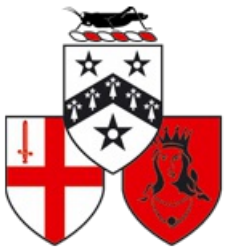

GRESHAM COLLEGE

Founded 1597

Big Data: The Broken Promise of Anonymisation Transcript

Date: Tuesday, 14 June 2016 - 6:00PM

Location: Museum of London

14 June 2016

Big Data:
The Broken Promise of Anonymisation

Professor Martyn Thomas

Big Data

Big Data is [i] one of the *Eight Great Technologies* identified by the UK Government as underpinning industrial strategy [ii] . (There are now nine – they forgot quantum technology).

https://s3-eu-west-1.amazonaws.com/content.gresham.ac.uk/data/library/14jun16MartynThomas_001.jpg

IBM estimates that 2,500,000,000 Gigabytes of data are created every day and that over 90% of all the data in the world were created in the last 2 years [iii] . Big data arises in many forms and from many sources, text, video, phone data, equipment monitoring, photographs, audio, store transactions, health monitors … almost everything that happens creates some data somewhere and much of it gets transmitted, copied and stored.

Computer processing and storage have advanced to the point where previously unimaginable volumes of data can be processed immediately or stored for later analysis. Twitter users create 400 million tweets each day and some organisations buy access to all of these tweets and process them in real time to extract information. A credit card transaction will create about 70 items of data, to identify the customer, the credit card, the goods purchased, the retailer, the time, location, whether the PIN was input, the currency, the tax amounts, transaction codes and identifiers. If the transaction is a purchase from an online shop, much more data will be created, recording for example

- all the website pages that were visited and how long was spent on each;
- the previous website visited and the where the user goes next;
- the browser version, operating system version and computer details;
- the user's IP address, location and internet service provider;
- previous access to any of the websites hosted on any of the pages visited (by checking stored cookies);
- advertisements viewed;
- any "likes" or other sentiment indicators;
- whether any social media sites were being viewed or postings being made;
- .. and potentially much more.

Visa alone handled 128 billion purchase transactions in 2015 [iv] . The digital data trail that results from all of our activities is commercially valuable, so it will often be stored for ever and processed many times for different purposes. This means that commercial companies hold a large amount of data about each of us. When an Austrian law student called Max Schrems insisted that Facebook send him all the data they held about him they initially resisted but, after a court decision, they sent him a CD containing a 1200 page PDF. This showed all the items on his newsfeed, all the photos and pages he had ever clicked on or liked, all the friends he could see and all the advertising that he had ever viewed [v] .

Some data are in no sense personal data. The Large Hadron Collider at CERN generates about 30 million gigabytes of data each year [vi] and none of it says anything about an individual human being. But much of the data that is generated and processed in the world is about individual people and their activities and could reveal things about individuals that they prefer or need to keep private.

This creates a conflict of interests, because data can have great value to organisations and to societies. For example, medical records are very valuable for research into the patterns and possible causes of illnesses; they are essential for individual healthcare and to manage health services efficiently. But medical records can also reveal highly personal information that could lead to discrimination in employment, to intrusive marketing, to family breakdowns, to risks of violence or to deportation.

As another example, phone companies need records that show what phone calls and texts were sent and received so that they can charge for services and plan and manage their networks; these same records are used by the police to discover who was in the neighbourhood of a crime. But phone records could also be analysed to reveal highly personal information such as who is attending a drug rehabilitation clinic, or who are spending the night together, and how often and where.

Governments and other regulators have attempted to resolve this conflict between personal privacy, commercial interests and the processing of data to provide vital services. The strengths and weakness of the strategies that they have used are the subject of this lecture but, first, we need to explore the right to privacy and just what is meant by identifiable *personal data.*

Privacy and Personal Data

Article 12 of the Universal Declaration of Human Rights states:

"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" ***[vii]***

Privacy is important. It is sometimes said that if you have nothing to hide, then you have nothing to fear from your personal information being made public [viii] , with at least a slight implication that people who care about their privacy must have done something they are ashamed about. But, most people prefer to choose what personal information they share and whom they share it with and for some people privacy may be very important indeed – it can even be a matter of life and death.

