COUNCIL FOR SCIENCE AND TECHNOLOGY

Science Landscape Seminars: Big Data, E-Infrastructure and Supercomputing

Background to the meeting

This seminar is one of a series convened by the Council for Science and Technology (CST), which is working to provide a map of the UK Knowledge Landscape as a whole. This mapping includes all areas of research carried out by academia, industry, charities and others.

The seminar series has brought together diverse sets of experts to discuss eight parts of the research landscape in depth; these areas are roughly aligned with the UK government's eight great technologies.

The aim of this work is to provide decision makers with a clearer picture of the whole landscape and enable better strategic decisions to be made. We would also like the reports to prompt communities to think more about what they can do to ensure their areas continue to make the best case for themselves and operate in a coherent way. The seminar series is limited in scope, but has revealed the importance of a clear articulation of the strengths and requirements of different parts of the UK research landscape. Specific research communities may wish to hold further sessions of their own.

The discussion took place under the Chatham House rule. This document represents the views of this group and is published alongside an infrastructure document (see below) which reflects the seminar's view of the Big Data landscape.

This meeting addressed Big Data research and development, and was asked to consider:

- Strengths and weaknesses of Big Data research in the UK;
- How the UK compares internationally; and
- Future concerns that exist for the discipline.

1. Infrastructure list

To seed discussion, attendees were provided with a draft list of infrastructure relevant to Big Data, Supercomputing and e-Infrastructure. The list is not exhaustive but does provide a summary of some of the key facilities for Big Data, Supercomputing and e-Infrastructure research in the UK. It was updated in the light of discussion at the seminar to include, for instance, a number of research institutes and datasets. The infrastructure list is available at: www.gov.uk/government/publications/science-landscape-seminar-big-data-e-infrastructure-and-supercomputing.

2. Strengths and weaknesses of the UK system

The view of the seminar was that the UK has an active research scene, both in terms of fundamental data research and applied research using data-intensive methods. The UK is also active in maintaining and growing large datasets which are often made available to researchers.

To give context to the strengths and weaknesses below, the seminar made the following points:

- Data knows no borders. The international nature of the field, including a heavy reliance on international private sector companies and particularly those in the US, is an integral aspect of the UK research landscape. Vast volumes of data are produced by the internet and, like other countries, the UK is heavily reliant on US-based multi-national corporations for access to this data. Cloud computing services and other hardware used by UK research and industry are often based internationally.
- Much international infrastructure is commercial and we lack legislative control over it.
- The fast-paced development of big data, e-infrastructure and supercomputing means that our assessment of UK strengths and weaknesses may quickly become dated. The UK must constantly pay attention to developments in this field and invest appropriately so as not to fall behind.

Key strengths of the UK landscape that seminar participants identified were:

- Strong capability in social research, coding and machine learning.
- Large and open public datasets. This includes government and administrative datasets, which have been made available through data.gov.uk and the Administrative Data Research Network, as well as research data. The on-going effort to make administrative and research data open is positive; however, this openness must continue to be balanced against privacy and security concerns.

  Examples where the UK performs world-leading research as a result of these open datasets include:

  - Large remote observation and other environmental datasets, which it uses effectively in conjunction with HPC facilities and other data infrastructure. This strength is aided by good communication between research councils and research institutions.
  - High quality bioinformatics research. The UK does this through the complementary use of large volumes of public health data and other medical data such as genome sequences.

Areas where attendees felt there was a weakness, or cause for concern, were:

- Skills and the need for more data-related skills throughout the whole research landscape. *(See section 3)*
- A reliance on international and commercial data infrastructure. This may limit the UK's ability to shape future developments in the field. To avoid this we must be ready to exploit innovations and use them to our full advantage.

