EFFECTIVE POLICY AND REGULATION OF BIG DATA

1 Abstract

In the digital economy there is ever more data available and ever more powerful tools for analysing it. This combination of raw materials and processing power has the potential to unlock substantial economic and social welfare. But is there a risk of a public backlash against the invasion of personal privacy? This paper considers the policy and regulatory measures that need to be taken to ensure ongoing public trust and confidence in the Big Data revolution.

2 What is Big Data?

Big Data offers a new perspective on reality, and therefore will affect and shape all sectors of our economy, especially those that play a role in the capturing and/or relaying of data and information. But Big Data's likely impact is broader than the economy; it affects how our societies make sense of the world, and decide important policy challenges, and innovation[1].

Like its cousin "Net Neutrality" the term "Big Data" is widely used but seldom defined. Various attempts at a definition, or at least a description, have been made including:

- "Big data" refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyse (McKinsey, 2011)
- "Big Data" refers to large amounts of different types of data produced from various types of sources, such as people, machines or sensors (EC, 2016)
- There are many definitions of "big data" which may differ depending on whether you are a computer scientist, a financial analyst, or an entrepreneur pitching an idea to a venture capitalist. Most definitions reflect the growing technological ability to capture, aggregate, and process an ever-greater volume, velocity, and variety of data (White House, 2014).

It is clear from these descriptions that Big Data is characterised by scale, by complexity and by an ever-evolving nature. As stated in the White House paper on Big Data, "Most definitions reflect the growing technological ability to capture, aggregate, and process an ever-greater volume, velocity, and variety of data."[2] These "three Vs" of Big Data are generic characteristics that are commonly mentioned within economic literature:

[1] As examined in the 2015 edition of ITU Trends in Telecommunication Reform, chapter 4 on Big Data – Opportunities or Threat? (ITU, 2015)

[2] White House, 2014, p2

- **Volume** refers to the vast quantity of data can now be gathered, stored and analysed cost-effectively. According to Cisco for the first time in 2016 global IP traffic exceeded 1 zettabyte (that is one trillion Gigabytes) of data[3]. The almost unlimited availability of cheap cloud-based storage means that there is little incentive to delete or destroy old data, with the result that the amount of data in the world is doubling every 3-4 years.
- **Velocity** refers to the speed at which organisations can accumulate, analyse and use new data. The rate of creation of new data is increasing as subscriber numbers for computers, tablets and mobiles increases and each user makes increasing use of a variety of digital applications, in particular social media applications. Vast data banks are also created with almost every sector of society: medical records, educational courses, architectural designs, retail sales records etc. Even more significant than data created by human action is the rise of automated data creation from cameras, sensors and monitors. There is almost unlimited potential for data creation, and improved analytical techniques are enabling organisations to harness the predictive power of data, sometimes instantaneously, using deep data mining tools such as Google's *MapReduce* or the open-source *HADOOP* software framework.
- **Variety** refers to the breadth of datasets, captured in different ways and from different places. Big Data is not just about the scale of databanks, but about the sophistication of the tools that can synthesise vast quantities of vastly different data sets. The power of these data sets is greatly enhanced when they are analysed together, e.g. to establish correlations and preferences and to predict future behaviour.

Another way of looking at this is the characterization of Big Data in three words: *more, messy and correlations*[4]. It's not just the amount of data that signifies the challenge, it is the fact that so many data points are gathered from so many sources thus creating a confused picture; and the natural tendency to find links between the data can, unless a lot of care is taken in the analysis, lead to erroneous conclusions.

In summary, the challenge in defining Big Data is like that of describing a massive animal, say a rhinoceros. It is hard to describe fully, but certain characteristics are obvious, and instantly you know one when you see it. Therefore, for most of the time, a detailed description is not especially important; what matters is how it is used. Whereas the rhino is today endangered by wrongful usage, the danger with Big Data is ignoring its potential usefulness and then being overwhelmed by its volume, velocity and variety. In both cases, immediate and critical action is required of policy makers.