Attitudes to privacy differ between cultures and age groups, but most people would like to retain some control over what information is collected about them and who has access to that information. We may have become used to carrying a tracking device [ix] with us wherever we go (rather as if we were electronically-tagged criminals) but few of us would welcome a live-streaming videocamera in our bedroom or bathroom (although that is what some parents and care-homes have installed [x]) and most people draw their curtains.

For a significant number of people, their privacy can be vitally important to their wellbeing or to their physical safety and that of their families, for example

· People who have suffered some trauma in their lives.

· People with spent criminal convictions [xi] .

· Children who have been taken into care and adopted and who may be at risk from their birth relatives.

· People escaping abusive relationships, who may be at risk from their former partners.

· Witnesses in criminal trials who may need protection.

· Anyone whose lawful actions would nevertheless be considered unacceptable in their culture, religion or family.

It is important to remember that data is persistent and that laws and attitudes change over time, so data that was once harmless can become a threat in future when one's personal circumstances change (for example, by becoming a celebrity or an adoptive parent) or when social norms change or one moves to (or even just visits) a country that has very different laws and culture. To dismiss concerns about privacy as unimportant, as some politicians do, is either ignorant or callous.

Data may be very valuable. Much of the stock market value of Google, Twitter and Facebook reflects the perceived commercial value of the data that they control. Clive Humby [xii] , of Tesco Clubcard fame, described data as "the new oil" at a conference at Kellogg School that was reported by Michael Palmer [xiii] .

Data is just like crude. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analyzed for it to have value. The issue is how do we marketers deal with the massive amounts of data that are available to us? How can we change this crude into a valuable commodity – the insight we need to make actionable decisions?

Sometimes, the conflict of interest between extracting the value from data and respecting privacy can be resolved by anonymising the data so that they can be shared and used without any breach of privacy. This is easiest if the data can be aggregated so that all individual data are lost in the aggregate. Unfortunately as we shall see, anonymisation can be very difficult or impossible if the data contain several facts about one individual, and lawmakers and public understanding have not kept up with the developments in data science. As a consequence, it has become difficult to say who owns the data that is collected about us and how they can be used legally and ethically. In this lecture, I shall ignore the minefield of *informed consent*, which is something we shall discuss in some detail in my next lecture on 18 October.

The UK Data Protection Act [xiv] (UK DPA) defines *personal data* as follows

"personal data" means data which relate to a living individual who can be identified—

(a) from those data, or

(b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller,

and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual;

The *data controller* is defined as *a person who (either alone or jointly or in common with other persons) determines the purposes for which and the manner in which any personal data are, or are to be, processed.* ***[xv]***

This is a significantly narrower definition of *personal data* than the one used in the EU Data Protection Directive [xvi] which states

"Personal data shall mean any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity".

What the Directive means by *an identifiable person* has been clarified by the Article 29 Data Protection Working Party (A29 WP) that was set up by the EU to support the Directive [xvii] . The A29 WP explains that account must be taken of all means that are reasonably likely to be used to identify the data subject, either by the data controller **or by any other person** at any time before the data is destroyed.

Recital 26 of the Directive pays particular attention to the term "identifiable" when it reads that "whereas to determine whether a person is identifiable account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person." This means that a mere hypothetical possibility to single out the individual is not enough to consider the person as "identifiable". If, taking into account "all the means likely reasonably to be used by the controller or any other person", that possibility does not exist or is negligible, the person should not be considered as "identifiable", and the information would not be considered as "personal data". The criterion of "all the means likely reasonably to be used either by the controller or by any other person" should in particular take into account all the factors at stake. The cost of conducting identification is one factor, but not the only one. The intended purpose, the way the processing is structured, the advantage expected by the controller, the interests at stake for the individuals, as well as the risk of organisational dysfunctions (e.g. breaches of confidentiality duties) and technical failures should all be taken into account. On the other hand, this test is a dynamic one and should consider the state of the art in technology at the time of the processing and the possibilities for development during the period for which the data will be processed. Identification may not be possible today with all the means likely reasonably to be used today. If the data are intended to be stored for one month, identification may not be anticipated to be possible during the "lifetime" of the information, and they should not be considered as personal data. However, it they are intended to be kept for 10 years, the controller should consider the possibility of identification that may occur also in the ninth year of their lifetime, and which may make them personal data at that moment. The system should be able to adapt to these developments as they happen, and to incorporate then the appropriate technical and organisational measures in due course. ***[xviii]***

