The big data dilemma an inquiry by the House of Commons Select Committee on Science and Technology

Evidence from the UK Computing Research Committee

Definitive. 1 September 2015

The UK Computing Research Committee (UKCRC), an Expert Panel of the British Computer Society (BCS), the Institution of Engineering and Technology (IET) and the Council of Professors and Heads of Computing, was formed in November 2000 as a policy committee for computing research in the UK. Its members are leading computing researchers who each have an established international reputation in computing. Our response thus covers UK research in computing, which is internationally strong and vigorous, and a major national asset. This response has been prepared after a widespread consultation amongst the membership of UKCRC, and as such is an independent response on behalf of UKCRC and does not necessarily reflect the official opinion or position of the BCS or the IET.

The Committee has asked for input on five topics, which we have addressed separately below.

1. the opportunities for big data, and the risks

Big data is here and the volume and variety of data that is available will continue to expand very rapidly for the foreseeable future. It is reasonable to assume that all data that could exist will exist, probably sooner than expected and those who can see the data will continue to increase. In the end everyone may be able to see everything.

The opportunities and the risks will arise from who has access to that data and who has

- the insight to see what new, more accurate, more timely or more cost-effective information can be extracted from the data than was possible hitherto;
- the expertise to devise effective ways of extracting that information,
- the resources (including access to other data) to do so and
- the ability to interpret that data and turn it into knowledge.

The opportunities arise from beneficial uses of the knowledge and the risks from damaging uses of the knowledge. These two categories are not mutually exclusive, because actions taken may benefit some people whilst simultaneously disadvantaging or injuring others. For example, when an organisation profiles individuals or groups and offers different services or prices depending on the profile, the discrimination will affect the organisation, separate groups and individuals in ways that may be very different.

Data manipulation creates a class of serious risks. Two illustrative examples are the manipulation of the LIBOR rate, and the report at this year's DEFCON 23 that many countries have inadequate controls on the process to register a birth or a death. This could allow the creation of fake IDs or the destruction of individuals' access to credit, government services or voting, for example.

There is also the risk of selection bias. Personal data collected in a particular way (for example, online) will over-represent some groups of people and under-represent others, leading to distortions in the analysis.

Self evidently, data analytics can only work with available data, which can only come from the past. UKCRC members have differing views on the extent to which it will be possible to have confidence in the subsequent analyses of data that contains different degrees of novelty. Such is the nature of research.

It is not possible to foresee the complete range of opportunities and risks. Both categories will be very large. If it were possible to foresee all future uses of data analytics and the use and abuse of big data then there could be no scope for innovation.

2. whether the Government has set out an appropriate and up-to-date path for the continued evolution of big data and the technologies required to support it

No government has the power to determine the path that innovative science and engineering will follow, because innovation occurs in all nations and all sectors of society in ways that are too diverse (and often too powerful) to be controlled. The Government has taken some important steps that will catalyse activity in data analytics in the UK (for example, data.gov.uk, the Turing Institute and the substantial indirect investments made by government through the research councils – e.g. at Leeds there is the LIDA institute with £12M of combined MRC and ESRC funding, combining under one roof health and consumer data centres). Although most developments will take place overseas and in large companies – because that is where the vast majority of the resources (including skilled staff) will be deployed – the UK needs to continue to invest so that we can exploit advances internationally as well as contributing to, and shaping, the landscape.

3. where gaps persist in the skills needed to take advantage of the opportunities, and be protected from the risks, and how these gaps can be filled

On the supply and demand of technical skills, UKCRC members report that there is a strong demand for MSc and PhD graduates with specialisms in machine learning, statistical language processing, and related data sciences. There are currently many more jobs than there are qualified applicants and it remains to

be seen whether the new data science MSC courses that are launching will adequately satisfy the demand.

4. how public understanding of the opportunities, implications and the skills required can be improved, and 'informed consent' secured

There is a continuum between the data that is impersonal (e.g., large-scale astrophysics data), through data that has little immediate personal relevance (e.g., climate data), through data that is not personal but has personal relevance (e.g., crime map data that says something important about where you live or work, and might affect decisions and house prices), through to the personal data about finances or health that you highlight. The care.data example illustrates a broader generic point about the need to engage both citizens and organisations in the discussion about what is and is not acceptable use of data. Arguably organisations (Facebook and Google are obvious examples) invest heavily in providing services that people use and value, but they are only viable because people (knowingly or otherwise) permit those organisations to use and profit from their data. If no-one shared any data, these industries would collapse. So it's not just a question of privacy and anonymisation (important though those questions are) but also one of costs and benefits, risks and rewards.

