Data Protection Act 1998

Big data and data protection

Information Commissioner's Office

Information Commissioner's Office

Contents

Introduction

1. Big data is currently a major topic of discussion across a number of fields, including management and marketing, scientific research, national security, government transparency and open data. Both public and private sectors are making increasing use of big data analytics. The ICO is interested specifically in the data protection and privacy risks posed by big data. In this paper, we set out what we understand by the term 'big data', consider what data protection issues it raises, and suggest how to comply with the Data Protection Act (the DPA).

2. This paper is intended to give an overview of the issues as we see them and contribute to the debate on big data and privacy. This is an area in which the capabilities of the technology and the range of potential applications are evolving rapidly and there is ongoing discussion of the implications of big data. Our aim is to ensure that the different privacy risks of big data are considered along with the benefits of big data - to organisations, to individuals and to society as a whole. It is our belief that the emerging benefits of big data will be sustained by upholding key data protection principles and safeguards. The benefits cannot simply be traded with privacy rights.

3. We carried out the research for this paper between June 2013 and June 2014. This involved desk research in the form of reading articles and reports (many of which are referenced in the footnotes), a small number of structured interviews with practitioners in other organisations, and discussions with experts at conferences and seminars. We are grateful to all those who contributed to this, as they have helped us to formulate our views on big data.

4. The area of big data is fast-evolving, and as such our guidance is likely to be subject to improvements and amendments in the future. We'd welcome any feedback on ways we can improve this document. Details on how to submit feedback are available in Annex 1.

5. In the paper we refer to a number of examples of big data applications used by companies and cite reports and other publications from companies. This information is taken from publicly available sources that are referenced in the footnotes. They are used for illustration and including them here does not imply any endorsement or otherwise by the ICO as the regulator of data protection in the UK.

Executive summary

6. Big data is currently a major topic of discussion across a number of fields, with both public and private sectors making increasing use of big data analytics. It is characterised by volume, variety and velocity of data, and by the use of algorithms, using 'all' the data and repurposing data.

7. The ICO is interested in big data as it can involve processing personal data. In this paper, we set out what we understand by the term 'big data', consider what data protection issues it raises, and suggest how to comply with the Data Protection Act (the DPA).

8. Many instances of big data analytics do not involve personal data at all. Using climate and weather data, for example, could enable new discoveries and improved services without using personal data.

9. However, there are many examples of big data analytics that do involve processing personal data, from sources such as social media, loyalty cards and sensors in clinical trials. Where personal data is being used, organisations must ensure they are complying with their obligations under the DPA.

10. One key data protection requirement is to ensure that processing of personal data is fair, and this is particularly important where big data is being used to make decisions affecting individuals. Fairness is partly about how personal data is obtained. Organisations need to be transparent when they collect data, and explaining how it will be used is an important element in complying with data protection principles. The complexity of big data analytics is not an excuse for failing to obtain consent where it is required.

11. Big data analytics can involve repurposing personal data. If an organisation has collected personal data for one purpose and then decides to start analysing it for completely different purposes (or to make it available for others to do so) then it needs to make its users aware of this. This is particularly important if the organisation is planning to use the data for a purpose that is not apparent to the individual because it is not obviously connected with their use of a service.

12. A key feature of big data is using 'all' the data, which contrasts with the concept of data minimisation in the data protection principles. This raises questions about whether big data is excessive, while the variety of data sources often used in the

analysis may also prompt questions over whether the personal information being used is relevant. The challenge for organisations is to address this by being clear from the outset what they expect to learn or be able to do by processing that data, as well as satisfying themselves that the data is relevant and not excessive, in relation to that aim.

13. Organisations must also be proactive in considering any information security risks posed by big data. Security depends upon the proper assessment of risk, and so responsible organisations should apply their normal risk management policies and procedures when they acquire new datasets or use existing ones for big data analytics. Big data can also be a tool to improve information security.

14. The proposed EU General Data Protection Regulation, if adopted, could improve the level of data protection for individuals in the context of big data analytics, in that it aims to increase the transparency of the processing, enhance the rights of data subjects and introduce a requirement for privacy by design and privacy impact assessments. The ICO stresses that these data protection benefits should be achieved through a risk based approach that avoids over-prescription.

15. There is evidence that some companies are developing an approach to big data that looks to place it in a wider and essentially ethical context. As well as a possible competitive advantage in being seen as a responsible and trustworthy custodian of customer data, adopting an ethical approach will also go some way towards ensuring that the analytics complies with data protection principles.

16. There are clear social benefits to be derived from big data analytics, for example in scientific and medical research. Being transparent about the purpose and impact of the analytics can also have benefits, in helping people to be confident as 'digital citizens' in a big data world.

17. We do not accept the argument that data protection principles are not fit for purpose in the context of big data. Big data is not a game that is played by different rules. There is some flexibility inherent in the data protection principles. They should not be seen as a barrier to progress, but as the framework to promote privacy rights and as a stimulus to developing innovative approaches to informing and engaging the public.

18. As well as looking extensively at the data protection issues presented by big data, this paper also suggests areas

organisations should address when considering big data analytics.

19. One route to consider is anonymisation, which, if done correctly, means the information being analysed is no longer considered personal data. This can assist big data analytics and help organisations to carry on research or develop products and services. It also enables organisations to give an assurance to the people whose data was collected that they are not using data that identifies them for big data analytics. In a world of multiple data sources effective anonymisation can be challenging and organisations must carry out a robust risk assessment.

20. Other practical aspects to consider when using personal data in big data analytics are summarised in this table:

Personal data	Does your big data project need to use personal data at all? If you are using personal data, can it be anonymised? If you are processing personal data you have to comply with the Data Protection Act.
Privacy impact assessments	Carry out a privacy impact assessment to understand how the processing will affect the people concerned. Are you using personal data to identify general trends or to make decisions that affect individuals?
Repurposing data	If you are repurposing data, consider whether the new purpose is incompatible with the original purpose, in data protection terms, and whether you need to get consent. If you are buying in personal data from elsewhere, you need to practice due diligence and ensure that you have a data protection condition for your processing.
Data minimisation	Big data analytics is not an excuse for stockpiling data or keeping it longer than you need for your business purposes, just in case it might be useful. Long term uses must be articulated or justifiable, even if all the detail of the future use is not known.
Transparency	Be as transparent and open as possible about what you are doing. Explain the purposes, implications and benefits of the analytics. Think of innovative and effective ways to convey this to the people concerned.

Subject access	People have a right to see the data you are processing about them. Design systems that make it easy for you to collate this information. Think about enabling people to access their data on line in a re-usable format.

What is big data?

21. It is difficult to produce a 'watertight' definition that would enable us to label definitively any particular instance of processing as big data or not. Big data has been described as a phenomenon rather than a technology[1], and this is a useful distinction. It is perhaps best to see it as a shorthand way of describing the features of certain data and how it is processed. The Gartner IT glossary defines it as follows:

> "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making"[2].

22. Following the Gartner definition, big data is often described in terms of the 'three Vs': volume, variety and velocity.

23. **Volume**. Big data uses massive datasets, including for example meta-data from internet searches, credit and debit card purchases, social media postings, mobile phone location data, or data from sensors in cars and other devices. The volume of data being produced in the world continues to increase rapidly. The Boston Consulting Group estimates total growth of 2.5 exabytes, which equals 2.5 billion gigabytes, per day[3]. It is increasingly possible to hold very large datasets, due to the decreasing cost of storage and the availability of cloud-based

[1] Wiggins, Lesley If big data and analytics exist in a silo, does the outcome matter? IBM Big Data and Analytics Hub, 25 February 2014 http://www.ibmbigdatahub.com/blog/if-big-data-and-analytics-exist-silo-does-outcome-matter Accessed 25 June 2014

[2] Gartner IT glossary Big data. http://www.gartner.com/it-glossary/big-data Accessed 25 June 2014

[3] Souza, Robert et al. How to get started with big data. BCG perspectives. The Boston Consulting Group, 29 May 2013 https://www.bcgperspectives.com/content/articles/it_strategy_retail_how_to_get_started_with_big_data/ Accessed 25 June 2014

services. These datasets may be so large that they cannot be analysed using 'traditional' methods, such as MS Excel spreadsheets, relational databases and SQL queries, but new tools have been developed to analyse them such as NoSQL and the open source software Hadoop.

24. **Variety**. Big data often involves bringing together data from different sources. Currently it appears that big data analytics mainly uses structured data[4], eg in tables with defined fields, but it can also include unstructured data. For example, it is possible to obtain a feed of all the data coming from a social media source such as Twitter. This is often used for 'sentiment analysis', ie to analyse what people are saying about products or organisations. A retailer might combine this data with their own in-house data collected from point-of-sale terminals and loyalty cards, to produce rich and detailed information for marketing. From an IT perspective, combining data from different sources in this way presents particular challenges. Technologies have been developed for big data that do not require all of the data to be put into a single database structure before it can be analysed.

25. In discussions we have had with practitioners, some have suggested that, of the 'three Vs', variety is the most important characteristic of big data. This view suggests that, if a company is analysing its own customer database, even if that database is particularly large, it may not necessarily raise any novel issues in terms of either analytics or data protection. However, when it combines its own information with data sourced externally (whether that be from a publicly accessible source or not), then it is doing something qualitatively different that can be called big data.

26. **Velocity**. In some contexts, it is important to analyse data as quickly as possible, even in real time. Big data analytics can be used to analyse data 'in motion', as it is produced or recorded, as well as data 'at rest' in data stores. A potential application of 'in motion' analysis is in credit card payments. For example, Visa[5] is looking at using big data analytics to develop a new ways of authorising credit card payments.

[4] Russom, Philip Managing big data. The Data Warehousing Institute, 2013. Available from http://www.pentaho.com/resources Accessed 25 June 2014
[5] The future of technology and payments. Edition 2. Visa Europe, April 2013 http://www.visaeurope.com/en/about_us/industry_insights/tech_trends.aspx Accessed 25 June 2014

27. There is no fixed definition of big data and some commentators have suggested additional criteria, but we take the 'three Vs' to be key elements of big data. To be described as big data, a particular instance of data processing has to be significant in terms of volume, variety or velocity, but it does not have to score highly on all three of these parameters.

28. In addition to the 'three Vs', big data analytics often has other characteristics that are different to those of 'traditional' processing.

29. **Use of algorithms.** Before the advent of big data, analysing a dataset involved, in general terms, deciding what you wanted to find out from the dataset and constructing a query to find it, by identifying the relevant entries. Big data analytics, on the other hand, often involves running a very large number of algorithms against the data in order to find correlations[6], rather than testing a particular hypothesis. Once relevant correlations have been identified, a new algorithm can be created and applied to particular cases. This is a form of 'machine learning', since the system 'learns' which are the relevant criteria from analysing the data. While algorithms themselves are by no means a new concept, their use in this way is a feature of big data analytics.