[3] Cisco, June 2016 – see: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.html

[4] ITU, 2015, p7

3 How and by whom is Big Data used?

Until very recently most data had a defined lifecycle. It was collected and used, then either destroyed or stored. Even when stored it would often be only temporarily, either because its usefulness declined or the data itself decayed, such were the technical limitations of storage devices (think of floppy disks and CD-ROMs). And, as with your desk, bookshelf or photo album, there comes a time when the effort of throwing out is less significant than the need to clear space for new data to arrive.

In the 21st century things are different. The era of Big Data has arrived. Each user device comes with multiple gigabytes of memory, and cloud computing offers cheap and virtually unlimited storage in vast online data centres, so there is no longer any need to delete data. It may be kept "just in case". However, as vast databanks are constructed so it becomes increasingly difficult to identify the important information held within them. Sophisticated analytical tools are needed to mine the data efficiently, and this requires a team of skilled operatives who are in short supply. All sectors of the economy are thus competing for scarce resources so as to make the most of the opportunities that Big Data creates.

Some organisations can make, and are making better, use of Big Data than others. Examples include:

- **Governments** are increasing using Big Data in crime prevention and national security. Collating information from multiple sources, such as surveillance cameras, financial transactions, phone calls, internet searches and social media, authorities can identify activities such as tax evasion and money laundering, track suspects and predict (and hopefully foil) terrorist attacks.
- **Companies** that routinely collect information about customer purchases (e.g. through loyalty cards) and use that information to make offers targeted at the individual (e.g. discount coupons for specific items). They also pay media companies such as Facebook and Google to target online advertising based on the user's pattern of internet usage – e.g. if you search the Web for "Victoria Falls" then offers of resorts and flights may appear on your Facebook feed.
- **Banks** have for many years led the way in collecting multiple data points about customers and using them to derive an aggregated credit risk score. They also use detailed analysis of financial transactions to detect and prevent potential fraud.
- **Educational establishments** are increasingly using deep data mining both in the selection of students and to identify students at risk of failing exams or dropping-out. Certain behaviours that are correlated with success or failure can be detected through data patterns and allow early intervention to be taken to resolve the potential harm.

- **Healthcare providers** use Big Data both at the individual level (e.g. to predict a person's predisposition to a particular disease, or to identify a patients propensity not to respond to a particular treatment plan) and on a societal level (e.g. to predict the outbreak or spread of an epidemic, and identify mitigating actions).

Given these, and a host of other benefits, there is no doubt that Big Data has the potential to bring very substantial welfare benefits to individuals and to society as a whole. Big Data leads to better and better-informed decisions. As the European Data Protection Supervisor has stated: "Big data, if done responsibly, can deliver significant benefits and efficiencies for society and individuals not only in health, scientific research, the environment and other specific areas."[5]

McKinsey & Co is one organisation that is a strong advocate for Big Data, and has been tracking its developments over recent years. In 2011 it published a paper entitled *Big data: the next frontier for innovation, competition and productivity*, which identified a wide range of sectors and applications where Big Data had the potential to bring significant benefits – see Figure 1 below.

Figure 1: Potential transformational benefits from Big Data

Archetype of disruption	Domains that could be disrupted
Business models enabled by orthogonal data	• Insurance • Healthcare • Human capital/talent
Hyperscale, real-time matching	• Transportation and logistics • Automotive • Smart cities and infrastructure
Radical personalisation	• Healthcare • Retail • Media • Education
Massive data integration capabilities	• Banking • Insurance • Public sector • Human capital/talent
Data-driven discovery	• Life sciences and pharmaceuticals • Material sciences • Technology
Enhanced decision-making	• Smart cities • Healthcare • Insurance • Human capital/talent

