both organisational barriers and established data governance mechanisms, obscuring end-to-end accountability. Technology has also enabled a paradigm shift in surveillance from a "one to many" to a "many to many" on a previously unimagined scale. Commercial and security pressures are increasing the business need to analyse and monetise big data and effective data governance has been left behind.

If we consider the surveillance aspects of big data (as distinct from CCTV surveillance), the belief that you are being watched influences behaviour, so if big data develops so that people are consciously aware data is being collected about them and this becomes associated with surveillance, they may lie or seek to be anonymous in many more situations. In addition, the more that data is mined in situations where individuals have not given consent (consciously) and then find that their data is being used, the more likely it is that data quality will reduce

The Value of big data – humanitarian use

Big Data analysis over the Internet is being carried out increasingly by International NGOs and the private sector as well as governments. BCS has warned that identity discovery through data aggregation is something that Internet users need to be aware of and they should consent to the use of their personal data online. Ethical scrutiny of both the collection and analysis of massive open data online is needed so that it can be used for social good. Since it is both cheap and possible people and organisations are going to do it. It is important to control the analytical risks and not forgo the opportunities to use such data for humanitarian reasons and individual benefits. These include such things as: using mobile phone geo-location to track victims of disasters, as in Haiti and after Hurricane Sandy, or using Internet performance monitoring data to identify such things as patterns of censorship in repressive regimes, like Iran.

Snowden is in danger of diverting everyone's attention away from what NGOs and commercial organisations are doing with Big Data and focussing attention solely on spying (which after all is what spies do).

Debate is urgently needed on the ethical issues on the collection, "anonymisation" and analysis of big data online.

Quality fitness for purpose

In the digital world "to be" is "to be recorded". However, we need to ask ourselves when big data is good enough to be acted upon. People are not just their textual or visual traces online. Big data sets may be dangerously skewed, both in terms of their representation of society (as with any opinion poll), since big data is largely individually posted and self-selecting, and in terms of the representation of those individuals in the data set.

Freeform data, generated without data definitions, is unlikely to be of high quality and consistency. This may not matter in some situations, such as getting a rough idea of what

people think about a new shopping centre or a conference talk from their tweets in response to a given hash tag. However, people often neglect to consider the self-selecting set of people making these responses, or that they are dashing them off with “finger trouble” and without deep thought. It matters more if the purpose of the analysis is to create a government policy or if one data set is aggregated with another inappropriately. The honesty, accuracy and completeness of what even the most conscientious individual puts into a data set, when it is not being used in the way the individual intended or anticipated when they posted their data, has to be suspect.

To understand the ramifications of this, we just have to look at the mistake of Target, who, on analysing a woman’s shopping basket, predicted that she was pregnant and helpfully sent her “relevant” marketing emails, only to find they had discovered her pregnancy before she had, and to have the emails discovered by the woman’s father.

Big data used as a precursor to actions and a predictor

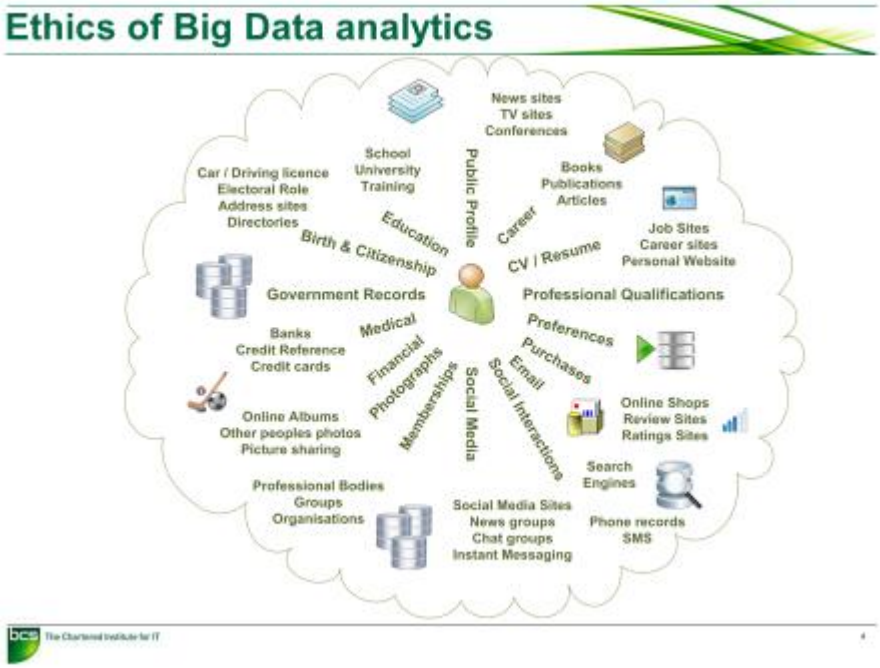
In the world of analysis of big data for marketing, the intention is often to pick up on the antecedent indicators to financial transactions (as in the anticipation of pregnancy spending above). Most commonly this is to identify individuals who will buy this or that by correlating the characteristics of those who have made a purchase with new individuals for whom you have data. This is always dangerous as correlation is not the same as causation. The classic exemplar of this is: school grades have reduced, alcohol consumption has gone up; ergo the drop in school grades is a result of drunk and hung over teachers. More broadly, this means that the data collected are used to put people into groups who are predicted to behave in certain ways. This channels people into becoming certain types of consumers. These predictive groups are often sub-divided with greater and greater granularity until logically they become groups of tens or even the construction of single individuals that are predicted to behave in certain ways.

