EPSRC Cross-SAT Big Data Workshop
Note of the meeting held on 5th August 2015

Attendees

Attendee	Organisation
Richard Pinch	GCHQ
Mark Girolami	University of Warwick
Dimitra Simeonidou	University of Bristol
Leigh Lapworth	Rolls Royce
Mike Chantler	Heriot Watt University
Tom McCutcheon	Dstl
Neil Chue Hong	University of Edinburgh
Sian John	Symantec Corporation
Sofia Olhede	University College London
Steven Kenny	Loughborough University
Gabriel Straub	Tesco
Alison McKay	University of Leeds
Rod Hose	University of Sheffield
Phil Darbyshire	BAE Systems Inc
Fiona Armstrong	RCUK
EPSRC Staff	**EPSRC Theme**
John Baird	PaCCS & Digital Economy
Liam Blackwell	ICT
Philippa Hemmings	Mathematical Sciences
Susan Morrell	Infrastructure
Claire Tansley	Digital Economy
Tracy Keys	Digital Economy
Chris White	Mathematical Sciences
Miriam Dowle	ICT
Iain Lamour	Infrastructure
Ruth Slade	ICT & Digital Economy
Sarah Hobbs	Healthcare Technologies
Lisa Coles	ICT

1. Welcome to the Workshop

Tracy Keys welcomed attendees to the workshop and thanked them for their participation in the "Well Sorted" pre-meeting study. It was reiterated that the purpose of the day was to identify major opportunities in Big Data research. It was timely for EPSRC, at this stage in the delivery planning process, to attempt to develop and share ideas and provide input into the Engineering and Physical Sciences challenges that exist around Big Data. The workshop provided the opportunity for an initial, relatively focussed discussion involving members from various SATs and equivalents to help provide us with insights for developing a compelling foundation of technical advice for how EPS skills and research underpin this area and are key to addressing the challenges associated with it.

2. Contextual presentations

The following presentations were given in order to provide attendees with context for the day:

i. **Data for Discovery**
Fiona Armstrong presented the RCUK-wide 'Data for Discovery' strategy, communicating the anticipated outcomes of the strategy, cross-cutting issues and rationale for implementation.

ii. **Alan Turing Institute**
Sofia Olhede talked about the Alan Turing Institute, which she has been heavily involved in so far. She discussed the scientific vision, the priorities for the institute, upcoming summits and its expected interactions.

iii. **Big Data: EPSRC Thinking on the Landscape**
Liam Blackwell discussed a range of potential EPSRC approaches for how EPSRC could contribute to Big Data, where the differences in the application domain affected the approach that should be taken. He also highlighted how Big Data would feed into the EPSRC delivery plan.

iv. **EPSRC Theme Priorities**
EPSRC theme priorities from a Big Data perspective were also briefly discussed. It was highlighted that there are elements of mathematical sciences, ICT and infrastructure that are developing methods to tackle Big Data problems, however there are also areas of EPS and beyond that primarily present problems to solve. It was recognised that, although the workshop may have had a strong focus on maths and ICT tools/techniques and methods, there was direct relevance to other Themes, both in EPSRC and beyond.

3. EPSRC Big Data Research Priorities

Attendees had contributed to a Well Sorted study in advance of the meeting, answering the question "*What are the top 3 opportunities for EPSRC research in Big Data in the next 5 years?*". The sorting from this study was used to facilitate group thinking around Big Data research priorities that EPSRC could sensibly focus on. The outputs of the session were later expanded by 'research priority convenors' identified during the meeting. These reports can be found in Annex 1.

4. Approaches to Big Data

During this session, attendees were asked to critically assess the potential EPSRC approaches that had been suggested earlier in the day (detailed below). Attendees identified relevant research priorities and application areas, as well as expressing advantages, potentials and concerns for each approach. The outputs of this session can be found in Annex 2.

Four main approaches were identified for how EPS could contribute, where the differences in the application domain affect the approach that should be taken:

- Sectors which need to improve their analytics capabilities to capture and exploit their data;
- Disciplines which need improve their own skills in data analytics to get the most benefit from their data;
- Domains where there are major changes in the non EPS aspects alongside advances in EPS relating to data analytics;
- As yet undefined and un-thought of research fields resulting from combining analytics and data in new ways.