Source: McKinsey

In 2016 it published a follow-up study (McKinsey, 2016) on how well those benefits were being captured. Only a fraction of the benefits that were predicted in 2011 have, as yet, been realised – typically 20-30% in most industries, although as much as 50% of the benefits related to location-based data (e.g. in transportation and logistics) have been achieved. There are many reasons why the transformation of Big Data has not moved forward apace, including the organisational transformation that is required fully to harness its potential and the human resource challenges created by the new discipline of "analytics". But a major drag co-efficient (that McKinsey's report does not adequately address) is the

[5] EDPS, 2015, p4

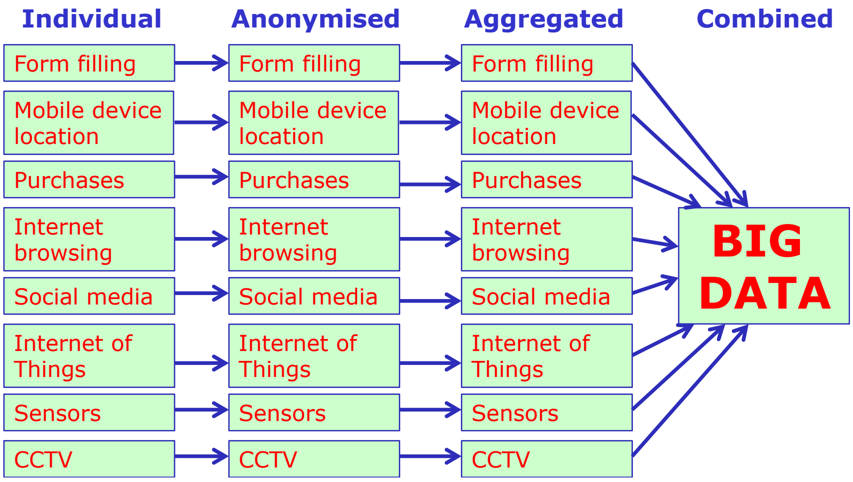
latent concerns many people and organisations have about the ethics of Big Data and its potentially deleterious impact on personal privacy. If Big Data is going to transform society for the better, these concerns need to be addressed head-on and with the utmost urgency.

4 What concerns does Big Data arouse?

The era of Big Data can easily turn into the era of Big Brother. With information gathered from so many different sources, and much of it invisible to the individual to whom the data relates, there are legitimate worries about the erosion of privacy. There is also a real concern that the advantages created by possession of Big Data, together with the cost of accessing and analyzing that data, will tend to concentrate the power of information in fewer but larger organisations. And as those organisations grow bigger, their data analytics ever more complex, how can their actions be witnessed never mind controlled?

Proponents of Big Data make re-assuring statements about anonymisation. They point out that as data is collated from a wide array of sources, so also it is stripped of details about the individual person, and analysis is only done on aggregated data so as to obtain information and make predictions about "herd" behaviour. See Figure 2.

Figure 2: How Big Data works



There is some re-assurance in these facts, but not enough to retain the widespread trust of citizens and consumers, upon which Big Data relies.

- How anonymous is it? As analytics becomes more sophisticated and as data is gathered from so many sources, there is every chance that the identity of an individual can be inferred from anonymised data. Not for nothing is the process sometimes called "pseudonymisation".
- How safe is it? Huge swathes of data are housed in the "cloud" – whether you've set your iPhone up to back up to the iCloud or whether you as an organisation

have contracted with a storage provider to back up and store your trade and customer data in their "cloud" (typically a remote data centre with connectivity between itself and all its customers), you actually don't have that data readily to hand. Accessing the data from the "cloud" means that it has to travel across third party connections again and risk decryption or corruption along the way. And what if the storage provider uses your data even if they say they won't or the data centre burns down and you don't have another copy? The Cloud Industry Forum6 was created a few years ago in much the same way and for the same purpose as the GSM Association, to set standards and agree on levels of protection for cloud storage, but it is self-regulatory.