> **Example**
>
> In the United States, Vree Health uses analytics to reduce hospital readmission rates for patients hospitalised for heart attack, heart failure or pneumonia. They use data collected throughout the hospital stay (this can amount to 12 million pieces of data per day) and combine it with data from follow-up visits by their representatives, patient interactions with their internet and phone-based help systems and third party data. Analysing all of this data enables them to spot characteristics or behaviours associated with readmission, and if a particular patient exhibits these they can put support in place.[7]

[6] Centre for Information Policy Leadership. Big data and analytics. Seeking foundations for effective privacy guidance. Hunton and Williams LLP, February 2013 http://www.hunton.com/files/Uploads/Documents/News_files/Big_Data_and_Analytics_February_2013.pdf Accessed 25 June 2014

[7] Ibid, pp 4-6

30. **Using 'all the data'.** Analysing data for research often necessitates finding a statistically representative sample, or carrying out random sampling. Big data analytics, on the other hand, tends to collect and analyse all the data that is available. For example, in a retail context it can mean analysing all the purchases made by shoppers using a loyalty card, and using this to find correlations, rather than asking a sample of shoppers to take part in a survey. This feature of big data has been made easier by the ability to store and analyse ever increasing amounts of data.

31. **Repurposing data.** Big data analytics often repurposes data that was obtained for a different purpose and in some cases by another organisation. Companies such as DataSift take data from Twitter, Facebook and other social media and make it available for analysis for marketing and other purposes. Social media data can also be used to assess individuals' credit worthiness[8]. In many cases the data that is being used for the analytics has been generated automatically, by interactive technology, rather than being consciously provided by individuals. For example, mobile phone presence data is used to analyse the footfall in retail centres[9]. Sensors in the street or in shops can capture the unique MAC address of the mobile phones of people passing by[10]. Although the MAC address does not itself identify a specific individual, it could be used to track repeated visits.

32. Some see big data merely as a continuation of the processing they have always done, even when they are handling very large volumes of data. This may be because they do not consider that what they are doing creates any new issues for them, especially in terms of data protection, or because they are sceptical of the 'hype' surrounding big data. Certainly, some examples may be closer to traditional business intelligence analytics even though they are reported as big data. Where the term is being used only as marketing, and the processing represents a continuation of what has been happening previously, then it is unlikely that it will raise any

[8] Deville, Joe. Leaky data: how Wonga makes lending decisions. Charisma, May 2013 http://www.charisma-network.net/finance/leaky-data-how-wonga-makes-lending-decisions Accessed 25 June 2014

[9] Smart Steps increase Morrisons new and return customers by 150%. Telefonica Dynamic Insights October 2013 http://dynamicinsights.telefonica.com/1158/a-smart-step-ahead-for-morrisons Accessed 25 June 2014

[10] Seward, Zachary M and Datoo, Siraj City of London halts recycling bins tracking phones of passers-by Quartz 12 August 2013 http://qz.com/114174/city-of-london-halts-recycling-bins-tracking-phones-of-passers-by/ Accessed 25 June 2014

new issues in terms of data protection. In practice it may be difficult to say definitively that a particular example of processing is or is not big data. Indeed, organisations are sometimes reluctant to label what they are doing as big data.

33. A difficulty in identifying instances of big data analytics may also be due to the fact that, despite the level of discussion of big data, the actual take-up of it in the UK appears to be still relatively low, although it is growing. A survey[11] by e-skills UK for SAS found that in 2012 only 14% of firms with more than 100 employees had adopted big data. This is expected to rise to 29% by 2017. The same survey estimated that take-up amongst small and medium-sized enterprises was less than 0.2%. The creation of the Alan Turing Institute for Data Science, announced in the 2014 Budget[12], is likely to further encourage the adoption of big data analytics in the UK.

Big data and personal data

34. Personal data is data that relates to an identifiable living individual. 'Identifiable' means that the individual can be identified from that data, either alone or in combination with other information. In assessing whether the data could be combined with other information to identify an individual, it is necessary to consider what means are reasonably likely to be used to identify them. The ICO has produced guidance that explains this definition further[13].

35. Data protection is concerned with personal data, but it is important to remember that many instances of big data analytics do not involve personal data at all. Examples of non-personal big data include: world climate and weather data; using geospatial data from GPS-equipped buses to predict arrival times; data from radio telescopes in the Square

[11] Big data analytics. Adoption and employment trends 2012-2017. e-skills UK, January 2013. Available from http://ec.europa.eu/digital-agenda/en/news/big-data-analytics-assessment-demand-labour-and-skills-2012-2017 Accessed 25 June 2014

[12] Department for Business Innovation and Skills. Plans for world class research centre in the UK. Press release 19 March 2014 https://www.gov.uk/government/news/plans-for-world-class-research-centre-in-the-uk Accessed 25 June 2014

[13] Information Commissioner's Office. Determining what is personal data. ICO, December 2012. http://ico.org.uk/for_organisations/guidance_index/~/media/documents/library/Data_Protection/Detailed_specialist_guides/PERSONAL_DATA_FLOWCHART_V1_WITH_PREFACE001.ashx Accessed 25 June 2014

Kilometre Array[14]; data from sensors on containers carried on ships. These are all areas where big data analytics enable new discoveries and improve services and business processes, without using personal data.

36. However, there are many examples of big data analytics that do involve processing personal data, for example: data from monitoring devices on patients in clinical trials, mobile phone location data, data on purchases made with loyalty cards and biometric data from body-worn devices.

37. There is a trend towards what is sometimes called the 'segment of one': fine tuning the offer of products of services to an individual based on their characteristics such as age, preferences, lifestyle etc. Findings derived from big data analytics can be applied to marketing to individuals (eg 'people with these characteristics tend to want these products or services at this time').

38. Big data analytics also has the potential to create new personal data. For example, social media and other data about an individual could be analysed to find out about that person's lifestyle as a factor in determining their credit rating, or whether they are at risk of developing a medical condition. Similarly, sensors in cars provide vast amounts of data about the car, but this can also be used to identify patterns in people's driving behaviour, which can help to inform decisions about their insurance premiums.

39. Big data can be used to identify general trends, rather than to understand more about individuals or to make decisions about them. For example, data from travel cards that record journeys made by individuals, such as Oystercard, could be combined with traffic data to plan new bus routes. In such cases the data may often be anonymised before being analysed.

Anonymisation

40. If personal data is fully anonymised, it is no longer personal data. In this context, anonymised means that it is not possible to identify an individual from the data itself or from that data in combination with other data, taking account of all the means that are reasonably likely to be used to identify them.

[14] Square Kilometre Array website https://www.skatelescope.org/ Accessed 25 June 2014

41. In some cases, the data used for big data analytics is anonymised for the purposes of the analysis. For example, Telefonica's Smart Steps[15] tool uses data on the location of mobile phones on its network in order to track the movement of crowds of people. This can be used by retailers to analyse footfall in a particular location. The data that identifies individuals is stripped out prior to the analysis and the anonymised data is aggregated to gain insights about the population as a whole and combined with market research data from other sources. Another example of this is in medical research, where data from clinical trials is rigorously anonymised before being made available for analysis. In practice, anonymised data may be used in a number of different scenarios: organisations may bring in anonymised data, or they may seek to irreversibly anonymise their own data before using it themselves or sharing it with others.

42. It may not be possible to establish with absolute certainty that an individual cannot be identified from a particular dataset in combination with other data that may exist elsewhere. Some commentators have pointed to examples of where it has been possible to identify individuals in apparently anonymised datasets, and so concluded that anonymisation is becoming increasingly ineffective [16]. On the other hand, Cavoukian and Castro[17] have found shortcomings in the main studies on which this view is based. The issue is not about eliminating the risk of re-identification altogether, but whether it can be mitigated so it is no longer significant. Organisations should focus on mitigating the risks to the point where the chance of re-identification is extremely remote. The range of datasets available and the power of big data analytics make this more difficult, and this risk should not be underestimated, but that does not mean anonymisation is impossible or that it is not an effective tool.

43. Organisations using anonymised data need to be able demonstrate that they have carried out this robust assessment

[15] Telefonica. Smart Steps http://dynamicinsights.telefonica.com/488/smart-steps Accessed 25 June 2014

[16] eg President's Council of Advisors on Science and Technology. Big data and privacy. A technological perspective. White House, May 2014 http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf Accessed 20 June 2014

[17] Cavoukian, Anne and Castro, Daniel. Big data and innovation, setting the record straight: de-identification does work. Office of the Information and Privacy Commissioner, Ontario, June 2014. http://www.privacybydesign.ca/index.php/paper/big-data-innovation-setting-record-straight-de-identification-work Accessed 20 June 2014.

of the risk of re-identification, and have adopted solutions proportionate to the risk. This may involve a range and combination of technical measures, such as data masking, pseudonymisation, aggregation and banding, as well as legal and organisational safeguards. The ICO's Anonymisation code of practice[18] explains these in more detail.

44. Anonymisation may be used when data is shared externally or within an organisation. For example, an organisation may hold a dataset containing personal data in one data store, and produce an anonymised version of it to be used for analytics in a separate area. Whether it remains personal data will depend on whether the anonymisation "keys" and other relevant data that enable identification are retained by the organisation. Even if the data remains personal data this is still a relevant safeguard to consider in order to enable processing to comply with the data protection principles.

45. Our Anonymisation code of practice[19] explains this and gives advice on anonymisation techniques. In addition, the UK Anonymisation Network (UKAN)[20] has an important role in providing expert advice on anonymisation techniques. This has been funded by the ICO and is co-ordinated by a consortium of the Universities of Manchester and Southampton, the Open Data Institute and the Office for National Statistics.

46. Anonymisation should not be seen merely as a means of reducing a regulatory burden by taking the processing outside the DPA. It is a means of mitigating the risk of inadvertent disclosure or loss of personal data, and so is a tool that assists big data analytics and helps the organisation to carry on its research or develop its products and services. It also enables the organisation to give an assurance to the people whose data it collected that it is not using data that identifies them for its big data analytics. This is part of the process of building trust which is key to taking big data forward. We return to the issue of building trust in the section on The business context.

[18] Information Commissioner's Office. Anonymisation: managing data protection risk code of practice. ICO, November 2012. http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf Accessed 25 June 2014

[19] Ibid

[20] UK Anonymisation Network website http://ukanon.net/ Accessed 25 June 2014

Fairness

47. Under the first principle of the DPA, the processing of personal data must be fair and lawful, and must satisfy one of the conditions listed in Schedule 2 of the DPA (and Schedule 3 if is sensitive personal data as defined in the DPA). The first question for organisations to consider when using personal data for big data analytics is whether the processing is fair. This is particularly important when the type of processing is novel and complex, as in big data analytics. Big data is sometimes characterised as sinister or a threat to privacy or simply 'creepy' because it involves repurposing data in unexpected ways, using complex algorithms, and drawing conclusions about individuals[21]. Assessing whether the processing meets the requirement of fairness involves directly addressing this issue.