Data ownership and right-to-use are important issues and the social and technical contexts are changing rapidly. To exercise some democratic control there is a need for education and for informed, ongoing debate. And people won't all come to exactly the same conclusion about what they are willing to share, with whom and for what purposes, now or at any time in the future.

The issue of informed consent is at the heart of the dilemma about data analytics. Most organisations' privacy and data policies are so long and so obscure that few people will ever read them. In any case, there is no individual choice – the option is to "take it or leave it". If anyone wants to use the products or services of the world's largest IT companies then they have to agree to a set of terms that few people will read, fewer will understand, and that no sensible person would accept willingly.

For consent to be really meaningful, one should be able to choose in some detail to what uses of one's data one consents and how long this consent will last . One should also be able to withdraw consent if one's circumstances or opinions change. This is gradually being recognised – the "right to be forgotten" and the proposal to allow children to delete their online indiscretions when they reach adulthood are current good examples: good because they show that the issue has been identified and good because they illustrate the inadequacy of policy objectives that cannot be implemented without a radical review of the way that data is generated, structured, owned and shared.

There are no simple solutions to this complex dilemma. It would be very helpful to have a wide and informed debate to agree the principles that a solution should incorporate.

The availability of data and the science and engineering of data analytics have developed much faster than the law or ethical frameworks. The uncertainty about the law and its inadequate enforcement have undoubtedly inhibited some commercial developments whilst allowing a degree of commercial exploitation of personal and sensitive data that has damaged public confidence and made it harder to gain consent for broadly beneficial Government policies such as care.data.

One important example is the whole matter of anonymised data. Many datasets contain information about individuals, some of which is public and some of which is private and could be very sensitive. For example, an individual's medical records will contain details of treatments, some of which are known to their family and business colleagues including time and place, and some of which they regard as private. If a database is "anonymised" in such a way that the records about an individual are linked (perhaps through a unique identifier) so that the identifier can be discovered by searching for an item of known information, then the private records can be immediately identified as well. Data analytics provide such power to combine and analyse different sources of data that it is now the case that any data about an individual should be treated as identifiable, no matter what strategies for anonymisation have been employed. This issue substantially impedes the use of existing health data for the public good, even though health is an area with relatively well developed ethical governance.

Data protection law is still evolving rapidly, through European legislation and through case law. Data usage that may be lawful in the UK today may be illegal next month or next year.

Furthermore, the medical profession has recognised that what is lawful may not be ethical – which is why they have a code of ethics that is broader than the requirements of the law. In the same way, data usage that is lawful may not be ethical.

The next EU Data Protection law seems likely to be a Regulation, with direct effect in the UK, so it is imperative that the UK plays a full part in shaping this law.

These issues would benefit from an informed, society-wide debate but it appears that few politicians understand the issues well enough to engage helpfully in such a debate, let alone to lead it. There is a need for Government to work with experts to facilitate this debate.

5. any further support needed from Government to facilitate R&D on big data, including to secure the required capital investment in big data research facilities and for their ongoing operation.

If the exploitation of big data is a revolution like the industrial revolution or the petrol engine revolution, or the IT revolution then the UK must be a leading player in the exploitation of the new world or be left behind and become a customer rather than a supplier.

We are already very strong in machine learning, computational statistics and web sciences and have strength and depth across the broad spectrum of fundamental computer science. These strengths position the UK to play a leading part in future big data research.

This is a fertile and very active field. Charles Babbage demonstrated almost 200 years ago[1] that it is necessary to understand *how and why* patterns of data are produced if you want to use them to predict the future, because any deduction from observations alone may overlook a deeper truth. The great success, for a while, of the Ptolemaic theory of planetary motion (based on fitting epicycles to observed data) is an example of how it is possible to find convincing patterns in data to confirm pre-existing prejudices. Despite the great successes to date, further data science research is essential to ensure that the information extracted from big data can be used with high and justified confidence.

There is much research yet to be done. The Internet of Things will bring new opportunities and risks, through the interconnection of millions of systems that allow actions in cyberspace to have a direct impact in the physical world.

Future research will need many different academic disciplines working together, from data scientists and economists to sociologists and ethicists and engineers. The UK has historic strengths in interdisciplinary working and investment in building these interdisciplinary research teams may create the opportunity for UK industry to take a lead in exploiting the emerging opportunities. Further investment, particularly for the growth and ongoing operation of the new centres, is a key aspect of making them sustainable. However, in the medium-long term big data analytics will become the norm, and so investment will become absorbed into everyday research funding schemes.

Martyn Thomas CBE FREng

For and on behalf of UKCRC

[1] Babbage, Charles. 1838. The Ninth Bridgewater Treatise. 2nd edn. London: John Murray.