The omission of *or by any other person* and *during the lifetime of the information* from the UK DPA means, for example, that according to the UK DPA (although not under European Law) a data controller need not treat records as personal data if they have been edited in a way that means that the data controller can no longer identify the individual person *even if it may be trivial for others to identify the person by using additional data held by them.* As we shall see shortly, it takes surprisingly little additional data to be able to re-identify individuals in detailed datasets that have been "anonymised" in the way this is usually done.

This difference between UK and EU law is unlikely to survive the introduction of the General Data Protection Regulation [xix] (GDPR) in 2018. Because this is a Regulation rather than a Directive, it will have legal force in every EU state without the need for national legislation (although several sections of the GDPR allow national legislation to modify the default legal positions under the GDPR, for example the age below which it is necessary to get parental agreement before processing the personal data of a child). The GDPR will be binding on all organisations inside *and outside* the EU that have a presence in the EU and that process the personal data of EU citizens. Any organisation that has been relying on the narrow definition of personal data in the UK DPA will need to review its processing of any data that contain details of EU citizens.

The GDPR introduces far-reaching and significant changes that are too big a subject to be covered as part of this lecture. I expect that Gresham College will devote a full lecture to the GDPR in 2018, when the details of the transposition into UK law will have been clarified. Meanwhile the UK Information Commissioner's Office (ICO) has issued a guide [xx] to the 12 steps that organisations should be taking now to prepare for the GDPR.

Open Data

The value of data can sometimes be maximised by making it available for anyone to use. I expect that most of us use one or more transport apps on our phones to find out the best routes, train timetables and fares, and when the next bus will arrive. These apps rely on data feeds from open data sources provided by the transport operators. To make the data easier to use, a company transport *api* ***[xxi]*** has consolidated as many data feeds as possible into one programming interface and they say they have over 1500 developers and organisations taking their data feeds and using them in their products and services. This is just one example of the power that open data has to stimulate innovation.

Sir Tim Berners-Lee (inventor of the world-wide web) and Sir Nigel Shadbolt (data scientist and Principal of Jesus College, Oxford) founded the Open Data Institute [xxii] to promote the use of open data. Their definition of *open data* is that it is Open data is *data that anyone can access, use and share*. To meet their definition, open data has to have a licence that says it is open data because, without a licence, the data could not legally be reused. The licence might also say that people who use the data must credit whoever is publishing it or that people who mix the data with other data have to also release the results as open data.

For example, the UK Department for Education (DfE) makes available open data about the performance of schools in England. The data is available as CSV [xxiii] and is available under the Open Government Licence (OGL) [xxiv] , which only requires re-users to say that they obtained the data from the Department for Education.

The DfE schools data is far from the only Government dataset that has been released under the OGL. The website https://data.gov.uk/ lists 22,732 OGL datasets (as of 6 June 2016) and about 10,000 others that have different availability. It is an extraordinary resource. A few example datasets are

· detailed road safety data about the circumstances of personal injury road accidents in GB from 1979 including the types (including Make and Model) of vehicles involved;

· all Active MOT Vehicle Testing Stations in England, Scotland and Wales including addresses, contact numbers and test classes authorised;

· planned roadworks carried out on the Highways Agency network;

· hourly observations for approximately 150 UK observing stations, daily site specific and 3 hourly site specific forecasts for approximately 5000 UK locations [xxv] ;

· National Statistics Postcode Lookup (NSPL) for the United Kingdom;

· all unclaimed estates held by the Bona Vacantia Division [xxvi] which are both newly advertised and historic;

· All MOT tests and outcomes, including make and model of vehicle, odometer reading and reasons for failure, since the MOT system was computerised in 2005

· the Accident and Emergency (A&E) Attendance data within Hospital Episodes Statistics (HES). It draws on over 18 million detailed records per year.