48. Fairness is partly about how personal data is obtained. The processing is unlikely to be fair if people are deceived or misled about how their data will be used at the point they are providing it. This means that transparency about how the date will be used is an important element in assessing whether big data analytics comply with the data protection principles. We discuss this further in the section on Transparency and privacy information.

49. It is also necessary to consider the effect of the processing on the individuals concerned. This does not simply mean whether it produces an outcome that they would want. For example, HMRC's Connect[22] system uses analytics, drawing on a wide range of data sources, to identify tax fraud and while this may be unwelcome to those identified it does not follow that it is unfair (in certain circumstances there is also an exemption in

[21] Eg Naughton, John Why big data has made your privacy a thing of the past Guardian online 6 October 2013 http://www.theguardian.com/technology/2013/oct/06/big-data-predictive-analytics-privacy Accessed 25 June 2014; Richards Neil M. and King, Jonathan H. Three paradoxes of big data 66 Stanford Law Review Online 41 3 September 2013 http://www.stanfordlawreview.org/online/privacy-and-big-data/three-paradoxes-big-data Accessed 25 June 2014; Leonard, Peter. Doing big data business: evolving business models and privacy regulation. August 2013. International Data Privacy Law 18 December 2013. http://idpl.oxfordjournals.org/content/early/2013/12/18/idpl.ipt032.short?rss=1 Accessed 25 June 2014

[22] Smarter Government Public Sector Fraud Taskforce. A fresh approach to combating fraud in the public sector National Fraud Authority. March 2010. http://www.eurim.org.uk/activities/psd/A-fresh-approach-to-combating-fraud-in-the-public-sector.pdf Accessed 25 June 2014

section 29 of the DPA from the fairness requirement for personal data processed for tax purposes). Fairness involves a wider assessment of whether the processing is within the reasonable expectations of the individuals concerned. For example, every aspect of analysing loyalty card data to improve marketing should not always automatically be considered fair, or within customer expectations. The issues that can arise in this context are illustrated by the well-publicised example of the US company, Target[23].

Example

The US retailer Target was interested in identifying points in consumers' lives at which they are open to changes in their buying habits. As part of this they wanted to be able to predict when a customer was going to have a baby, so that they could market relevant products to them in advance. They started by analysing the purchases of customers in their 'baby shower registry' who had given the company their due date. From this they were able to identify a group of products which indicated that a customer was expecting a baby, and they found a correlation between the dates of those purchases and their due date. They then applied this to purchases by every female shopper in their customer database to identify those who were likely to be pregnant as well as their likely due date. Target could then send them marketing and offers on pregnancy and baby-related products, and take advantage of any change in their buying habits to encourage them to go on to buy other products.

This was highlighted when a father complained to a Target store outside Minneapolis that his daughter, who was still in high school, had received these coupons when she wasn't pregnant. It transpired that Target's predictive model was accurate after all: his daughter was pregnant, but her father did not know it. The man subsequently apologised to the store.

This is an example from the USA, which does not have the same data protection regime as the UK and the EU, but it nevertheless highlights issues of fairness and customer expectations that can arise in the context of big data analytics.

[23] Duhigg, Charles. How companies learn your secrets. New York Times online 16 February 2012. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=all&_r=0 Accessed 21 July 2014.

50. There is also a difference between using personal data when the purpose of the processing fits with the reason that people use the service and one where the data is being used for a purpose that is not intrinsic to the delivery of the service. A retailer using loyalty card data for market research is an example of the former. A social media company making its data available for market research is an example of the latter. This does not mean that the latter is necessarily unfair; this depends on what people are told when they join and use the social media service. It is important to make people aware of what is going to happen to their data if they choose to use the service. This is discussed further in the section on Transparency and privacy information below.

51. How big data is used is an important factor in assessing fairness. Big data analytics may use personal data purely for research purposes, eg to detect general trends and correlations, or it may use personal data in order to make decisions affecting individuals, such as setting their insurance premium. Any processing of personal data must be fair, but if the analytics are being used to make decisions affecting individuals, the assessment of fairness must be even more rigorous.

52. There is a danger that the algorithms may be used in a way that perpetuates stereotyping or even bias. For example, an online customer's likes, purchases and search history can be used to present targeted advertisements to them when they are searching or viewing web pages. This may be seen as a benefit of big data analytics because it tailors their experience of the internet to their interests. However, it can also mean that they are being profiled in a way that perpetuates discrimination, for example on the basis of race[24]. Research in the USA suggested that internet searches for "black-identifying" names generated advertisements associated with arrest records much more frequently than those for "white-identifying" names[25]. Profiling can also be used in ways that have a more direct effect, for example if decisions about a

[24] Rabess, Cecilia Esther. Can big data be racist? The Bold Italic 31 March 2014. http://www.thebolditalic.com/articles/4502-can-big-data-be-racist Accessed 25 June 2014

[25] Sweeney, Laetitia. Discrimination in online ad delivery. Data Privacy Lab January 2013. http://dataprivacylab.org/projects/onlineads/1071-1.pdf Accessed 25 June 2014

person's creditworthiness are based on information about where they live, their shopping habits or social contacts. This is a complex area. It is not necessarily the case that there is discrimination because a person belongs to a particular social group, but they are being treated in a certain way based on factors that they share with members of that group.

53. The quality of the data being used for the analytics may also be an issue. This is not so much a question of whether the data accurately records what someone said or did, but rather to what extent that provides a reliable basis for drawing conclusions. The issue of the quality and reliability of the data has been recognised as an issue in discussions of big data analytics.[26]

54. Organisations should also remember that under section 12 of the DPA, individuals have certain rights to prevent decisions being taken about them that are based solely on automated processing of their personal data. This situation may not have arisen very often, since there is usually human oversight of the decision in some form, but the increased use of big data analytics may now lead to more situations in which decisions are taken by automated processing.

Conditions for processing personal data

55. Under the first principle of the DPA, the processing of personal data must not only be fair and lawful, but must also satisfy one of the conditions listed in Schedule 2 of the DPA (and Schedule 3 if is sensitive personal data as defined in the DPA). This applies equally to big data analytics that use personal data. The Schedule 2 conditions that are most likely to be relevant to big data analytics, particularly in a commercial context, are consent, whether processing is necessary for the performance of a contract, and the legitimate interests of the data controller or other parties. Our Guide to data protection[27] explains these conditions in more detail. Here we consider how they relate to big data analytics specifically.

---

[26] Eg Forrester Consulting. Big data needs agile information and integration governance. Forrester Research Inc, August 2013. Available from: http://www.ibmbigdatahub.com/whitepaper/big-data-needs-agile-information-and-integration-governance Accessed 25 June 2014

[27] Information Commissioner's Office. Guide to data protection. ICO, November 2009. http://ico.org.uk/for_organisations/data_protection/~/media/documents/library/Data_Protection/Practical_application/the_guide_to_data_protection.pdf Accessed 25 June 2014

Consent

56. If an organisation is relying on people's consent as the condition for processing their personal data, then that consent must be freely given, specific and informed[28]. This means people must be able to understand what the organisation is going to do with their data and there must be a clear indication that they consent to it. If an organisation has collected personal data for one purpose and then decides to start analysing it for completely different purposes (or to make it available for others to do so) then it needs to make its users aware of this. This is particularly important if the organisation is planning to use the data for a purpose that is not apparent to the individual because it is not obviously connected with their use of a service. For example, if a social media company were selling on the wealth of personal data of its users to another company for other purposes.

57. It may be possible to have a process of graduated consent. A study commissioned by the International Institute of Communications[29] found that some users wanted to be able to give consent (or not) to different uses of their data throughout their relationship with a service provider, rather than having a simple 'binary' choice at the beginning. For example, they could give an initial consent to opt in to the system and then separate consent for their data to be shared with other parties. Furthermore, they wanted a value exchange, ie to receive some additional benefit in return for giving their consent.

58. It may be reasonable for organisations to use consent as a condition for processing in a big data context but they have to be sure that it is the appropriate condition. Furthermore, if people do not have a real choice and are not able to withdraw their consent if they wish, then the consent would not meet the standard required by the DPA.

59. If an organisation buys a large dataset of personal data for analytics purposes, then it becomes a data controller in respect of that data. The organisation needs to be sure that it has met

[28] For a detailed discussion of consent, see Article 29 Data Protection Working Party Opinion 15/2011 on the definition of consent. European Commission 13 July 2011. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2011/wp187_en.pdf Accessed 25 June 2014

[29] International Institute of Communications. Personal data management: the user's perspective. International Institute of Communications, September 2012. http://www.iicom.org/open-access-resources/doc_details/226-personal-data-management-the-users-perspective Accessed 25 June 2014

a condition in the DPA for the further use of that data. If it is relying on the original consent obtained by the supplier as that condition, then it should ensure that this covers the further processing it plans for the data.

60. The apparent complexity of big data analytics should not become an excuse for failing to seek consent where it is required. Organisations must find the point at which to explain the benefits of the analytics and present users with a meaningful choice - and then respect that choice when they are processing their personal data. Paul Ohm argues[30] that Google "breached a wall of trust" by using search data to find correlations between search terms and recorded cases of flu in its Google Flu Trends[31] project. He suggests that if they had explained that they were doing this "to help avoid pandemics" or "to save lives", users would have been likely to agree to it. This may be debateable, but his argument does raise the issue that if an organisation can identify potential benefits from using personal data in big data analytics, it should be able to explain these to users, and seek consent, if it is required as a condition for the processing. Furthermore, while consent may be the condition for processing, this does not mean that by consenting people trade all their privacy rights. The processing must still be fair and lawful, and they retain their rights under the DPA.

Contracts and legitimate interests

61. While consent is one condition for processing personal data, it is not the only condition available under the DPA. Organisations do not have to seek consent in all cases, and other conditions may be relevant.

62. Condition 2 of Schedule 2 of the DPA is that the processing is necessary for the performance of a contract to which the data

---

[30] Ohm, Paul. The underwhelming benefits of big data. 161 University of Pennsylvania Law Review Online 339 (2013). http://www.pennlawreview.com/online/161-U-Pa-L-Rev-Online-339.pdf Accessed 25 June 2014

[31] Ginsberg, Jeremy at al. Detecting influenza epidemics using search engine query data. Google, 2009. http://static.googleusercontent.com/media/research.google.com/en//archive/papers/detecting-influenza-epidemics.pdf. Accessed 25 June 2014. See also: Fung, Kaiser. Google flu trends' failure shows good data > big data HBR Blog Network 25 March 2014. http://blogs.hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data/?utm_source=Socialflow&utm_medium=Tweet&utm_campaign=Socialflow Accessed 25 June 2014

subject is a party. This is relevant, for example, when someone makes a purchase online, and their name, address and credit card details have to be processed by the website in order to carry out the purchase. Specific consent is not required for this. The problem in applying this in a big data context is that the processing must be "necessary". Big data analytics, by its nature, is likely to represent a level of analysis that goes beyond what is required simply to sell a product or deliver a service. It often takes the data that is generated by the basic provision of a service and repurposes it. It may therefore be difficult to show that the big data analytics are strictly necessary for the performance of a contract. However, given the potential development of new payment methods[32], it is conceivable that there may be situations in which big data analytics are necessary for authentication purposes.