5. Connected Nation

This session focussed on obtaining attendees' input to the proposed "Connected Nation" outcome as part of the delivery plan.

6. Close of Workshop

Attendees were thanked for their participation and engagement throughout the day. The next steps for the workshop outputs were clarified as follows. Big Data research priorities, once fleshed out by research priority convenors, would feed into the delivery plan, as well as other future EPSRC activities in this space. The higher-level delivery plan outcomes would be discussed with SATs and equivalents, and the resulting outcomes fed into SAN.

Annex 1. EPSRC Big Data Research Priorities

TIPS for Big Data (Group A)

Research Priority Convenor

- Richard Pinch

Members of group at workshop

- Neil Chue Hong
- Phil Darbyshire
- Sian John

Summary of Research Priority and challenges within it

Trust, Identity, Privacy and Security:

- Developing the underlying concepts and techniques
- Classifying requirements at an abstract level
- Developing efficient and effective methods at scale

What needs to be done now, near, next?

NOW (Current state of play):

- Anonymisation techniques for people not data; pseudonymisation in data; management of distributed identities
- Controlled access to data; scale up from file level classification and policies; preserving efficiency of access
- Fully Homomorphic Encryption – are we there yet?

NEAR (Obvious opportunities, how to make them happen):

- Calculus of privacy, risk and costs
- Controlled computation: data as a service not a product vs open data
- Data assurance in the cloud: Integrity, curation, access, auditing
- Anonymisation and controlled deanonymisation
- Trust management, reasoning and calculus of risk
- Scale factors

NEXT (Exciting long-term potential):

- Identities – factors and components
- Social and cultural factors in crowd-sourcing
- Detecting manipulation and data subversion
- ESRC linkages

Contextual Data Capture and Exploitation (Group B)

Research Priority Convenor

- Leigh Lapworth

Members of group at workshop

- Mark Girolami
- Gabriel Straub

Summary of Research Priority and challenges within it

Contextualising data:

- Self-discovery of meta-data – ensuring meta-data is consistent and current over the lifetime of the data which may be several decades in some cases.
- Self-describing meta-data – particularly future-proofing and curating data.
- Anonymisation of data – how do we make commercially sensitive data available for research? Both the security of the data from hacking but also at what level does it become possible to de-anonymise the data.

Trusting data:

- Provence and pedigree of data
- Confidence and uncertainty (epistemic and aleatory) in data – making sure data values never get separated from their statistical variation.
- Self-validating data – removal of errors in data and detection of bugs in the data processing software.

Understanding data:

- Using data out of context, e.g. using manufacturing data to improve product design. Using data in domains that have differing confidence requirements.
- Standardising metadata across multiple disciplines.
- Converting data into information, particularly for consumption in different disciplines or by lay consumers, e.g. senior decision makers. “Informative metaphors to communicate complex data”
- Linking data especially across large disparate organisations – “lots of little bits of data make big data”.

Exploiting data:

- Data driven decision making – confidence in the data and the decision.

What needs to be done now, near, next?

- Focus on cross-discipline re-use of data.
    - How do I know the context of the data I have obtained from another discipline?
    - How do I know I can trust the data for my planned usage?

- What confidence can I give to the understanding I have gleaned from processing the data?

NOW (Current state of play)

- Data is: disconnected, dispersed, hidden and incomplete with unknown quality.

NEAR (Obvious opportunities, how to make them happen)

- Standardise meta-data across time, discipline, geography, sector whilst enabling decisions to be made now and in the future.

NEXT (Exciting long-term potential)

- Understanding the difference between a good decision and a good outcome.
- Archiving decisions and their rationale. Enabling retrospective analysis of decisions and rationales. Future proofing data and decisions.
- Exploration of decision spaces and modelling the complexity thereof.
- Machine learning from linked data, including discovery of dispersed or hidden data.
- Visualisation in a probabilistic world including data filtering and compression, particularly for large scale physical simulations on HPC.
- Understanding and modelling data complexity including variable degrees of trust in individual data items.
- Dynamics of data driven systems including emergent behaviour, instabilities and bifurcations.

