UK COMPUTING RESEARCH COMMITTEE

UKCRC/2014/3

Commons Science & Technology Committee
Social Media Data and Real Time Analytics

UKCRC Response

The UK Computing Research Committee (UKCRC), an Expert Panel of BCS The Chartered Institute for IT, the Institution of Engineering and Technology 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 from UK academia and industry. Our evidence reflects the experience of researchers who each have an established international reputation in computing.

The terms of reference for this committee provides the following context for the consultation:

"Big data has been announced as one of the Government's eight great technologies with priorities for funding and research. …Traditional data storage systems were not designed for real-time analysis but new technologies can now provide live information and data analysis can accomplished in real-time. Social media data offers the possibility of studying social processes as they unfold at the level of populations as an alternative to traditional surveys or interviews. The data from social media is described as 'qualitative data on a quantitative scale' and requires innovative analysis techniques."

1. This gives a focus on social media data but, of course, there is not a clean separation between this form of data and other forms of data; nor is there a crisp transition between traditional data architectures and architectures specific to social media. This is an area of rapid expansion, some of it revolutionary but much of it evolutionary from existing computer systems, design practices and standards.

   How can real-time analysis of social media data benefit the UK? What should the Government be doing to maximise these benefits?

2. Government has already taken an encouraging step in encouraging open data through the release of data sets (via data.gov and other routes) and in supporting facilitators of open data (such as the Open Data Institute). This is important in itself but is also an enabler of broader social media analysis because the availability of open data feeds social media.

3. The field of social media (as it has now developed) is new, diverse and fast changing so it is difficult to define an optimal strategy to encourage its growth. We can be sure, however, that it pushes the boundaries of computing science on a variety of levels (from communications networks, through information architectures to HCI) so stimulating

research in relevant areas across the computing landscape helps to maximize benefit. This sort of research is interdisciplinary and we are already seeing deep interactions with "non-computational" disciplines (such as economics and sociology) that are leading to the formation of new research communities.

How does the UK compare to other EU countries in funding for real-time big data research?

4. All major funding organisations internationally (including the EU) are ramping up funding of data intensive research. Some of this increase is in initiatives that specifically target data research (or, even more specifically, real-time data) but a large contribution to the increase is in initiatives directed at particular industry sectors or societal problems (such as energy or healthcare). This makes it difficult to give precise, quantitative comparisons. It is, however, obvious that spending on data intensive research in Horizon 2020 will dwarf UK spending (and real-time data research is only one element of the UK computing research landscape). The issue is, then, how the UK can invest wisely in research given an across the board increase internationally.

5. Recent investments in data networks and centres by research councils (MRC, ESRC, etc) provide some stimulus for research based on real-time data analysis. There are also industry focused activities (the Catapults and regional Innovation Centres) that could provide vehicles for industry engagement. These are not set up to push the boundaries in pure research in algorithms for real-time data analytics and modelling or in developing new architectures to support this activity. The mission of existing networks and centres is primarily to do more with current technologies (in healthcare, administrative data, etc) with the advancement of core computing science being a beneficial secondary effect. Impact on domains of application is important (and getting attention) but we need also to retain UK strength in core algorithm and architecture design. We hope that the recently announced £42M Government commitment to a Turing Institute will help to address this, assuming that it can reach out across the UK.

What are the barriers to implementing real time data analysis? Is the new Government data-capability strategy sufficient to overcome these barriers?

6. The Government's report on data-capability strategy is welcome and provides a well balanced analysis of the issues and opportunites afforded by the data revolution. The issue is how to deliver fully on the actions identified by the report. In the items below we consider four actions that are of particular importance.

7. The report says that "… government will work with employers, e-skills UK, Nesta, Universities UK and the Open Data Institute to explore the skills shortages in data analytics and set out clear areas for government and industry collaboration." There is a huge gap to fill here if we are to develop the skills we need in data in the UK workforce (and remember that understanding of data is important in all areas for endeavor, not just engineering).