63. Alternatively, under condition 6, the processing may be necessary for the legitimate interests of the organisation collecting the data (or others to whom it is made available). An organisation may have a number of legitimate interests that could be relevant, including, for example: profiling customers in order to target its marketing; preventing fraud or the misuse of its services; physical or IT security. However, having established that it has a legitimate interest, the organisation then has to carry out a balancing exercise between those interests and the rights, freedoms and legitimate interests of the individuals concerned. Organisations seeking to rely on this condition will therefore have to pay particular attention to the impact of the analytics on people's privacy. This can be a complex assessment involving a number of factors. The recent opinion of the Article 29 Working Party[33] on legitimate interests sets out in detail how to assess these factors and carry out the balancing exercise.

64. Furthermore, to meet this condition the processing must be "necessary" for the legitimate interests. This means that it must be more than just potentially interesting. The processing

[32] As discussed for example in The future of technology and payments. Edition 2. Visa Europe, April 2013 http://www.visaeurope.com/en/about_us/industry_insights/tech_trends.aspx Accessed 25 June 2014

[33] Article 29 Data Protection Working Party. Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. European Commission 9 April 2014. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf Accessed 25 June 2014

is not necessary if there is another way of meeting the legitimate interest that interferes less with people's privacy.

Purpose limitation

65. The second data protection principle creates a two-part test: firstly the purpose for which the data is collected must be specified and lawful, and secondly, if the data is further processed for any other purpose, it must not be incompatible with the original purpose.

66. It has been suggested[34] that big data challenges the principle of purpose limitation, and that the principle is a barrier to the development of big data analytics. This reflects a view of big data analytics as a fluid and serendipitous process, in which analysing data using many different algorithms reveals unexpected correlations that can be used for new purposes. It is suggested that the purpose limitation principle restricts an organisation's freedom to make these discoveries and innovations.

67. However, what is called 'purpose limitation' could more accurately (if more awkwardly) be described as 'non-incompatibility'. The DPA does not say that processing for a new purpose is not permissible, nor does it say that the new purpose must be the same as the original purpose, nor even that it must be compatible with the original purpose: it says that it must not be incompatible with it.

> The Article 29 Working Party Opinion on purpose limitation[35] says:
>
> "By providing that any further processing is authorised as long

[34] Eg in World Economic Forum Unlocking the value of personal data; from collection to usage. World Economic Forum February 2013 http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf Accessed 25 June 2014 ; and Rubinstein, Ira S Big data. The end of privacy or a new beginning? International Data Privacy Law 25 January 2013. http://idpl.oxfordjournals.org/content/early/2013/01/24/idpl.ips036.full.pdf+html Accessed 25 June 2014

[35] Article 29 Data Protection Working Party. Opinion 03/2013 on purpose limitation. European Commission 2 April 2013. http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf Accessed 25 June 2014

> as it is not incompatible (and if the requirements of lawfulness are simultaneously also fulfilled), it would appear that the legislators intended to give some flexibility with regard to further use. Such further use may fit closely with the initial purpose or be different. The fact that the further processing is for a different purpose does not necessarily mean that it is automatically incompatible: this needs to be assessed on a case-by-case basis …" (p. 21)

68. The Article 29 Working Party Opinion sets out a detailed approach to assessing whether any further processing is for an incompatible purpose. It also addresses directly the issue of repurposing data for big data analytics. It identifies two types of further processing: firstly where it is done to detect trends or correlations and secondly where it is done to find out about individuals and make decisions affecting them. In the first case it advocates a clear separation between the analytics. In the second it says that "free, specific, informed and unambiguous 'opt-in' consent would almost always be required, otherwise further use cannot be considered compatible"[36]. It also emphasises the need for transparency, and for allowing people to correct and update their profiles and to access their data in a portable, user-friendly and machine-readable format.

69. In our view, a key factor in deciding whether a new purpose is incompatible with the original purpose is whether it is fair. In particular, this means considering how the new purpose affects the privacy of the individuals concerned and whether it is within their reasonable expectations that their data could be used in this way. If, for example, information that people have put on social media is going to be used to assess their health risks or their credit worthiness, or to market certain products to them, then unless they are informed of this and asked to give their consent, it is unlikely to be either fair or compatible. Where the new purpose would be otherwise unexpected, and it involves making decisions about them as individuals, then in most cases the organisation concerned will need to seek specific consent, in addition to establishing whether the new purpose is incompatible with the original reason for processing the data.

70. If an organisation is buying personal data from elsewhere for big data analytics, it needs to practice due diligence. It should first consider whether it needs to use personal data at all, or whether it could take the data in anonymised form. If it is

[36] *Ibid* p.46

acquiring personal data, then it becomes the data controller for it and has to meet the requirements of DPA. The organisation should establish whether the individuals concerned have in fact consented to this further use of their data, or whether it can rely on another data protection condition. If not, it will need to tell those individuals what it is doing and seek consent for the new use. It will also need to assess whether the new processing is incompatible with the original purpose for which the data was collected.

Data minimisation: collection and retention

71. Data protection legislation embodies the concept of data minimisation – that is, organisations should minimise the amount of data they collect and process, and the length of time they keep the data. Principle 3 of the DPA says that "personal data shall be adequate, relevant and not excessive in relation to the purpose or purposes for which they are processed", while principle 5 says that "personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes".

72. By contrast, big data analytics tends to involve collecting as much data as possible. Mayer-Schönberger and Cukier[37] describe this as "N=all". This means that big data is about collecting and analysing all the data points in a particular set, for example the logs of all the phone calls made via a wireless operator in a particular country for a four month period. Where the N=all approach is applied to smaller datasets, the key point remains that all the data points are used for analysis, rather than a sample.

73. Big data is also about the variety of data sources that are used in the analysis. In terms of principle 3, this raises questions not only about whether the data is excessive, but also whether it is relevant. Big data may discover unexpected correlations, for example between data about people's lifestyles and their credit worthiness, but that does not necessarily mean that any information that can be obtained about those individuals is necessarily relevant to the purpose of assessing credit risk. Finding the correlation does not retrospectively justify obtaining the data in the first place. Organisations therefore need to be able to articulate at the outset why they need to collect and

[37] Mayer-Schönberger, Viktor and Cukier, Kenneth, in Chapter 2 of Big data. A revolution that will transform how we live, work and think. John Murray, 2013

process particular datasets. They need to be clear about what they expect to learn or be able to do by processing that data and satisfy themselves that the data is relevant and not excessive, in relation to that aim. The challenge is therefore to define the purposes of the processing and establish what data will be relevant to them.

74. The principle 5 requirement, that personal data shall not be kept longer than necessary for the purpose for which it is being processed, supports the data privacy of individuals and it also reflects good practice in records management. However, in the world of big data it may potentially be challenged on two fronts. Firstly, the capacity to store data increases all the time and the cost of storing it is falling. Ten years ago IBM kept a record of all the data warehouses of more than one terabyte that they knew of[38] ; this is now the standard storage capacity of a new home pc. Secondly, the ability of big data analytics to process very large volumes of data may encourage data controllers to keep long runs of historical data, beyond the period required for normal business purposes.

75. While this may be a potential issue, we have not seen clear evidence that commercial organisations are changing their practices to keep data longer than necessary, just in case it proves to be useful for big data analytics. If organisations wish to retain data for long periods for reasons of big data analytics they should be able to articulate and foresee the potential uses and benefits to some extent, even if the specifics are unclear. Retention periods are specified in general records management and accounting standards, and in some sectors there are regulatory requirements as to how long records should be kept. In some research contexts, such as clinical trials, long term studies are important but in a commercial context organisations are more likely to be interested in analysing the most current data rather than historical records. Finally, even though the cost of storage is falling, it still represents an overhead that organisations will want to minimise.

76. It is important to remember that the data protection principles only apply to personal data and not to fully anonymised data. Anonymisation can therefore be a tool to help organisations to carry out innovative analytics or storage.

[38] Zikopoulos Paul C et al Understanding big data. Analytics for enterprise class Hadoop and streaming data. McGraw Hill, 2012. Available from: https://www14.software.ibm.com/webapp/iwm/web/signup.do?source=sw-infomgt&S_PKG=500016891&S_CMP=is_bdebook1_smarteranalytics Accessed 25 June 2014

Subject access rights

77. The DPA gives individuals powerful rights to access personal data that is being processed about them, and these rights still apply in the world of big data. People have the right to be told what personal data is being processed about them, the purposes for which it is being processed and who it may be disclosed to. They have the right to receive a copy of the information that constitutes their personal data, as well as information about the sources of that data. This right is only limited by exemptions set out in the DPA itself and in associated secondary legislation.

78. It may be thought that the volume and variety of big data and the complexity of the analytics makes it more difficult for organisations to meet this obligation. However, these cannot be an excuse for not meeting legal obligations. The existence of this right compels organisations to practice good data management. They need adequate metadata and the ability to query their data in order to find all the information they have on an individual and also to know whether the data they are processing has been truly anonymised or whether it can still be linked to an individual. A common problem for organisations historically in dealing with subject access requests is that the information is held in different places. It has been suggested to us that if an organisation's move to big data means that disparate data stores are brought together, then this may make it easier to find all the data on an individual.

79. Under section 8(2) of the DPA, data controllers do not have to give the data subject a copy of their information if supplying the copy would be impossible or involve disproportionate effort, but we consider this is only likely to arise in exceptional cases. Our subject access code of practice[39] explains this further.

80. If an organisation is using and/or buying in a range of data sources, including unstructured data, then it can be challenging to produce all the data on one individual. However, in other cases, the actual data held on an individual may not necessarily be very extensive or difficult to identify, even though the analytics applied to it are complex and extensive. For example,

[39] Information Commissioner's Office. Subject access code of practice. ICO, February 2014. http://ico.org.uk/for_organisations/data_protection/~/media/documents/library/Data_Protection/Detailed_specialist_guides/subject-access-code-of-practice.PDF Accessed 30 June 2014

the data in question may be a record of phone calls which is also available to a customer on an itemised phone bill.

81. Some organisations already make this data available to their customers on request or proactively on line, through a secure log in. In the USA, the data broker Acxiom is piloting a web portal which will enable people to see the data that is held about them for marketing purposes and the sources of that data[40]. If it is possible to make personal data available in this way, then this will help the organisation to meet its data protection obligations, and it could also help to reassure them as to the amount and type of information that is being held about them.

82. Individuals also have rights in relation to direct marketing. Organisations that use big data analytics for direct marketing to individuals should be aware that under the DPA an individual can require them to stop doing this, and they have to comply with this. There are also specific regulations relating to electronic marketing. These are explained in our Guide to the Privacy and Electronic Communications Regulations[41].