Connecting up the data threads (Group C)

Research Priority Convenor

- Alison McKay

Members of group at workshop

- Rod Hose
- Dimitra Simeonidou

Summary of Research Priority and challenges within it

Users of big data work with threads of data where each thread is made from multiple strands. Each strand could come from highly heterogeneous data infrastructures, harvested under a wide range of conditions in terms of dynamics (real time or historical), times, agents, contexts, etc. The data itself comes in many formats and with many contexts [through data pipelines]. Key research challenges are:

- How can data strands and threads be connected to form useful data sets that meet the needs across different application domains? For example, different types and timeframes of decisions create different requirements for the data sets.
- What are commonalities across sectors and what is special to specific sectors?
- How can we use this data to create predictive models?
- How can use real time data to compose real time services (including actuation)?
- Can we create generalised frameworks for platform and data convergence, for example through generic information or semantic translation models across heterogeneous data platforms
- Could the data be used to create new forms of empirical models that could be used in new ways by human decision makers?

How can models of engineering and physical processes best be exploited to integrate with and assist in the interpretation of big data?

What needs to be done now, near, next?

NOW (Current state of play)

- Overlay platform model, disjointed data sets and federated data; industry is doing database aggregation
- Social networks for databases – a company already provides tools (in the form of a layer across database(s)) to do this as a means of establishing the provenance of data

NEAR (Obvious opportunities, how to make them happen)

- New empirical models
- Improved characterisation of systems, including human systems spanning biological, physiological, environmental and social interactions
- Finding new patterns and trends in data, including by the integration of engineering and physics-based models

- Convergence of data and analysis platforms; more generally, there are likely to be different convergence requirements through the data pipeline, e.g., from data creation (where data could be kept where it was created) through compilation of data sets
- Linking data with infrastructure and spatiotemporal issues

NEXT (Exciting long-term potential)

- New kinds of service offering built on data that was not previously accessible
- Real time data operations/programmable platforms
- We expect to gain new insights (emergent things) when we can see newly accessible data in new ways. A next step would be to build tools to allow the sea of data to be searched again for things that emerged from previous searches ... supporting new kinds of hypothesis-driven research ...
- Improved methods for data curation, and for the handling of missing, decaying and/or incorrect data and meta data

Maths, Algorithms and Machine Learning across Scales (Group D)

Research Priority Convenor

- Steven Kenny

Members of group at workshop

- Mike Chantler
- Tom McCutcheon

Summary of Research Priority and challenges within it

- Improving our understanding of what the space of algorithms is and through this understanding choice of algorithm for a particular problem
- Need to drive interest from those engaged with blue skies research in the fundamental challenges in the area
- Need an understanding of how new algorithm development is being driven

What needs to be done now, near, next?

NOW (Current state of play)

The current state of play is that there are limited silos of good practice in terms of cross-cutting research across disciplines (and within them). It is very difficult to understand the lessons coming out of these projects and how they can best be transferred to other problems.

It is currently very difficult to compare different approaches, without implementing the approach yourself and performing the testing. There is no library for big data algorithms and little to no community assessment of the impact of the algorithms.

NEAR (Obvious opportunities, how to make them happen)

One aspect that is important is understanding what algorithms are bad at, as well as what they work well for. This is essential for building a resource for researchers who are not experts in data science to help them make decisions about the correct approach for their problem. There are a number of barriers to this including publishing negative results in most fields is not something that is common for either journals or authors. In this area it was felt that this is an issue where the journal editors could be engaged with to enable this. Another approach that could enable this is through the use of research fish to encourage both positive and negative result reporting.

To further address the issue of the usage of algorithms it was suggested to build a community website where the code is available but most importantly it also allows crowd assessment to occur. This would be a useful resource for both code and algorithm selection.

One important aspect for this area is capturing the impact that the development of new algorithms and code has on both the academic and business community. We need to make it much easier to capture this and incorporate it into REF case studies, as this will drive its perceived value.

To solve the fundamental problems in the area it is essential to engage with blue skies researchers working in other fields. To enable this to occur a workshop is needed that highlights some of the deep challenges in the area with the aim of inspiring academics to study those problems.

NEXT (Exciting long-term potential)

One of the grand challenges in the area is to bring together different technique and problem owners. This has the potential for spreading best practice and discovering commonality in diverse problem spaces.