- Is personal data for sale? It is increasingly obvious that mega-corporations such as Facebook and Google are making hay from the data they possess about their users. The cost of their free services comes in a reduction of privacy. For many users that may be acceptable, but the exchange is not transparent, the price paid neither obvious not sanctioned by the user.
- Leaks, leaks and more leaks. The top ten data leaks of the 21st Century affected over 400 million people[7] and resulted in the unlawful distribution of credit card details, account details, employee records, disability records, passwords and email addresses.
- Prediction being treated as reality. Probabilistic interpretations of Big Data lead some companies (e.g. in setting insurance premiums or offering credit) not only to predict how groups of consumers might behave, but then to engage with individual consumers as if they have already behaved in the predicted manner. This may result in unfair discrimination.

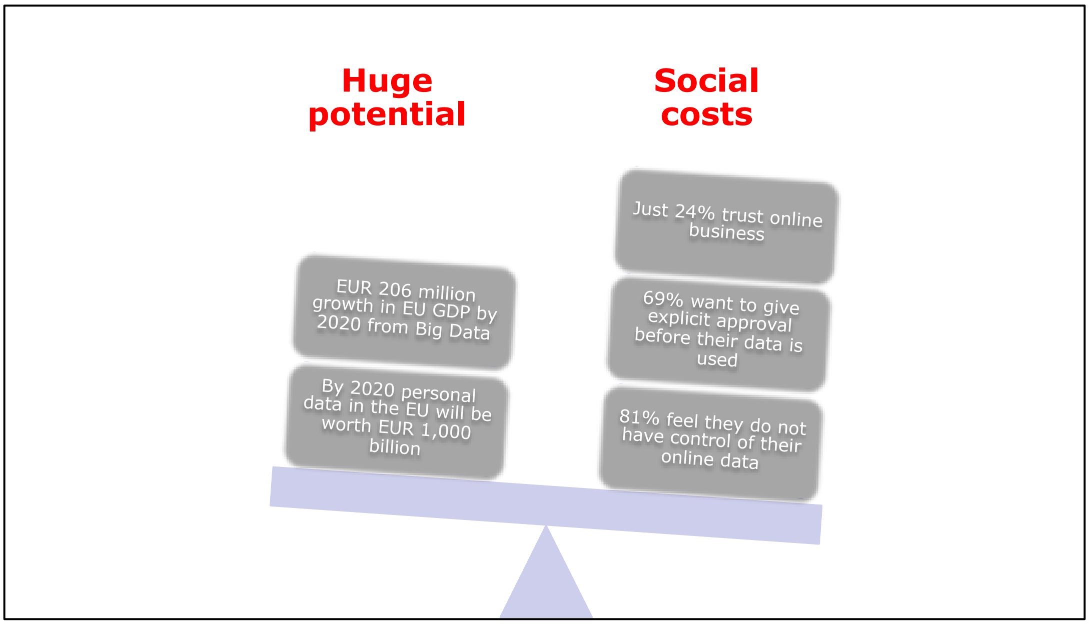
The risk is that trust in Big Data will evaporate. Already there is some evidence of consumers migrating from applications that retain data (e.g. Facebook) to others that undertake to delete data within a short time (e.g. Snapchat)[8] and there remains a sizeable proportion of the population that refuses to engage with social media on privacy grounds. If this trickle of user dissent becomes a torrent, then Big Data will become less effective and society ultimately will miss out on its benefits. The Big Data balancing act (see Figure 3) is at a tipping point; without prompt and decisive regulatory action there is a chance that societal confidence will be lost; and once lost it will be hard to regain.

[6] https://www.cloudindustryforum.org/

[7] ITU, 2015, p25

[8] ITU, 2014, p17

Figure 3: The Big Data balancing act


Huge potential
Social costs
Just 24% trust online business
EUR 206 million growth in EU GDP by 2020 from Big Data
69% want to give explicit approval before their data is used
By 2020 personal data in the EU will be worth EUR 1,000 billion
81% feel they do not have control of their online data

Source: EC, 2016

5 How can policy and regulation help?

Policy and regulation has a vital role to play in ensuring the trust of citizens and consumers in Big Data, so as to achieve its potential and avert its threat. This role should not be seen as orthogonal to the interests of industry participants; rather a well-crafted and properly-implemented data protection policy will support the Big Data industry by allowing it operate within an environment of consumer confidence. This has been recognised by authorities in both the United States and the European Union, which hope to attract inward investment in data-rich organisations precisely because of their plans for strong privacy and data protection rules. The European Commission reports that organisations such as Apple, Salesforce and Zettabox have opened data centres in the EU because of its comprehensive data protection regulations[9].