- There is a lack of incentives for researchers to spend time properly archiving data and follow strict data standards. This means that the value of some datasets may diminish over time.
- Access to data or infrastructure. This issue may present a particular problem for those researchers who are not already familiar with access procedures. While structures are in place to ensure that researchers *can* access data sets, the *practicalities* of actually accessing this data are often complex and difficult to navigate. Having some infrastructure around access to data and facilities may be helpful.
- Movement of data and the physical location of data. Currently, moving large volumes of data is difficult; much more so than storing it. This means that UK researchers wishing to use the larger datasets are likely to move to the data, rather than move the data to them, or move their smaller UK-based datasets abroad to combine them. This problem is shared internationally and may change with the increased uptake of cloud computing.

3. Skills

Seminar attendees noted the following points related to big data skills:

- There is a perception that computer science graduates lack the business skills required to make them employable in some sectors. Outside of academia, skills are often concentrated in government intelligence organisations and the financial sector.
- Big data, e-infrastructure and supercomputing are fast moving subject areas. It may be the case that university curricula are not able to update quickly enough to match this rapid evolution. Alternatively, university courses may be lacking direction from industry to ensure their relevance.
- SMEs often exploit cloud computing services, because the financial burden of accessing computing hardware may be too great. This places these companies at the forefront of research and development, and fostering skills in this area.
- Digital skills gaps affect research in almost every field, since data and computing techniques are now so fundamental to the way new discoveries are made.
- There is a need for flexible software skills. This enables software engineers to exploit new hardware setups as they appear.
- Currently, it is perceived that government could take a more joined-up approach to funding hardware, data and skills, which may reflect the increased need for digital skills to allow government to act as an ‘intelligent customer’.
- A lack of digital curation skills means that our current datasets risk being underused or becoming inaccessible in the future.
- In disciplines where the UK has huge, high quality datasets, researchers will need to be skilled in data analysis techniques in order to fully exploit this resource. Many subjects increasingly rely on data analytics so an increase in digital skills for researchers in all subject areas would be advantageous. One method that seminar participants suggested to achieve this was for all relevant undergraduate and postgraduate degree programmes to contain a digital skills focus, including dataset curation. An alternative approach suggested was to

encourage further collaborative working between data specialists and subject specialists, since novel collaborations are likely to yield innovative solutions to problems. The new Turing Institute may help address this issue, along with continuing work by the Knowledge Transfer Network and other bodies.

4. International comparisons

Seminar participants noted the following points:

- China, Japan and Germany are highly capable in high performance computing (HPC) and Singapore is at the forefront of Big Data investment. The US has made significant investments in both areas, as was highlighted in the 2013 Information Economy Industrial Strategy. The US also invests heavily in HPC access for industrial research and development.
- The US hosts many of the internet giants (such as Google and Facebook) and these household names dominate internet data creation. It was suggested that the logistical difficulty of moving large datasets, compared with moving smaller amounts of data to combine with a larger pre-existing datasets gives a strategic advantage to the US in this instance, whose vast datasets act as a 'magnet' for other global data.
- Equally, the UK's strong public data offering may act as a carrot for attracting new strategic collaboration.

5. Future priorities

The fast-paced and internationally competitive nature of this field demands constant re-evaluation of our Big Data capabilities. Participants considered the following points to be priorities for the future:

- Having the skills available to use and develop cloud-computing services. The recent shift towards cloud computing, where hardware and the companies which own it are usually based abroad, particularly in the US, leaves its development somewhat out of UK control. Nonetheless, these services can, and do, provide accessible high-level capability to academia and industry in the UK.
- Making effective use of healthcare data. In the UK there exists a wealth of public health and clinical data. In combination with consumer data collected through, for example, loyalty card schemes, there is an opportunity to understand and effectively influence public health policy decisions.
- Strategic decisions with regards to HPC computing. In terms of HPC capacity, seminar participants believed that the UK is currently well-positioned internationally, but will not maintain this status quo without investment. The race to the most powerful computer is always on and the UK should be strategic and decide whether it needs to be involved in this race, or whether skills development and agility to exploit the latest hardware developments are a better option.

This publication available from www.gov.uk/cst

Contacts us if you have any enquiries about this publication, including requests for alternative formats, at:

Council for Science and Technology
1 Victoria Street
London SW1H 0ET

Email: cstinfo@go-science.gsi.gov.uk