It was also felt to be important that acceptance of risk is embedded into some calls. For instance if you want someone working a different field to see whether their work could have impact on big data algorithms you need to accept this might not work, or that its impact may be different than first envisaged by the proposal. It was acknowledged that this was not an EPSRC problem as much as one with the mindset of referees, but this could be addressed through statements in the call document.

Annex 2. EPSRC Approaches to Big Data

Approach: Up-skilling sectors

Description:

Sectors such as retail, finance, manufacturing and transport want and need to exploit existing and new techniques in data analytics. To do this they need:

- Skilled people who can apply their expertise and have the ability develop new techniques suitable to the sector concerned
- The development of new tools and techniques which can be applied in their sectors
- To use these tools and techniques to improve their operations and develop new services

Which research priorities are relevant here:

- Algorithm selection
- Methodologies
- Talking to people
- Visualisation & Interaction
- Representation and visualisation of uncertainty
- Validity (how's it the right model?)
- How do you support decision making & value generation?
- Modelisation
- Cyber security
- Psychology of decision makers using tools/models/visualisation (between different sectors and between different people)

Which application areas are relevant here:

- Getting to decision makers who understand stats
- Tertiary education for up skilling sectors rather than higher
- **Data science for public policy**

Pluses:

- Better interaction & impact
- Linking of science to problem solvers/users
- Co-production
- Centres of excellence
- Data scientists don't exist, data teams do. Need courses that are split sensibly so can hire A+B+C+D=data team
- T people (wide big data analytics knowledge and deep sector knowledge) and H people (deep understanding of a bit of data analytics and deep sector knowledge but with ability to link) needed.

Potentials:

- Skill in hardware and software
- Ability to access distributed computing

- Outreach to earlier cohorts in schools to inspire and start journey sooner
- InnovateUK and KTPs
- Awareness raising within sector leaders cohort
- Upskill in ability to ask the right questions rather than answering the questions
- Public sector KTPs? Big data influence on government will be large

Concerns:

- Differences between sector needs (solution with impact – even if messy) and academic approach (journey to get best approach)
- EPSRC have indirect impact via university courses
- Lag between research and inclusion in undergrad courses
- Is this EPSRC's job? Can influence but can they dictate?
- How do you incentivise people to look sideways/at other options?
- Universities just re-badging current content as data science, they need to look at it properly.

Other comments:

- Prioritisation of sectors to up-skill, can't do them all at once.
- EPSRC influence/shape at margins of (i.e. if a proposal gets a score of six from four reviewers it's likely to be funded irrespective of research area it's in, only at the margins does balancing come into it) therefore it takes a very long time to shape if only working on the margins. Would EPSRC refuse to fund an excellent proposal if it was in a de-prioritised area?
- How do you get people to change when their answer is "good enough" already?
- Up-skill AI researchers in ethics, philosophy
- EPSRC job to build one end of bridge, not bridge itself. i.e. train people
- Delivering impact is an output & outcome of core EPSRC activities
- CASE, KTN, Industrial fellowships etc.
- EPSRC exchange fellowships
- PhDs are EPSRC drivers/influence

Approach: Up-skilling disciplines

Description:

Disciplines which produce large amounts of data can benefit from exploiting existing and new techniques in data analytics. To do this they need:

- To enhance their own data analytics skills as these are often lacking
- Data analytics practitioners who can join research teams
- The development of new tools and techniques which can be applied in their sectors

Which research priorities are relevant here:

- Methodology
- Also demonstrators – bring two domains together – identify and pull out generic methodology & then move to up skilling strategy

Which application areas are relevant here:

- All areas have data
- Text analysis for humanities

Pluses:

- Creating opportunities for virtuous cycles of big data use
- Educate people in the domain

Potentials:

- Data services unit (like statistics services unit found currently in Universities) – people show them their problems – big data problem solving unit
- Skills through whole pipeline of big data – getting, using, contextualising etc..
- Start with a small group of people who can work in both areas
- Creating opportunities for people from user disciplines to play with big data
- Co-fund to ensure right mix of skills from the top, from the start
- Research data engineer role could be created – a new cross disciplinary career pathway

Concerns:

- Need to capture things that don't work - across disciplines (people are working on classes of problems that other disciplines have shown not to work)
- Reinventing the wheel – there is a need for disciplines to do it but not too many times – this is a cross disciplinary issues
- There are a lot of open source tools already out there
- Developing the same function but in 50 different programming languages is inefficient – there is a role for best practice in programming and software
- Why is astronomy still using Fortran – How do we move people into good practice