· … and there are tens of thousands more!

As Open Data becomes more and more widely used, each dataset becomes a single point of failure for all the services that depend on it. Sir Nigel Shadbolt has said that open data is so important that it has become part of the country's critical national infrastructure [xxvii] . Because the ownership of the more than 32,000 datasets is spread across very many organisations and because anyone can use the 22,000 OGL datasets, no-one can have oversight of the dependencies that are accumulating and no-one can have overall responsibility for ensuring that the data remains available and that it has not been altered for criminal purposes. The implications of this for another widely-used and freely available data source, the GPS satellite signal have been described in detail in a Royal Academy of Engineering report [xxviii] on the widespread dependence on Global Navigation Space Systems and the extraordinary vulnerabilities that are the result.

Some Big Data sources suffer from *selection bias* in that the data may not accurately represent the underlying population. Health data, for example, will under represent the healthy if it is sourced from hospital and GP records, and data from on-line activity will not represent the 25% of the population who have little or no online presence (and who tend to be poorer, older and less healthy that the overall population. Data analytics have

great value when the right data can be analysed in the right way, but selection bias may lead to the wrong policy decisions if it is not recognised and controlled.

Some commercial companies deliberately release data that they collect, for many reasons. I shall shortly show examples of how this can go badly wrong.

Anonymisation

We have seen that international Human Rights law, EU law and UK law all require that personally identifiable data must be protected, although they have different definitions of what makes personal data identifiable.

Privacy is a fundamental human right recognized in the UN Declaration of Human Rights, the International Covenant on Civil and Political Rights and in many other international and regional treaties. Privacy underpins human dignity and other key values such as freedom of association and freedom of speech. It has become one of the most important human rights issues of the modern age ***[xxix]*** .

We have also seen that the collection, analysis and sharing of personal data can provide great benefits to society (through better health care, more effective law enforcement and improved efficiency, for example) and to businesses and other organisations. The pressure to share and use personal data whilst preserving privacy has led many organisations to anonymise the data they collect, hoping and expecting to be able to get the benefits from the data without creating harm.

Anonymisation of data about an individual involves removing the details that identify the person whilst preserving the details that are needed so that the data is still useful. This is harder than it may seem, because it takes surprisingly few details to identify someone with a high degree of probability. In America, 61% of the population in 1990 and 62% in 2000 were uniquely identified by their approximate address (their ZIP code), their gender and their birth date [xxx] . The same is probably true in the UK if you replace the American ZIP code with the first three characters of the UK postcode.

Paul Ohm, Associate Professor at the University of Colorado Law School, explains the anonymisation challenge succinctly in the introduction to his seminal paper on *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization* ***[xxxi]*** .

Imagine a database packed with sensitive information about many people. Perhaps this database helps a hospital track its patients, a school its students, or a bank its customers. Now imagine that the office that maintains this database needs to place it in long-term storage or disclose it to a third party without compromising the privacy of the people tracked. To eliminate the privacy risk, the office will anonymize the data, consistent with contemporary, ubiquitous data-handling practices.

First, it will delete personal identifiers like names and social security numbers. Second, it will modify other categories of information that act like identifiers in the particular context—the hospital will delete the names of next of kin, the school will excise student ID numbers, and the bank will obscure account numbers.

What will remain is a best-of-both-worlds compromise: Analysts will still find the data useful, but unscrupulous marketers and malevolent identity thieves will find it impossible to identify the people tracked. Anonymization will calm regulators and keep critics at bay. Society will be able to turn its collective attention to other problems because technology will have solved this one. Anonymization ensures privacy.