The research exemption

83. Section 33 of the DPA contains certain exemptions for when personal data is used for research. These may be relevant to big data analytics. The term 'research' is not defined in the DPA, but we consider that it can include not only historical or scientific research, but also research for commercial purposes such as market research. However, the exemptions can only apply if the research is not used to make a decision affecting an individual and if it is not likely to cause substantial damage or distress to an individual.

84. The exemptions only relate to principle 2 (incompatible purposes), principle 5 (retention) and the subject access right. This means that if personal data is obtained for one purpose and then re-used for research, the research is not an

[40] Singer, Natasha A data broker offers a peak behind the curtain. New York Times 13 August 2013. http://www.nytimes.com/2013/09/01/business/a-data-broker-offers-a-peek-behind-the-curtain.html?pagewanted=all Accessed 25 June 2014

[41] Information Commissioner's Office. Guide to the Privacy and Electronic Communications Regulations. ICO, September 2013. http://ico.org.uk/for_organisations/privacy_and_electronic_communications/~/media/documents/library/Privacy_and_electronic/Practical_application/the-guide-to-privacy-and-electronic-communications.pdf Accessed 25 June 2014

incompatible purpose. Furthermore, the research data can be kept indefinitely. The research data is also exempt from the subject access right, provided, in addition, the results are not made available in a form that identifies any individual. Our Anonymisation code of practice[42] discusses the research exemption in more detail.

85. As section 33 only provides an exemption from certain parts of the DPA, an organisation using it will still have to comply with the other data protection principles, including the duty to process personal data fairly and lawfully.

86. The research exemption is only relevant if the organisation is using personal data for research. If it uses anonymised data for research, then it is not processing personal data and the DPA does not apply to that research.

Security

87. Big data uses large volumes of data that may be held in the cloud and it may involve distributed processing across several servers. It has been suggested that the growth of big data increases the threats to the security of information. The European Union Agency for Network and Information Security (ENISA) has identified[43] a number of emerging threats arising from the potential misuse of big data by so-called 'adversaries'. These arise primarily from data breaches and leakages of information, which, it is suggested, will enable adversaries to carry out attacks to target further information. ENISA consider that the 'threat trend' is increasing in this area.

88. The ENISA report says that "uncontrolled collection, usage and dissemination of user and systems data are the perfect playground for malicious activities". This implies that a key issue is how far the growth of big data is "uncontrolled". Security depends upon the proper assessment of risk. If responsible organisations apply their normal risk management policies and procedures when they acquire new datasets or use

[42] Information Commissioner's Office. Anonymisation: managing data protection risk code of practice. ICO, November 2012. Chapter 9 http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf Accessed 25 June 2014

[43] Marinos, Louis ENISA threat landscape. ENISA, December 2013. http://www.enisa.europa.eu/activities/risk-management/evolving-threat-environment/enisa-threat-landscape-2013-overview-of-current-and-emerging-cyber-threats Accessed 25 June 2014

existing ones for big data analytics, then this should not be considered "uncontrolled". On a practical level, it is unlikely that big data analytics will be carried out in a context that is completely separate from an organisation's other data processing and analytics operations. In many cases it will be an extension of those activities, as when externally sourced datasets are combined with an organisation's in house data assets. In this way, the data will again benefit from an organisation's existing risk management policies and procedures.

89. In some cases it will be more practical to carry out big data analytics in the cloud. In this case, organisations must obtain sufficient guarantees from the cloud provider as to the security measures it has in place. In that sense, doing big data analytics in the cloud is no different to using a cloud provider for other data storage and processing. Our guidance on the use of cloud computing[44] explains the data protection issues involved.

90. There is also some evidence[45] that big data can be used to improve information security, and the ENISA report does recognise this. The ability of big data analytics to analyse very large volumes of data very quickly means that it can be used to analyse network traffic, transactions and log files that are too big to handle with other technologies in order to detect patterns and anomalies to rapidly identify security threats.

91. It is therefore too simplistic to say that big data in general either increases or mitigates information security risk. While there is a potential for increased risk, this can be mitigated by applying normal security procedures and by the use of big data in security analytics.

[44] Information Commissioner's Office. Guidance on the use of cloud computing. ICO, October 2012. http://ico.org.uk/for_organisations/data_protection/topic_guides/online/~/media/documents/library/Data_Protection/Practical_application/cloud_computing_guidance_for_organisations.ashx Accessed 25 June 2014

[45] Eg Big Data Working Group Big data analytics for security intelligence. Cloud Security Alliance, September 2013 https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf Accessed 25 June 2014; and Curry, Sam et al Big data fuels intelligence-driven security. RSA Security Brief. EMC, January 2013. http://www.emc.com/collateral/industry-overview/big-data-fuels-intelligence-driven-security-io.pdf Accessed 25 June 2014

Data controllers and data processors

92. Some organisations moving into the field of big data may want to outsource the analytics rather than doing it in-house. Big data is a rapidly developing field, and it is recognised that there is currently a shortage in the UK workforce of the skills required for big data analytics[46]. This skills shortage may have the effect of increasing the outsourcing of analytics.

93. The DPA lays down specific requirements where personal data is processed by a third party. 'Processing' in this context means storing the data as well as carrying out any analysis of it. The data controller must not only take appropriate technical and organisational measures itself to protect the data, but it must also choose a data processor that provides sufficient guarantees regarding security measures. It must also take reasonable steps to ensure that the data processor complies with those measures. There must also be a contract that specifies that the data processor is subject to the same security obligations as the data controller and acts only on the instructions of the data controller.

94. It is important to establish which organisation is the data controller and which is the data processor. In some cases, rather than a relationship between a controller and a processor, it may be that the data is being shared between two data controllers, acting jointly or in common. Our guidance document on Data processors and data controllers[47] explains this in more detail.

95. Big data analytics may represent a novel form of processing which an organisation contracts out because it does not have the skills to do in-house. The duties and safeguards in the DPA apply in this context as to any other form of personal data processing. Moreover, where an organisation is contracting out the analysis of its personal data to a third party, it is the responsibility of the organisation to ensure that the third party is applying appropriate security measures.

---

[46] HM Government. Seizing the data opportunity. A strategy for UK data capability. Department for Business Innovation and Skills, October 2013. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254136/bis-13-1250-strategy-for-uk-data-capability-v4.pdf Accessed 25 June 2014

[47] Information Commissioner's Office. Data controllers and data processors. ICO, May 2014. http://ico.org.uk/for_organisations/guidance_index/~/media/documents/library/Data_Protection/Detailed_specialist_guides/data-controllers-and-data-processors-dp-guidance.pdf Accessed 25 June 2014.

Overseas transfers

96. This paper focusses on the requirements of the DPA and the EU Data Protection Directive[48]. However, big data analytics may often be carried out internationally, with processing being done outside the EU. This may also be the case if the big data processing is done in the cloud. The eighth principle of the DPA restricts the transfer of personal data outside the European Economic Area. This issue of overseas transfers of personal data is a complex one and we have not discussed it in this paper, but general guidance on this is available in our Guide to data protection[49], which in turn links to more detailed guidance documents.

97. Our guidance document on cloud computing covers the international transfer issues raised by this technology[50].

Tools for compliance

98. In the previous sections we have discussed a number of key data protection issues in relation to big data. We now turn to some of the tools that can help to ensure that processing complies with data protection principles and that people's data privacy rights are respected.

Privacy impact assessments (PIAs)

99. Big data analytics can involve novel, complex and sometimes unexpected uses of personal data. In that context, in order to establish whether the processing is fair, it is particularly important to assess, before processing begins, to what extent it is likely to affect the individuals whose data is being used. The

---

[48] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:HTML Accessed 25 June 2014

[49] [49] Information Commissioner's Office. Guide to data protection. ICO, November 2009. http://ico.org.uk/for_organisations/data_protection/~/media/documents/library/Data_Protection/Practical_application/the_guide_to_data_protection.pdf Accessed 25 June 2014

[50] Information Commissioner's Office. Guidance on the use of cloud computing. ICO, October 2012. http://ico.org.uk/for_organisations/data_protection/topic_guides/online/~/media/documents/library/Data_Protection/Practical_application/cloud_computing_guidance_for_organisations.ashx Accessed 25 June 2014

tool to use for this analysis is a privacy impact assessment. Our code of practice on Conducting privacy impact assessments[51] gives practical advice on how to do this, and it links the privacy impact assessment to standard risk management methodologies.

100. Assessing privacy risk involves being clear at the outset about the benefits and aims of the big data project, as well as the impact on individuals' privacy. In many cases, the benefits in question are benefits to the organisation that is proposing to process the personal data, but it is important to factor in also benefits that may accrue to individuals or to society more broadly. When solutions to mitigate privacy risk have been identified, it is necessary to assess whether the final impact on those individuals, after those solutions have been applied, is proportionate to the aims of the project. Polonetsky and Tene have suggested parameters for assessing the potential benefits of big data projects[52] in particular.

101. It will be important that a range of people of involved in big data projects understand PIAs. The organisation's data protection officer may need to co-ordinate the process but other staff, such as data scientists, need to understand how to apply PIA techniques to their work. For a PIA to be effective in a big data environment those who have the technical expertise in designing and applying algorithms must have an understanding of privacy impact.

Privacy by design

102. The balance between the benefits of a project and the protection of privacy should not be seen as a zero-sum game, in which more of one means less of the other; there can be a positive sum[53]. To put is another way, it needn't be a case of the benefits of big data inevitably coming at the cost of a loss

[51] Information Commissioner's Office. Conducting privacy impact assessments code of practice. ICO February 2014, http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/pia-code-of-practice-final-draft.pdf Accessed 25 June 2014

[52] Polonetsky, Julius and Tene, Omer Privacy and big data. Making ends meet. 66 Stanford Law Review Online 25 3 September 2013. http://www.stanfordlawreview.org/online/privacy-and-big-data/privacy-and-big-data Accessed 25 June 2014

[53] Cavoukian, Ann and Jonas, Jeff. Privacy by design in the age of big data. Information and Privacy Commissioner, Ontario, Canada June 2012. http://www.ipc.on.ca/images/Resources/pbd-big_data.pdf Accessed 25 June 2014

of privacy. If a privacy risk is identified, then this can be an opportunity to find creative technical solutions that can deliver the real benefits of the project while protecting privacy.

103. Privacy by design solutions can involve not only anonymisation techniques but a range of both technical and organisational measures. These include access controls and audit logs, data minimisation, data segregation and purpose limitation and separation. These are intended to protect privacy by mitigating the risk of re-identification and the risk of data misuse. Further information on privacy by design principles and solutions is available from the Office of the Information and Privacy Commissioner of Ontario, Canada[54].

104. The privacy by design approach is about finding ways to build privacy controls into systems from the start. Cavoukian and Jonas have argued that is possible to "bake in" privacy–enhancing technologies at the outset when planning big data projects[55], while Tene and Polonetsky[56] discuss ways of obscuring the identity of individuals in a dataset. A further approach is to record individuals' preferences and corporate rules within the metadata that accompanies that data[57]. Mayer and Narayanan refer to solutions that mitigate a risk to privacy as 'privacy substitutes'[58]. These are essentially ways of carrying out transactions that only require the collection and storage of the minimum information that is actually needed to validate a customer or complete a transaction.