There are four main aspects to the creation of a regulatory environment in which Big Data can thrive while consumers are safeguarded:

- Ensuring data protection and privacy:
  - Data given by a consumer must be used only for the purposes authorised by the consumer

[9] EC, 2016, p2

  - User consent must be reasonably obtained – not via legalese that no user can reasonably be expected to read
  - Consumers must be given an unconditional opt-out (or/and a proactive opt-in) for any usage of personal data
  - Opt-outs should be built into the design of devices and applications, so users can easily switch their consent on or off, either permanently or temporarily.

- Informing consumers and giving them control:

  - Anonymisation in data analytics is of limited value: a person's identity may be inferred by combining supposedly anonymous data sources, often without the user's awareness
  - Users must be informed in clear and simple terms how their data will be used: "Big data analytics may create such an opaque decision-making environment that individual autonomy is lost in an impenetrable set of algorithms" (White House, 2014)
  - Users should be informed of the personal data held by the organisation, the source of that data and the logic of decision-making based on that data.

- Limiting the use if probabilistic predictions:

  - Secondary use of data for trend analysis should be <u>functionally separated</u> from any application of the data concerning the individual
  - Organisations must be accountable for how they use secondary data for research purposes, beyond simple anonymisation:
    - Consistent with purpose for which the data was given and consent obtained?
    - Is there a legitimate interest in processing the data (absent explicit user consent)?
    - Is re-use of data in a new context adequate, relevant, and proportionate?
  - Accountability is to ethic boards internally and to regulatory authorities, and ultimately courts, externally.

- Keeping data markets fluid:

  - Users should have a right to <u>data portability</u>. They should have right to access to all personal data in machine-readable format, and given the ability to modify, delete or transfer their own data. They should also be allowed to switch providers and to use third-party applications to analyse their own data.
  - Data on an individual could be held in a <u>personal data space</u>, that is a user-centric safe and secure place to store and trade personal data. This would allow individuals to participate in and potentially benefit financially from the sharing of their own data.

While some of these ideas have been proposed by the Federal Trade Commission in the USA (in particular with an eye to curbing the harmful effects of Big Data on social inclusion) and by the Executive Office of the President (especially with regard to facilitating innovation and while protecting privacy)[10], the European Union has been in the vanguard of these developments. In particular, The European Union issued Regulation 2016/679 of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (the "General Data Protection Regulation")[11] that currently sets an appropriate standard for the regulation of Big Data.

6 Collaborative methods of regulating Big Data

Big data is especially challenging to regulate because it cuts across virtually every sector of the economy and involves both private and public sectors. Regulatory solutions are therefore necessarily collaborative. All of the organisations shown in Figure 4 have to participate in defining and implementing solutions, engaging in a national and international conversation to achieve best outcomes that maximize economic welfare. All sides need to approach the challenge of Big Data with a generosity of spirit and an openness of mind, listening out for genuine concerns and eager to grasp opportunities. In this regard the ITU has a significant role to play, particularly its ITU-T Study Group 13[12] on Future networks, with focus on IMT-2020, cloud computing and trusted network infrastructures. This Study Group is responsible for studies relating to the requirements, ecosystem, and general capabilities for cloud computing and big data, as well as functional architecture for cloud computing and big data.[13].