Other comments:

- Statistics should be mandated from birth through to Uni – it should be seen as a core discipline for all data relevant subjects
- Creating new interdisciplinary problems (challenges) is a good route to getting teams with different skill sets to work together and arbitrage expertise
- There should be a route for people to tell other people what their real problem is
- Is a data analytics toolkit required – much like the complexity toolkit that exists for understanding complex problems
- There is a skills gap in software engineering
- Train undergrads in techniques – not tools – so they can work in best practice open sources platform

Approach: New Domains

Description:

In domains where profound changes are occurring in the non-data analytics aspects alongside advances in the data analytics aspects, research in all aspects needs to be done together.

To enable this we need:

- Researchers from all disciplines involved to work closely with one another and form new communities who are experts in this new domain
- Priorities within this domain which are the product of researchers from across its different flavors to shape the research and research training within it
- The new domain to span existing disciplines so that all researchers feel joint ownership of it

The balance of research and research training within the domain to reflect the need for advances across all aspects of it.

Which research priorities are relevant here:

- User empowerment to make decisions
- Adoption & Perception
- Internet of Things
- Intelligence Community

Which application areas are relevant here:

- Academic w/Business Models
- Data Science w/Psychology – why do people make bad/optimal decisions?
- Quantified Self and Digital Representation
- Data Science w/Social Science – e.g. policy impact, Trust in digital voting for referendums
- Connected Healthcare - Individualised drugs

Pluses:

- It's less difficult to get mathematicians interested in applications than the other way round – but it's still not easy!

Potentials:

- There is lots of public mistrust of this area, if we can understand the psychology then we can exploit this.

Concerns:

- Stability in funding – if we fund a new domain, must make sure it will be sustainable beyond the original funding or else it will not survive
- Will people get out of touch if they move away from their original discipline?
- There are cultural differences between disciplines which are hard to overcome – CS/Maths/Applications

- The REF is very focused on traditional disciplines and this will put people off doing something different
- Once someone has 'embedded' they don't come back – there may be a capacity problem

Other comments:

- There are two examples where this approach has already happened and a new domain has been formed:
    - Genomics – the genomics community is ahead of maths in certain areas of algorithms
    - Neuroinformatics
- Everyone is an end user in some sense
- Focusing on outcomes will bring in the right people
- Multidisciplinary domains – not parachuting in someone else

Approach: New Opportunities

Description:

- There will be areas where combining analytics and data in new ways opens up new research fields. To identify and pursue these we need:
    - To bring together researchers and users from diverse areas to develop thinking on what the new opportunities could be, and follow this up with research to explore these ideas
    - Give researchers in data analytics opportunities to embed themselves in potential application domains and try out new things so that they can explore potential new research fields

Which research priorities are relevant here:

All 4 priorities from the morning session are relevant. In addition:

- Agreeing objective functions to optimise for and implementing this eg. driverless cars – developing safety and accident procedures
- Abstraction
- Exploration of big data space
- Exploration of high dimensional non-linear data
- Presenting visualisations that allow problem owners to recognise patterns

Which application areas are relevant here:

- Data analytics where statistical (or current statistical) methods don't apply (including situations where it is inappropriate to assume a particular model)
- Engineering and Physical Sciences applications
- User-driven / personalised data environments
- Ethics/philosophy of big data
- Finance, Biology, Urban Design, policy, communities, service providers
- One time experiments – linking data; identifying gaming or when people catch on
- Validating simulations
- Testing/validating of big data social approaches

Pluses:

- We know from experience that putting people from different disciplines and sub-disciplines together does lead to new research ideas/fields
- Ever more feasible due to data availability & computing power

Potentials:

- Mixing divergent science & user cultures to stimulate new approaches
- More interaction between application areas & researchers
- Pilots / feasibility studies / time for initial discussions
- Something like a Gordon conference – generate new theories & hypotheses

Concerns:

- Needs a concentrated effort
- Needs a responsible + ethical research – easy to go down a dark path
- People are not incentivised to take risks
- Barriers between different disciplines
- Reviewing of interdisciplinary proposals

Other comments:

- Different application areas are at different stages of big data analytics