---

[54] http://www.privacybydesign.ca/ Accessed 18 July 2014

[55] Cavoukian, Ann and Jonas, Jeff. Privacy by design in the age of big data. Information and Privacy Commissioner, Ontario, Canada June 2012. http://www.ipc.on.ca/images/Resources/pbd-big_data.pdf Accessed 25 June 2014

[56] Tene, Omer and Polonetsky, Jules. Judged by the tin man: individual rights in the age of big data. 15 August 2013 Journal of Telecommunications and High Technology Law, Available at SSRN: http://ssrn.com/abstract=2311040 Accessed 25 June 2014

[57] Nguyen, Caroline et al. A user-centred approach to the data dilemma: context, architecture and policy. In Digital Enlightenment Yearbook 2013 –The value of personal data. Digital Enlightenment Forum September 2013. http://www.digitalenlightenment.org/publication/def-yearbook-2013-value-personal-data Accessed 25 June 2014

[58] Mayer, Jonathan and Narayanan, Arvind. Privacy substitutes. 66 Stanford Law Review Online 89 3 September 2013. http://www.stanfordlawreview.org/online/privacy-and-big-data/privacy-substitutes Accesed 25 June 2014

Transparency and privacy information

105. We have noted already that the complexity of big data analytics can mean that the processing appears opaque to citizens and consumers whose data is being used. It may not be apparent to them their data is being collected, for example their mobile phone location, or how it is being processed, for example when their search results are filtered based on an algorithm (the so called "filter bubble" effect[59]). Similarly, it may be unclear how decisions are being made about them, such as credit scoring. This opacity can lead to a lack of trust that can affect people's perceptions of, and engagement with, the organisation doing the processing. We return to the issue of building trust in the section on The business context.

106. Organisations carrying out big data analytics therefore need to think about promoting transparency at an early stage. The DPA contains a specific transparency requirement, in the form of a 'fair processing notice', or more simply a privacy notice. This is where the organisation tells people what it is going to do with their data when it collects it. It should state the identity of the organisation collecting the data, the purposes for which they intend to process it and any other information that needs to be given to enable the processing to be fair. Our Privacy notices code of practice[60] explains this further with practical examples.

107. If an organisation buys in personal data from another organisation in order to use it for big data analytics, then it also needs to ensure that the original privacy notice that was given to the individuals by the seller covers this further use of the data. If it does not, then the buyer will need to give the individuals concerned its own privacy notice, making clear the new purpose for which the data is going to be processed, and give them an opportunity to opt out.

---

[59] Pariser, Eli. Beware online "filter bubbles". TED Talk, March 2011. http://www.ted.com/talks/eli_pariser_beware_online_filter_bubbles/transcript Accessed 18 July 2014

[60] Information Commissioner's Office. Privacy notices code of practice. ICO, December 2010. http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Detailed_specialist_guides/PRIVACY_NOTICES_COP_FINAL.ashx Accessed 25 June 2014

108. It is sometimes suggested[61] that it is not feasible to give privacy notices in relation to big data analytics. This is argued on a number of grounds:

- People are unwilling to read lengthy privacy notices. When they want to download an app or purchase something on line they simply tick "I agree" without reading the conditions. This is often linked to the view that people are becoming less concerned about the use made of their data, particularly in view of the fact that they share more and more information about themselves on social media.

- The analytics used in big data, including complex algorithms, are too difficult to explain in simple terms.

- Given that big data analytics often involves repurposing data, it is not possible to foresee at the outset all the uses that may be made of the data.

We will discuss each of these arguments.

- Not reading privacy notices

109. The fact that people very often do not read privacy notices does not necessarily mean that they are unconcerned about how their data will be used. It may also indicate that they do not expect, from previous experience, that the privacy notice will give them useful information in an understandable form. The fact that there are poorly written privacy notices does not remove the responsibility on organisations to explain to customers what they are doing. Rather, it challenges them to be as innovative in this area as they are in their analytics, and to find new ways of conveying information concisely. Channel 4's use of a YouTube video[62] to accompany their privacy notice is an example of an innovative approach.

110. This is not to underestimate the growing challenges of providing privacy notices, in a context where the methods of collecting personal data are changing. A recent OECD roundtable[63] discussed a proposal for a new way of classifying

[61] Eg Rubinstein, Ira S Big data. The end of privacy or a new beginning? International Data Privacy Law 25 January 2013. http://idpl.oxfordjournals.org/content/early/2013/01/24/idpl.ips036.full.pdf+html Accessed 25 June 2014
[62] Channel 4 website http://www.channel4.com/4viewers/ Accessed 25 June 2014
[63] OECD Working party on Privacy and Security in the Digital Economy. Summary of the OECD privacy expert roundtable. Protecting privacy in a data-driven

personal data, based on how the data originated. This taxonomy distinguishes between provided, observed, derived and inferred data:

- Provided data is consciously given by individuals, eg when filling in an online form.

- Observed data is recorded automatically, eg by online cookies or sensors or CCTV linked to facial recognition.

- Derived data is produced from other data eg calculating customer profitability from the number of visits to a store and items purchased.

- Inferred data is produced by using analytics to find correlations between datasets and using these to categorise or profile people eg calculating credit scores or predicting future health outcomes.

111. Big data increasingly uses observed, derived and inferred, rather than provided data. This can be problematic in terms of providing privacy information, because individuals may be unaware that this data is being collected and processed, and the processing may be done by organisations that are not directly customer-facing. However, this does not remove the need for transparency; it is even more important because the processing is not obvious to the individuals concerned. The issue is finding the point at which to communicate this information and the most effective way to do it.

112. Privacy information does not need to be provided by just one method; a combination and mix can be used. Innovation will be needed to support different types of data collection. This will need to include consideration of in-product and just-in-time notices. There is also a strong case to consider at an early stage how this information will be provided, eg the relationship between usability and privacy by design[64].

113. Furthermore, research suggests that the view that people are unconcerned about the use of their personal data is too

---

economy: taking stock of current thinking. OECD, May 2014. http://www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=dsti/iccp/reg%282014%293&doclanguage=en Accessed 1 July 2014

[64]Information and Privacy Commissioner of Ontario. Privacy by Design and User Interfaces. http://www.ipc.on.ca/images/Resources/pbd-user-interfaces_Yahoo.pdf Accessed 26 June 2014

simplistic. Research commissioned by the International Institute of Communications[65] shows that people's willingness to give personal data, and their attitude to how that data will be used, is context-specific. That context depends on a number of variables, eg how far an individual trusts the organisation, what information is being asked for, etc. Furthermore, people use various strategies to protect their privacy when giving their personal data. The Boston Consulting Group[66] found that for 75% of consumers in most countries, the privacy of personal data remains a top issue, and that young people aged 18-24 are only slightly less cautious about the use of personal online data than older age groups.

- Complex algorithms

114. The view that it is too hard to explain the algorithms perhaps stems from a misunderstanding of the purpose of a privacy notice. The DPA does not require the privacy notice to describe *how* the data is processed (ie the technical details of how the algorithms work), but the *purposes* for which it is processed. The DPA is also clear that processing cannot be fair if people are deceived or misled about those purposes[67].

- Unforeseen purposes

115. If there is a problem about privacy notices in a big data context, it is not really about the complexity of the analytics but rather about uses of their data that individuals concerned would not expect. The ability to analyse data for different purposes, such as using the location of mobile phones to plot movements of people or traffic is an important characteristic - and a benefit - of big data analytics. If an organisation has collected personal data for one purpose and then starts to use that personal data for a completely different purpose, it needs to update its privacy notice accordingly and ensure that people are aware of this. Furthermore, the idea that in a big data context it is not possible to tell people about the possible uses of their data needs to be challenged. In general terms, big data

[65] International Institute of Communications. Personal data management: the user's perspective. International Institute of Communications, September 2012. http://www.iicom.org/open-access-resources/doc_details/226-personal-data-management-the-users-perspective Accessed 25 June 2014

[66] Rose, John et al. The trust advantage: how to win with big data. Boston Consulting Group November 2013. https://www.bcgperspectives.com/content/articles/information_technology_strategy_consumer_products_trust_advantage_win_big_data/ Accessed 25 June 2014

[67] Data Protection Act 1998 Schedule I Part II

analytics allows data to be used in innovative ways, but this does not mean that an organisation cannot foresee what use it is going to make of that data, and tell people about it.

116. If what was originally personal data is being used in an anonymised form, this does not necessarily mean that the organisation can ignore this when it is writing a privacy notice. Given the complexity of big data analytics, the processing is potentially opaque, and it may not be readily apparent to people whether their personal data is being used. It may therefore be helpful in particular cases not only to tell people what is being done with their personal data, but also to tell them when their personal data is not being used, ie if the data is anonymised for analysis. This helps to dispel some of the mystery surrounding big data, and this openness may help to build trust in the analytics.

EU General Data Protection Regulation

117. The proposed EU General Data Protection Regulation[68] contains a number of provisions that would have a bearing on the use of personal data in big data analytics. We have published our detailed comments[69] on the proposed Regulation elsewhere, here we draw out some specific points relevant to big data. In particular, these points relate to: data minimisation and the anonymised data; an onus on data controllers to justify the processing; the need for transparency; building in data protection by design and default; a shift in the balance of power; and a possible extension of data protection duties to organisations outside the EU.

118. The principle of data minimisation and the need for organisations to justify their processing of personal data emerge strongly in the proposed regulation. Personal data must be "limited to the minimum necessary in relation to the purposes for which they are processed" and shall only be

[68] European Commission. Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) COM(2012) 11 final. European Commission. 25 January 2012 http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf Accessed 25 June 2014

[69] Information Commissioner. Proposed new EU Data Protection Regulation: article-by-article analysis paper. ICO, February 2013. http://ico.org.uk/news/~/media/documents/library/Data_Protection/Research_and_reports/ico_proposed_dp_regulation_analysis_paper_20130212_pdf.ashx Accessed 25 June 2014

processed "if and as long as the purposes could not be fulfilled by processing information that does not involve personal data"(article 5 (c)). Similarly, personal data must be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed". Furthermore, the 'right to be forgotten' under article 17 means that data subjects can obtain the erasure of personal data if it is no longer necessary for the purposes for which they were collected or processed. The recent judgment of the European Court of Justice in the case of Google Spain[70] , made under the current Directive (which the DPA implements), also supports this direction of travel.

119. This challenges the idea that personal data can be collected and stored in case it might be useful in some future analysis. Instead, organisations would have to justify why they are collecting and holding personal data. Furthermore, these provisions would encourage them to use anonymised data for the analytics unless it is necessary to use data that identifies individuals.