Big Data gives rise to a host of policy and regulatory issues. Two of the most important are:

- Skills development. In much the same way as the development of ICT skills have been placed at the heart of National Broadband Plans, to ensure that economies can take full advantage of the digital economy, so also they now need to focus on enhancing skills in data analytics. This needs to be a combined effort from policy makers, educational establishment and the industry itself, with appropriate resource devoted to it.

[10] The Obama administration stated that "No matter how serious and consequential the questions posed by big data, this Administration remains committed to supporting the digital economy and the free flow of data that drives its innovation" (White House, 2014, p9), but it remains to be seen whether this policy perspective will be maintained by the Trump administration.

[11] EU, 2016

[12] http://www.itu.int/en/ITU-T/studygroups/2017-2020/13/Pages/mandate.aspx

[13] Recommendation ITU-T Y.3600 provides requirements, capabilities and use cases of cloud computing based big data as well as its system context. Cloud computing based big data provides the capabilities to collect, store, analyze, visualize and manage varieties of large volume datasets, which cannot be rapidly transferred and analysed using traditional technologies. http://www.itu.int/ITU-T/recommendations/rec.aspx?rec=12584&lang=en

- Prevention of anti-competitive behaviour. The electronic services sector has taken a lead in developing ex-ante regulation to prevent abuses of dominance that might reasonably be expected to occur and that cannot effectively be dealt with through the ex-post application of competition law. A similar two-pronged approach is necessary for Big Data, which is another market characterised by network effects. The tendency for concentration of market power – the more data you own the more power you have and the easier it is to obtain even more data – mean that there is a reasonable expectation of foreclosure, simply as a result of holding a dominant market position. Regulators need to sharpen their teeth with purveyors of Big Data, as indeed some have recently been doing[14].

Important though these issues are, they are nothing like as critical as the issues of data protection and privacy. The political earthquakes of 2016 in both the US and the UK arose in large part because rule makers were perceived as out of touch by a majority of the population. The movers and shakers of Big Data should be very careful not to make the same mistake. Privacy and the careful handling and protection of sensitive personal data are fundamental human rights.

Data protection regulation has to be at the heart of Big Data, for otherwise user confidence and trust will disappear and Big Data will implode. This is a perfectly achievable outcome for, as the European Data Protection Supervisor has said: "privacy and data protection are not in competition with economic growth and international trade, nor with great services and products - they are part of the quality and value proposition"[15].

The new General Data Protection Regulation of the European Union should therefore be seen as an important reference in Big Data regulation. It is based on five principles:

- Purpose limitation – data should be collected for a specified, explicit and legitimate purpose and should be used only for that purpose;
- Necessity – only such data that is necessary for the specified purpose should be collected, processed and stored;
- Data minimisation – no more data than is required for the stated purpose should be collected or processed, and such data should be held for the minimum time period that is required for that purpose;
- Proportionality – appropriate measures should be taken, in the context of the data processing activities and the risks associated with a breach of privacy, to ensure data protection;
- Transparency – the subject of personal data must be fully informed as to the purpose to which the data will be put and the period for which it will be stored,

[14] For example, the European Union has charged Google over abuse of dominance in the websearch market.

[15] EDPS, 2015, p9

and must be allowed access to the data so that the subject may at any time make or request change or deletion of the data.

Embedded in policy, rigorously enforced through regulation, and respected in practice by both commercial and institutional practitioners, these five principles when refined for use in other jurisdictions will provide the bedrock of a thriving Big Data industry and create the impetus for a beneficial social and economic revolution.

7 References

http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf (EU, 2016)

http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world (McKinsey, 2016)

https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf (FTC, 2016)

http://ec.europa.eu/justice/data-protection/files/data-protection-big-data_factsheet_web_en.pdf (EC, 2016)

http://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world?cid=analytics-alt-mgi-mck-oth-1612 (McKinsey, 2016)

https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Opinions/2015/15-11-19_Big_Data_EN.pdf (EDPS, 2015)

http://www.itu.int/pub/D-PREF-TTR.16-2015 (ITU, 2015)

https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf (White House, 2014)

http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (McKinsey, 2011)