Unfortunately, this rosy conclusion vastly overstates the power of anonymization. Clever adversaries can often reidentify or deanonymize the people hidden in an anonymized database.

He concludes that *Data can be either useful or perfectly anonymous but never both.*

The following examples are discussed in his paper, which I strongly recommend as an authoritative source for analysis and references about anonymisation. The examples are American but *mutatis mutandis* they transpose with ease to other countries.

Example 1: AOL Search Data

In August 2006, AOL announced that they were releasing some data to help build an open research community.

https://s3-eu-west-1.amazonaws.com/content.gresham.ac.uk/data/library/14jun16MartynThomas_002.jpg

AOL had tried to anonymise the data they released by removing the searcher's IP address and replacing the AOL username with a unique random identifier linking of the searches by any individual, so that the data was still useful for research.

It did not take long for two journalists to identify user 4417749, who had searched for people with the last name Arnold, "homes sold in shadow lake subdivision gwinnett county georgia" and "pine straw in lilburn ga." as Thelma Arnold, a widow living in Lilburn, Georgia. AOL were strongly condemned [xxxii] for releasing the data and apologised, calling it a "screw up" but claiming that "there was no personally identifiable information linked to these accounts" [xxxiii] .

https://s3-eu-west-1.amazonaws.com/content.gresham.ac.uk/data/library/14jun16MartynThomas_003.jpg

AOL took the database offline but it was too late: the internet never forgets. Many mirror sites had already been set up and the data is still online today at http://www.not-secret.com/, where you can see (by searching for 4417749) the revealing picture of Thelma Arnold's life that her searches provide.

Other users' searches were similarly revealing; user 114037 seems to have been contemplating wife swapping. Another user has murder on their mind:

https://s3-eu-west-1.amazonaws.com/content.gresham.ac.uk/data/library/14jun16MartynThomas_004.jpg

Maybe user 17556639 was a writer looking for plot details.

What picture do your searches paint? Or are you careful to use a privacy sensitive search engine such as Startpage [xxxiv] or DuckDuckGo [xxxv] for all your searches? Explore the AOL data using the interface at http://www.not-secret.com/ and you will never feel the same about search engines that retain your search history.

Example 2: The Netflix Prize

Netflix is a movie rental company that benefits from its ability to recommend films that subscribers would like to watch, based on their previous ratings. In October 2006, Netflix launched a $1m prize for an algorithm that was 10% better than its existing algorithm Cinematch [xxxvi] . Upon registration, participants were given access to the contest training data set of more than 100 million ratings from over 480 thousand randomly-chosen, anonymous customers on nearly 18 thousand movie titles. The data were collected between October, 1998 and December, 2005 and reflected the distribution of all ratings received by Netflix during this period. The ratings were on a scale from 1 to 5 (integral) stars. Netflix stated that *to protect customer privacy, all personal information identifying individual customers has been removed and all customer ids have been replaced by randomly-assigned ids. The date of each rating and the title and year of release for each movie are provided. No other customer or movie information is provided* .

Two weeks after the prize was launched, Arvind Narayanan and Vitaly Shmatikov of the University of Texas at Austin announced that they could identify a high proportion of the 480,000 subscribers in the training data. Their paper [xxxvii] detailing their methods and results was published in 2008. The 10-page paper is fascinating and I recommend it strongly. The conclusions would surprise most people.

In the case of the Netflix movie ratings dataset, the attacker may already know a little bit about some subscriber's movie preferences — the titles of a few of the movies that this subscriber watched, whether she liked them or not, maybe even approximate dates when she watched them. Anonymity of the Netflix dataset thus depends on the answer to the following question:

How much does the attacker need to know about a Netflix subscriber in order to identify her record in the dataset, and thus completely learn her movie viewing history?