120. The Regulation places a notable emphasis on transparency. We have already discussed the need for organisations carrying out big data analytics to be as transparent as possible, including the role of privacy notices, in the section on Transparency and privacy information. Under the proposed Regulation, the data controller would need "transparent and easily accessible policies" on processing personal data, and communicate with data subjects in "an intelligible form, using clear and plain language, adapted to the data subject" (article 11). We expect this provision would mean that, where big data analytics involve novel or unexpected processing, data controllers should actively alert people to this. The privacy notice would also be expected to contain more detail than at present, including how long the personal data will be stored and whether the data controller intends to transfer the personal data outside the European Economic Area (article 14). However, it remains to be seen how practicable it would be to communicate all of the stipulated information in some of the contexts in which big data is gathered.

[70] Google Spain SL and Google Inc v Agencia Española de Protección de Datos and González C-131/12. http://curia.europa.eu/juris/document/document.jsf?text=&docid=152065&pageIndex=0&doclang=en&mode=lst&dir=&occ=first&part=1&cid=243691 Accessed 25 June 2014

121. The Regulation also proposes practical measures to protect privacy rights. It includes a specific requirement to build in "data protection by design and default" (article 23). Data controllers would have to implement mechanisms to ensure that throughout the analytics only the minimum amount of personal data is used and is kept no longer than needed for the processing. They would also have to carry out a privacy impact assessment (article 33). This reflects some of the developments we have noted in the section on Tools for compliance and they have an important role in trying to ensure that the processing of personal data in the analytics is not unfair, excessive or unwarranted. In terms of practical measures, the Regulation also refers to data portability and to certification, which we discuss further in the section on The role of third parties.

122. Big data is sometimes characterised as a power relationship that favours corporations and governments[71]. The Regulation suggests a desire to shift the balance of power in favour of the individual by giving them more explicit rights over the processing of their personal data. The individual would have a right to object to processing carried out for certain purposes (article 19) and the right not to be subject to automated profiling which has "legal effects" on them or "significantly affects" them (article 20). What constitutes a significant effect is open to question, but these provisions are potentially relevant to the processing of personal data in big data analytics.

123. This desire is also reflected in the provisions dealing with consent as a condition for processing personal data. Under article 7, consent cannot provide a legal basis for the processing where there is a "significant imbalance" between the position of the data controller and that of the data subject. This supports the requirement that consent must be freely given (article 4), but it is arguable that an imbalance would not necessarily mean that people cannot give genuine consent.

124. The Regulation would also extend the scope of data protection to apply to data controllers outside the EU that are processing the personal data of people in the EU, if the processing relates to offering them goods or services or monitoring their behaviour (article 3). In principle this could extend the scope

[71] Richards, Neil M and King, Jonathan Three paradoxes of big data 66 Stanford Law Review Online 41 3 September 2013 http://www.stanfordlawreview.org/online/privacy-and-big-data/three-paradoxes-big-data Accessed 25 June 2014

of data protection to big data analytics that is carried out internationally and outside of the EU, for example profiling people using internet services. Nevertheless, there are questions as to what constitutes "monitoring", and moreover as to how this would be controlled in practice.

125. The ICO has advocated that these provisions in the Regulation are drafted with a risk-based approach in mind. Whilst overall the Commissioner supports the protections introduced by the Regulation, it is important to ensure the provisions are effective in practice, which includes further consideration about what this level of prescription will achieve.

126. It is clear that the Regulation is intended to address some of the key data protection issues that have been identified in relation to big data analytics, but it remains to be seen whether it will be enacted in its current form.

A challenge to data protection?

127. A number of commentators[72] have argued that the rise of big data represents a fundamental challenge to established data protection principles. In this view, the current model based on stating purposes at the outset and obtaining consent for the processing no longer works because of the complexity of the analytics and people's perceived lack of interest in how their data is used. The principles of purpose limitation and data minimisation are said not to fit with the propensity of big data to re-use data for different purposes. Furthermore, there are claims we risk losing the benefits that can be derived from big data if we attempt to confine it within an outdated framework of data protection.

128. A US study[73] also suggests that companies overestimate customers' concerns about the use of their personal data. It

---

[72] Eg Rubinstein, Ira S Big data. The end of privacy or a new beginning? International Data Privacy Law 25 January 2013. http://idpl.oxfordjournals.org/content/early/2013/01/24/idpl.ips036.full.pdf+html Accessed 25 June 2014; World Economic Forum Unlocking the value of personal data; from collection to usage. World Economic Forum February 2013 http://www3.weforum.org/docs/WEF_IT_UnlockingValuePersonalData_CollectionUsage_Report_2013.pdf Accessed 25 June 2014; Tene, Omer and Polonetsky, Julius. Big data for all; privacy and user control in the age of big data analytics 11 Northwestern Journal of Technology and Intellectual Property 239 (2013). http://scholarlycommons.law.northwestern.edu/njtip/vol11/iss5/1/ Accessed 25 June 2014

[73] Forbes Insights and Turn. The promise of privacy. Respecting consumers' limits while realizing the marketing benefits of big data. Forbes Insights 2013

claims what people are actually concerned about is not what the company is planning to do with their data but the potential misuse of that data by a 'rogue individual'. In other words, privacy concerns are really security concerns.

129. Against this background it is argued that, rather than trying to control what happens at the point when the data is collected, the emphasis should be on how the data is used, or misused. This idea was also put forward in the recent White House report on big data[74], which suggested that focussing on controlling the collection and retention of personal data may no longer be sufficient to protect personal privacy. This view emphasises the need for organisations to review internally the impact of what they are planning to do on people's privacy and the need for legal sanctions against misuse[75].

130. Our view is that the basic data protection principles already established in UK and EU law are still fit for purpose in the big data world. The view that current data protection principles are not adequate underestimates their inherent flexibility. Applying those principles involves assessing the impact of the processing on individuals and whether it is proportionate to the aim being pursued in any particular case. It is true that the current European data protection law was drawn up in the early days of the internet and it is right to look to update it to take account of how personal data is processed now. However, this does not mean that basic data protection principles are no longer fit for purpose in the big data world, or that a new data protection paradigm is required. Big data is not a game that is played by different rules.

131. In the sections on fairness, consent, purpose limitation, data minimisation and transparency above, we have set out our views on these topics. In summary, there is a challenge to organisations to be helpful and innovative in telling people what they are doing with personal data, to seek clear consent if it is required, and to recognise when they are repurposing the data for something that is different and unexpected. At the root

---

http://images.forbes.com/forbesinsights/StudyPDFs/turn_promise_of_privacy_report.pdf Accessed 25 June 2014

[74] Executive Office of the President. Big data. Seizing opportunities, preserving values. White House, May 2014. http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf Accessed 25 June 2014

[75] Mayer-Schönberger, Viktor and Cukier, Kenneth, in Chapter 9 of Big data. A revolution that will transform how we live, work and think. John Murray, 2013

of these requirements is the fundamental legal duty to be fair, which involves taking account of the users' perspective.

132. Organisations need to keep under review how their processing of personal data is impacting on people's privacy. However, this is not an alternative to carrying out the duties set by the data protection principles. Rather, carrying out a privacy impact assessment is a means to ensure that the analytics are fair in data protection terms, and helps the organisation to assess whether a new use of personal data is incompatible with the original purpose, and when it needs to seek people's consent. It is a tool to use to meet existing obligations, rather than an alternative approach.

The role of third parties

133. There has also been some discussion of the role of what has been described as 'personal data services'. This refers to technology that enables people to have easy access to the data that is held about them, so that they can re-use it themselves for their own purposes, or monetise it by making it available to other organisations in return for some reward. The proposed EU Regulation includes a provision on data portability that would enable data subjects, under certain conditions, to obtain their personal data in "an electronic and structured format which is commonly used" and transfer it to other systems (Article 18). Our Subject access code of practice[76] supports this by encouraging data controllers, when responding to subject access requests, to provide personal data in open re-usable formats where it is practicable to do so.

134. This suggests a role for intermediary organisations that can manage this data on behalf of individuals. The concept of personal data services is being developed in the UK through the government's midata initiative[77]. This is currently focused on the banking sector, mobile phone companies and energy companies. This is still at a relatively early stage, and the

[76] Information Commissioner's Office. Subject access code of practice. ICO, February 2014. http://ico.org.uk/for_organisations/data_protection/~/media/documents/library/Data_Protection/Detailed_specialist_guides/subject-access-code-of-practice.PDF Accessed 25 June 2014

[77] Department for Business Innovation and Skills. The midata vision of consumer empowerment. Press release. DBIS, 3 November 2011. https://www.gov.uk/government/news/the-midata-vision-of-consumer-empowerment Accessed 25 June 2014

government is reviewing progress. It has been suggested[78] that personal data services are a possible solution to the privacy challenges posed by big data. While there is clearly potential for this to complement existing data protection rights and to support consumer empowerment, there is not yet enough evidence from this to justify seeing personal data services as part of an alternative to the current data protection framework.

135. It has also been proposed that there is a role for a third party in ensuring that organisations carrying out big data analytics have proper regard to privacy issues. Mayer-Schönberger and Cukier[79] envisaged a role for both external and internal 'algorithmists'. These would have an audit-type role in ensuring the quality of big data applications, but they would also act as a resource for people who believe they have been harmed by big data 'predictions'. Internally they would have a role similar to that of an organisation's data protection officer. Shaw[80] suggests there are a number of missing roles in the 'big data ecosystem' that could be fulfilled by a third party, including educating and advising consumers, helping firms to obtain and share data, as well as supporting regulators and privacy and consumer organisations. The UK Anonymisation Network, referred to in the section on Anonymisation above, could also be seem as a relevant third party, since it provides advice on how to anonymise datasets, which is a key aspect of big data analytics.

136. One of the third party functions proposed by Shaw is that of the awarder of a trust mark that would show a firm uses consumer data in line with regulatory requirements. The ICO is currently looking into the feasibility of setting up a privacy seals scheme for data protection. This would certify a particular service, product or process (rather than the organisation as a whole) to show that it complies with data protection requirements. It would be awarded on the basis of rigorous testing and follow-up. We are looking to work with an established certification body, which would carry out the testing and award the privacy seal, following guidelines laid down by

[78] Rubinstein, Ira S Big data. The end of privacy or a new beginning? International Data Privacy Law 25 January 2013. http://idpl.oxfordjournals.org/content/early/2013/01/24/idpl.ips036.full.pdf+html Accessed 25 June 2014;
[79] Mayer-Schönberger, Viktor and Cukier, Kenneth, in Chapter 9 of Big data. A revolution that will transform how we live, work and think. John Murray, 2013
[80] Shaw, Duncan R. Personal big data; is there a missing third party in our emerging big data society? March 2014 http://duncanrshaw.co.uk/2014/06/27/privacy-is-there-a-missing-third-party-in-our-emerging-big-data-society-new-white-paper/ Accessed 27 June 2014

the ICO. While this was not planned specifically for the big data context, it has the potential to be applied as an assurance tool for specific big data applications. The proposed EU Regulation also encourages the establishment of data protection certification mechanisms and seals to allow people to more easily assess the level of data protection provided by data controllers and processors (article 39).