Often a firm will be pleased that their marketing efforts based on big data analysis have increased take up of an offer or new product by say 1% of those targeted. However, they fail to realise that 99% of people were probably just irritated by the random email or text message that was of no interest to them. If this happens repeatedly it can become a disincentive to purchase the product from the supplier.

The consumer example may seem trivial and not to involve ethical considerations. However, this type of big data analysis is the basis of the government identifying groups (and therefore the individuals that belong to such groups) who might commit fraud in the tax and social security areas. Those groups of individuals and businesses are then targeted for special scrutiny by HMRC or DWP. In the same way, the banks use big data analysis to identify the likelihood of a credit or debit card being used fraudulently, because the purchase is outside an individual’s normal spending habits. While many people are reassured by this, it can be embarrassing if the unusual and one off jewellery purchase is an engagement ring and it is queried or blocked.

Some people have even suggested that big data can identify people who are likely to commit a criminal act, before they do so. If that potential criminal act is very serious, like murder, child abuse or terrorism, to what extent should it be acted upon and the individual apprehended and locked up before they commit the crime? We are on the cusp of doing this based solely on big data analysis without specific cause in child protection and terrorism cases.

Conclusions



Big Data analysis over the Internet is a growing activity raising concern. This is being carried out by both Governments and the private sector. BCS has warned that identity discovery through data aggregation is something that Internet users need to be aware of and they should consent to the use of their personal data online.

Ethical scrutiny of both the collection and analysis of massive open data online is needed so that it can be used for social good. Because it is possible it is not necessarily ethical.It is important to control these risks and not forgo the opportunities to use such data for humanitarian reasons (such as to track victims of disaster as in Haiti, or use Internet performance monitoring data to identify such things as patterns of censorship in repressive regimes).

UK legislators should lead discussion on the ethical issues on the collection, “anonymisation” and analysis of big data online. There are problems deciding who is responsible for shared data: governments, telcos, NGOs, ISPs, Ap developers, in all contexts, even humanitarian relief. Some key issues are data granularity and the retention of metadata.

The privacy issues surrounding data collection and analytics are enormous and require a rational, unemotional debate about when societal good outweighs personal privacy.

The BCS position on an ethical approach to the use of big data is summarised as follows:

- Transparency is the key to the ethical use of big data. BCS will press for individuals to be made aware and give their consent when data related to them is being collected, aggregated and analysed. BCS will push government, corporations and the not for profit sector to be open about their collection, use and monetisation of big data. The OECD‘s seven principles for the protection of personal data (1980) on which the Data Protection Act 1998 was based remain core to the necessary transparency in relation to big data. These are:
    1. Notice - data subjects should be given notice when their data is being collected;
    2. Purpose - data should only be used for the purpose stated and not for any other purposes;

3. Consent - data should not be disclosed without the data subject's consent;
4. Security - collected data should be kept secure from any potential abuses;
5. Disclosure - data subjects should be informed as to who is collecting their data;
6. Access - data subjects should be allowed to access their data and make corrections to any inaccurate data; and
7. Accountability - data subjects should have a method available to them to hold data collectors accountable for following the above principles.

• BCS believes that analysts and policy makers must understand the limitations associated with the use of massive largely unstructured data sources and ensure that they derive evidence based policies from them in a way that is both scientifically and statistically correct, fair and ethical to contributors and non-contributors to those databases alike. This includes ensuring they engage with all demographic groups to understand their views on the consumption of Big Data by both businesses and government. They also need to consider the dangers related to the use of Big Data for the purpose of predicting individual actions;
• We feel that more research needs to be undertaken to understand better how mass observation and constant surveillance affects the quality of data and whether certain types of policy and service decisions are better made with smaller rather than bigger quantities of data;
• Users of Big Data must be made aware of the likelihood that attribute combination will de-anonymise Big Data sets in whole or in part and that when such data sets are de-anonymised, they becomes personal data and must be treated as required by the Data Protection Act. This is particularly important where sensitive data are concerned;
• BCS endorses the push by governments and business to grow the online economy and ensure that the UK is in the forefront of deriving benefits from the online world.
• BCS will push for ethical scrutiny of the collection and analysis of massive open data online so that it can be used for social good, in a way that is proportionate to any potential individual privacy concerns;
• BCS will strive to educate the general public about the value of their personal data and how it is or may be used or abused by users and organisations, so that the public can make their own personal decisions about the balance of risks and rewards for themselves.
• BCS supports the view that the Internet should remain an open platform available to all and will resist attempts at Balkanisation;
• BCS will promote the European Union and OECD approach to Internet governance encapsulated in the COMPACT acronym, which states that the Internet is a space of:
o Civic responsibilities,
o One un-fragmented resource governed via a
o Multi-stakeholder approach to
o Promote democracy and Human Rights, based on a sound technological
o Architecture that engenders
o Confidence and facilitates a
o Transparent governance both of the underlying Internet infrastructure and of the services which run on top of it.