In the rest of this paper, we investigate this question. In brief, the answer is: very little. For example, suppose the attacker learns a few random ratings and the corresponding dates for some subscriber. We expect that the dates when the ratings are entered into the Netflix system are strongly correlated with the dates when the

subscriber actually watched the movies, and can thus be inferred by the attacker. To account for the imprecision of date knowledge, in our analysis the attacker's knowledge of the dates has either a 3-day, or 14-day error.

With 8 movie ratings (of which we allow 2 to be completely wrong) and dates that may have a 3-day error, 96% of Netflix subscribers whose records have been released can be uniquely identified in the dataset. For 64% of subscribers, knowledge of only 2 ratings and dates is sufficient for complete deanonymization, and for 89%, 2 ratings and dates are enough to reduce the set of plausible records to 8 out of almost 500,000,which can then be inspected by a human for further deanonymization. Less popular movies are extremely helpful. For example, even with a 14-day error in the dates, approximate knowledge of 8 ratings (2 of which are wrong) on movies that are not among the top 100 most rated movies is enough to completely deanonymize 80% of the subscribers in the dataset.

They were able to calculate these percentages just by analysing the training data and applying some probability theory. The small amount of knowledge required to deanonymize the data may be found from casual conversation or social media, or in the public lists in the Internet Movie Database [xxxviii] . Narayanan and Shmatikov tested the IMDb approach by testing the ratings given by only 50 IMDb users; with just this small sample of data they were able to identify two subscribers in the Netflix training data [xxxix] .

Does this matter, when the data only represent movie ratings? The great importance of Narayanan and Shmatikov's results is their demonstration of the small amount of extra knowledge that may be needed before anonymised data can be reidentified with high confidence, and the probabilistic methods they developed for calculating this. There results have implications for the release of any "anonymised" personal data (and it is worth recalling that even half a dozen movie ratings can reveal far more than one might expect: research [xl] by psychologists at Cambridge University have shown that as small number of seemingly innocuous Facebook *Likes* can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender).

Example 3: Medical Records in Massachusetts

In the 1990s, the Group Insurance Commission (GIC), a Government agency in Massachusetts, started to release records to researchers that summarised the hospital visits of approximately 135,000 state employees and their families. All explicit identifiers, such as name, address, and Social Security number were removed, so the data were falsely believed to be anonymous and was then given freely to researchers and industry. Latanya Sweeney, Director of the Data Privacy Laboratory at Carnegie-Mellon University, obtained the data and compared it with the public voter registration records. She has said

One case stood out. William Weld was governor of Massachusetts at that time and his medical records were in the GIC data. Governor Weld lived in Cambridge Massachusetts. According to the Cambridge Voter list, six people had his particular birth date; only three of them were men; and, he was the only one in his 5-digit ZIP code. ***[xli]***

Conclusion

We have seen that privacy is not only a legal right but it may be essential to the safety or wellbeing of individuals and their families.

As these examples have illustrated, if data is released that contains details about individuals, then these individuals' identities can often be found by comparing the data with other available data. The computing term for the simplest version of this is *performing an inner join* between datasets [xlii] and there are many other analyses that can be performed if the simple approach is not possible. Moreover, each person that is identified adds to the ease with which further individuals can be identified.

This raises a number of issues.

Firstly, the release of any data that contains information about individuals (or the relationships between individuals) may enable those individuals to be identified, even if the data that has been released seems to contain no personally identifiable information.

Secondly, as more data is collected, stored and shared, the ability to anonymise personal data is further weakened. Yet the collection and use of *big data* is widely seen to be important for personalising services, improving efficiency and stimulating innovation.

Thirdly, the definition of *personally identifiable information* that is at the centre of data protection laws and regulations is too fluid to be useful. Paul Ohm concluded that *Data can be either useful or perfectly anonymous but never both* and it is hard to disagree, so the current basis for regulating data protection is probably unsustainable although the need to protect privacy remains. We need to find a different approach

---

[i] Data is a Latin plural noun but (just like *agenda*) usage has led to it being treated as singular in most contexts. I apologise if this offends any readers.