The business context

Building trust

137. There is evidence that some companies are developing an approach to big data that focusses on the impact of the analytics on individuals. This approach is not concerned solely with the capabilities of the analytics or their compliance with data protection legislation, but also looks to place big data in a wider and essentially ethical context. In other words, they are asking not only "can we do this with the data?", ie does it meet regulatory requirements, but also "should we do this with the data?" ie is it what customers expect, or should expect?

Example

Aimia[81] are a global company in the field of loyalty management and run Nectar and other loyalty programmes. As such, they deal with a very large volume of customer data. Their research showed high levels of concern amongst consumers about privacy and a desire for control over their personal data, and, contrary to a commonly expressed view, this was shared by consumers aged 19-29.

They have developed a set of data values with the acronym TACT, which stands for Transparency, Added value, Control and Trust. *Transparency* means telling customers what data is being collected, how it is being collected and how it is being used. *Added value* means making customers aware that they will receive rewards for their participation. *Control* is about giving customers control over the data they provide and enabling them to share it and to opt out. *Trust* means giving customers confidence that the data will only be used in way that you say you will use it and only share it with partners you

[81] Johnson, David and Henderson-Ross, Jeremy The new data values Aimia, 2012. http://www.aimia.com/content/dam/aimiawebsite/CaseStudiesWhitepapersResearch/english/WhitepaperUKDataValuesFINAL.pdf Accessed 25 June 2014

have identified.

Aimia describe TACT as "a philosophy" and "a matter of business culture". It is intended as a set of principles that the company should apply in all data-related business decisions.

Example

IBM have been developing an ethical framework for big data analytics[82]. They say that big data has "widened the gap between what is possible and what is legally allowed". Their ethical framework takes account of the context in which the data will be collected and used; whether people will have a choice in giving their data; whether the amount of data and what will be done with it is reasonable in terms of the application; the reliability of the data; who owns the insights to be gained from the data; whether the application is fair and equitable; the consequences of processing; people's access to the data; and accountability for mistakes and unintended consequences.

138. It is significant that these frameworks have been developed not by regulators but by companies themselves, as a response to the situation in which they find themselves in the big data world. They are approaches that are derived from the relationship between a company and its customers, and consider how to put that relationship on an equitable footing, rather than being derived from the company's need to comply with statutory obligations. Nevertheless, it is notable that many aspects of these frameworks echo key data protection principles and requirements. They reflect the importance of telling people what is being done with their data and who it will be shared with, considering whether the uses of that data are within people's expectations, giving people access to their data and some control over the use of it, and considering the impact of the analytics on the people to whom the data relates. In short, adopting an ethical approach of the type outlined in these examples will also go some way towards ensuring that

[82] Chessell, Mandy. Ethics for big data and analytics. IBM Big Data and Analytics Hub, 2014 http://www.ibmbigdatahub.com/sites/default/files/whitepapers_reports_file/TCG%20Study%20Report%20-%20Ethics%20for%20BD%26A.pdf Accessed 25 June 2014

the analytics complies with data protection principles. In particular, it helps to meet what we would see as the key issues of fairness and transparency.

139. If, as suggested here, there is a trend amongst companies towards the development of what can be seen as an ethical approach to big data analytics, it is driven by commercial imperatives as much as regulatory requirements. It would harm a company's reputation if it were the subject of a media story about the misuse of personal data, while consumers can also publicise their views to the world instantly. This is an important consideration in a competitive world. There may well be a competitive advantage in being seen as a responsible and trustworthy custodian of customer data. It is notable that responses from the public to the government's consultation on its Information Economy Strategy included the view that, "brands should be developed carefully as 'trusted custodians of customer data' "[83]. The Boston Consulting Group[84] has identified what it calls the "trust advantage":

> **Example**
>
> "Two companies in the same industry, using the same data in the same new ways, will likely achieve fundamentally different results, with the more trusted organization able to access at least five to ten times more data than the less trusted one. This, in turn, will lead to better online recommendations, more accurate targeting, faster development of new products and services, and other tangible benefits to consumers. This is the trust advantage."

140. They argue that building trust depends on two main activities: mastering data stewardship and engaging consumers. The former includes establishing guidance principles for collecting and using data and translating these into a code of conduct. The ethical frameworks developed by Aimia and IBM could be

[83] HM Government. UK government information economy strategy. A call for views and evidence. Summary of responses. Department for Business, Innovation and Skills. May 2013 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/203959/bis-13-779-information-economy-strategy-call-for-evidence-summary-of-responses.pdf Accessed 25 June 2014

[84] Rose, John et al. The trust advantage: how to win with big data. Boston Consulting Group November 2013. https://www.bcgperspectives.com/content/articles/information_technology_strategy_consumer_products_trust_advantage_win_big_data/ Accessed 25 June 2014

seen as examples of this approach. This ethical approach includes a respect for the privacy rights of individuals, and some commentators have argued that privacy itself will become a brand value.[85] Whether that is the case or not, it is evident that the development of an approach focussed on building trust can support the task of ensuring that big data analytics are compliant with data protection principles.

Information governance

141. In parallel with the development of a trust-based ethical approach, there is evidence of a growing emphasis on the issue of data quality and information governance in relation to big data analytics. Forrester Consulting[86] carried out a survey of senior executives in companies dealing with big data and asked them "what best describes how you govern big data today?" Their reported concerns included not only security, but also protection and masking of sensitive data, data quality, profiling data sources and data life cycle management, including archiving. Although the terminology here appears familiar from the data protection context, within the big data world they are primarily technical considerations. They are about how far decision makers trust the quality of the data being analysed and the validity of the insights gained from that analysis rather than how far consumers or data subjects trust the company to use their data appropriately. Nevertheless, even addressing this in a technical context inevitably addresses core data protection questions. Managing data properly and using it effectively, for example, means considering whether it identifies individuals, whether it is sensitive personal data, the circumstances in which it was collected and how long it should be retained. The Forrester report supports the view that information governance is increasingly important to big data analytics, and to be effective this has to take account of data protection and privacy issues.

[85] Satell, Greg. 5 things managers should know about the big data economy. Forbes Online 26 January 2014. http://www.forbes.com/sites/gregsatell/2014/01/26/5-things-managers-should-know-about-the-big-data-economy/ Accessed 25 June 2014

[86] Forrester Consulting. Big data needs agile information and integration governance. Forrester Research Inc, August 2013. Available from: http://www.ibmbigdatahub.com/whitepaper/big-data-needs-agile-information-and-integration-governance Accessed 25 June 2014

142. These issues are moving up the corporate hierarchy. The Boston Consulting Group has argued[87] that the potential gains from using big data are so great that the management of personal data (including issues of how data is collected, consent, purposes and security) is a "c-suite" issue, that is, an issue that should be addressed at chief officer level within an organisation. This positioning of personal data issues means that there is a convergence between the data management agenda and the data protection and privacy agenda.

143. Big data is an arena for innovation and discovery, but it is crucial to eliminate the possibility of someone in an organisation using personal data in unwarranted and intrusive ways, that would incur both regulatory action and reputational damage. Internal governance is therefore an important issue for any organisation using personal data in big data analytics. In particular the organisational relationship between those responsible for regulatory compliance and those responsible for analysing and managing data is key. Proposals for acquiring new datasets or repurposing data for big data analytics should be subject to the same risk management, security and compliance checks as any other new data initiative.

Explaining the benefits

144. Having both an appropriate values-based framework and rigorous governance arrangements in place are clearly important steps, but it is also important to educate people as citizens and as consumers. This means explaining the benefits of the analytics, in terms of improved services, more relevant marketing or enhanced rewards, and looking to foster a value exchange, in which people are happy to provide data if they are informed and have trust in how it will be used[88].

145. Given the opportunities created by big data and the commercial advantages that it can deliver to companies, there may be a tendency to take the benefits as a given. However, these are not beyond question. Google Flu Trends has often been cited as a key example of big data analytics. Initially it appeared that by analysing Google searches it was possible to produce a

[87] Dean, David et al Unleashing the value of consumer data. BCG Perspectives. Boston Consulting Group 2 January 2013. https://www.bcgperspectives.com/content/articles/digital_economy_consumer_insight_unleashing_value_of_consumer_data/ Accessed 25 June 2014

[88] Boston Consulting Group. The value of our digital identity. Liberty Global November 2012 http://www.libertyglobal.com/PDF/public-policy/The-Value-of-Our-Digital-Identity.pdf Accessed 25 June 2014

reliable indicator of the spread of flu in advance of official statistics. New research[89] has called this into question and highlighted its inaccuracies. There needs to be a realistic assessment of the benefits of big data.

146. A recent report from Sciencewise[90] showed that people have concerns about how their personal data is used by both government and companies, but in practice they provide their data because they see little alternative. Offering people a specific personal or public benefit can significantly increase their acceptance of the collection, sharing and use of their data, but they still want safeguards to be put in place, eg around misuse and security, as well as more information about how the data is used.

147. We recognise that there are many social benefits to be derived from big data analytics, eg in scientific and medical research. These are in addition to improved products and services for consumers, and the commercial advantages for companies. At the same time, these benefits should not be achieved at the cost of unfairly impacting on privacy. Data protection principles should not be seen as a barrier to progress, but as the framework to promote privacy rights and as a stimulus to developing innovative approaches to informing and engaging the public. Being transparent about the purpose and impact of the analytics is not only a legal requirement; it can also help people to be confident as 'digital citizens' in a big data world.

[89] Fung, Kaiser. Google flu trends' failure shows good data > big data HBR Blog Network 25 March 2014. http://blogs.hbr.org/2014/03/google-flu-trends-failure-shows-good-data-big-data/?utm_source=Socialflow&utm_medium=Tweet&utm_campaign=Socialflow Accessed 25 June 2014

[90] Sciencewise. Big data. Public views on the collection, sharing and use of big data by governments and companies. Sciencewise April 2014. Available from: http://www.sciencewise-erc.org.uk/cms/big-data/ Accessed 25 June 2014

Annex 1: Feedback questions

To help us plan further work in this area, we would like to receive your answers to the following questions:

1. Does this paper adequately reflect the data protection issues arising from big data or are there other relevant issues that are not covered here? If so, what are they?

2. Should the ICO produce further guidance documents to help organisations that are doing big data analytics to meet data protection requirements? If so, what should they cover?

3. This paper refers to a number of practical measures and tools that can help to protect data privacy in the context of big data analytics: anonymisation, privacy impact assessments, privacy by design, privacy notices, data portability and privacy seals. Are other practical measures and tools needed? If so, what are they?

Please email your comments to consultations@ico.org.uk by 17 October 2014, with the subject heading 'Big data and data protection paper'.

If you would like further information on this please telephone 0303 123 1113 and ask to speak to Carl Wiper or email consultations@ico.org.uk.

We may publish a summary of responses received. Information people provide in response to our consultations, including personal information, may be disclosed in accordance with the Freedom of Information Act 2000 and the Data Protection Act 1998. If you want the information that you provide to be treated as confidential please tell us, but be aware that we cannot guarantee confidentiality.

Thank you