8. The report promises that "Universities UK will review how data analytics skills are taught across different disciplines and assess whether more work is required to further embed these skills across disciplines." Skills in data analytics are not embedded across disciplines to anything near the level needed to understand (far less exploit) the opportunities afforded by technology. A key problem is that computing science has advanced very rapidly in this area, and the approaches to "small" data familiar to those in many disciplines do not scale to "big" data problems.

9. The report informs us that "The EPSRC is developing a proposal for a national network of centres in big data analytics … The centres will develop world-leading capability and capacity in new, transformative tools and techniques to enable UK companies and the research community to be at the forefront of extracting knowledge and value from data." These centres could help us retain UK strength in core algorithm and architecture design (see earlier comment) but to do this they need to focus on core science rather than duplicating other, existing, application-oriented networks.

10. The report states that "Working with the Information Economy Council, the government will look at options to promote guidance and advice on the rights and responsibilities of data users." This must look carefully at the practical needs of data intensive research and attempt to strike a balance between legitimate regulation through data protection regulations versus the need by the academic research community for accountable but agile means of conducting empirical experiments on realistic data.

What are the ethical concerns of using personal data and how is this data anonymised for research?

11. The recent public reaction to Care.data demonstrates how vulnerable even well supported and well intentioned research initiatives can become when whey fall out of step with public opinion. There is likely always to be a tension between privacy and the pressure on researchers to publish their data sets to permit peer reanalysis of the data to confirm results, so we cannot expect simply to solve this problem and move on. We should maintain a broad view of the ethics of data sharing as technologies and cultures continue to adapt.

12. It is very difficult to provide guarantees of anonymity across data sets without restricting access to those data in some way. Even with pseudonomised data and secure means of data transfer, it is too easy to use data analytics to reveal identity through correlations between/within data sets. In recognition of this, researchers who use sensitive personal data often work in environments where their access is restricted and monitored. This is not ideal because it limits access to research data but is prudent as a starting position for developing improved, responsible methods. Recent initiatives such as the MRC Farr network for medical data sharing and the ESRC Administrative Data Research Network provide a basis for developing this agenda.

13. Despite the great work being done in opening data for research, much of the data we could analyse will not and cannot be open and much of that data currently resides with

big businesses or governments. We have to develop pragmatic approaches to analysis which recognise this and enable researchers to get more limited access to such data for the purposes of research.

14. Meanwhile, the need to manage and integrate personal data continues to grow. Many of us now possess multiple devices (smartphones, health monitors, smartwatches, etc) that store, infer and transmit personal data. Some forms of deep personal data (such as personal genome sequences) are now affordable for the affluent in society and may become commonplace. This raises difficult technical issues of maintaining and controlling access to such data and also ethical issues in its exploitation and protection. These technical and ethical strands are closely intertwined.

What impact is the upcoming EU Data Protection Legislation likely to have on access to social media data for research?

15. The impact depends on the interpretation of the legislation. Research would become very cumbersome if it were to require explicit consent from each individual to opt-in for each new form of data processing plus the right any individual to transfer or remove their data at any time from any experiment a researcher might undertake. On the other hand, people should (and, we think, will) demand greater rights of control over their personal data. A practical balance is required between the legitimate need for personal control versus the impracticality of complete control.

16. Complete control over personal data becomes difficult as the variety of our personal data expands; social media is an example of that expansion. Although a "right to be forgotten" is a laudable aim for legislation, it is very difficult to guarantee this right via conventional computing infrastructures. Data spreads fast between systems and, although its trajectories may in theory be possible to reconstruct, it is not always practical to trace where each data element goes. Data changes as it spreads – it is annotated, revised, translated and transformed – so it is seldom easy to know when we have "the same" data. For these, and other, technical reasons it is difficult to "reel back" data introduced to social media ecosystems, and difficult to attribute responsibility for so doing.