[ii] https://www.gov.uk/government/speeches/eight-great-technologies

[iii] http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html

[iv] https://www.nilsonreport.com/upload/PurchTransGlobalCards2015.gif

[v] http://www.wired.co.uk/article/privacy-versus-facebook

[vi] http://home.cern/about/computing

[vii] http://www.ohchr.org/EN/UDHR/Documents/UDHR_Translations/eng.pdf (Article 12)

[viii] https://en.wikipedia.org/wiki/Nothing_to_hide_argument

[ix] a mobile phone leaves a persistent data trail of where it is relative to network access points all the time it is switched on, and many apps will add GPS co-ordinates to their data

[x] http://www.independent.co.uk/life-style/gadgets-and-tech/baby-monitors-cctv-cameras-and-webcams-from-uk-homes-and-businesses-hacked-and-uploaded-onto-russian-9871830.html and http://www.bbc.co.uk/news/technology-30121159

[xi] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/216089/rehabilitation-offenders.pdf

[xii] https://www.marketingweek.com/2015/11/19/dunnhumby-founder-clive-humby-customer-insights-should-be-based-on-passions-as-well-as-purchases/

[xiii] http://ana.blogs.com/maestros/2006/11/data_is_the_new.html

[xiv] http://www.legislation.gov.uk/ukpga/1998/29/section/1

[xv] To avoid this definition identifying ministers and MPs as data controllers, the DPA adds this rider: *Where personal data are processed only for purposes for which they are required by or under any enactment to be processed, the person on whom the obligation to process the data is imposed by or under that enactment is for the purposes of this Act the data controller.*

[xvi] Directive 95/46/EC

[xvii] This Working Party was set up under Article 29 of Directive 95/46/EC. It is an independent European advisory body on data protection and privacy. Its tasks are described in Article 30 of Directive 95/46/EC and Article 15 of Directive 2002/58/EC.

[xviii] http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf

[xix] http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

[xx] https://ico.org.uk/media/for-organisations/documents/1624219/preparing-for-the-gdpr-12-steps.pdf

[xxi] http://www.transportapi.com/

[xxii] http://theodi.org

[xxiii] the data format *Comma-Separated Values*.

[xxiv] http://www.nationalarchives.gov.uk/doc/open-government-licence/version/1/

[xxv] https://data.gov.uk/dataset/metoffice_uklocs3hr_fc

[xxvi] https://www.gov.uk/unclaimed-estates-bona-vacantia/overview

[xxvii] Presentation at the Foundation for Science and Technology, Royal Society, London, 25 May 2016.

[xxviii] http://www.raeng.org.uk/publications/reports/global-navigation-space-systems

[xxix] http://gilc.org/privacy/survey/intro.html (accessed 18 May 2016)

[xxx] Philippe Golle, Revisiting the Uniqueness of Simple Demographics in the US Population, 5 ACM WORKSHOP ON PRIVACY IN THE ELEC. SOC'Y 77, 78 (2006)

[xxxi] http://www.uclalawreview.org/pdf/57-6-3.pdf (last accessed 7 June 2016)

[xxxii] http://techcrunch.com/2006/08/06/aol-proudly-releases-massive-amounts-of-user-search-data/

[xxxiii] http://techcrunch.com/2006/08/07/aol-this-was-a-screw-up/

[xxxiv] https://startpage.com/

[xxxv] https://duckduckgo.com/about

[xxxvi] http://www.netflixprize.com/rules

[xxxvii] *How To Break Anonymity of the Netflix Prize Dataset* http://arxiv.org/pdf/cs/0610105v1.pdf

[xxxviii] http://www.imdb.com/

[xxxix] http://cobweb.cs.uga.edu/~perdisci/CSCI6900-F10/BMeyer_Presentation3.pdf

[xl] http://www.pnas.org/content/110/15/5802.full.pdf

[xli] http://dataprivacylab.org/dataprivacy/talks/Flick-05-10.html (see recommendation 4)

[xlii] http://www.tutorialspoint.com/sql/sql-inner-joins.htm

Gresham College

Barnard's Inn Hall

Holborn

London

EC1N 2HH