Report of the EPSRC New Approaches to Data Science Call Information Day
London, 15 November 2016

1. Background to the call

Dr Liam Blackwell, ICT Theme lead at EPSRC, set out the wider context of the call and introduced the landscape in which it exists.

The "connected nation" is central to the UK's success and prosperity and this call sits firmly within that space. Enabling a competitive and data-driven economy depends on delivering the next generation of big data capabilities. This will require novel mathematical and statistical techniques and data analytics to generate insight and value for future business and society, enabling efficient, agile responses to a rapidly changing global market.

The EPSRC, as highlighted in its 2015 high-level position statement, sees big data, and the potential it offers, as an important opportunity. It has already invested approximately £150m into this area, spread across the UK and across a number of different fields, from visualisation and databases to artificial intelligence and statistics. It also funds a research training portfolio, with centres for doctoral training. There have been other calls and priorities, including a Making Sense from Data call in 2015 and fellowship priority areas.

EPSRC has been involved in several engagement activities with the community:

- EPSRC Cross-SAT Workshop on Big Data
- RCUK Data for Discovery Workshop
- ICT Perspectives on Big Data Analytics Workshop
- Alan Turing Scoping Workshops
- Statistics and Applied Probability Review Day
- Applied Mathematics Evidence and Engagement Workshop

This call is about filling a gap by supporting new approaches to data science driven by real world challenges, with projects undertaken in close collaboration by teams of researchers from the mathematical sciences and/or ICT, together with researchers in other disciplines and end-users.

i. Scope and requirements of the call
Miriam Dowle and Michele Erat, EPSRC

Aims of the call
In a complex landscape the EPSRC wants to make a difference and support some exciting, transformative and new approaches that are driven by real world challenges. It wants to see teams of researchers joining together across maths and ICT and other disciplines, working in close collaboration with others with a real world challenge in mind. It expects to fund three to four programmes of research that should involve novel maths or ICT research (given that the call is funded primarily by the Maths, ICT and Digital Economy themes).
It wants to provide complementary support to existing investments and so applicants will need to explain how their approach will be complementary and will represent a culture change in terms of how people work together in this space.

Co-creation and users
This call seeks to support a co-creation, which aims to get people working together from the start to create solutions to real problems, rather than people creating things on their own in a room then trying to find a user at the end! A user is anyone who uses data and could be an SME, another discipline, a large company, third sector, data provider or data consumer. Summaries of users who have already contacted EPSRC to express an interest in the call will be provided.

Institutional requirements
EPSRC expects substantial support from institutions. This could be in the form of additional support for people, such as secondments, student support or short-term internships. It could be physical space, access to professional services, opportunities for wider engagement and networking, access to equipment or facilities or partnership with other institutions. There are many ways that institutions can offer support.

Assessment criteria in brief (further details in the call document)
- Fit to the scope of the call
- Quality of the research
- National importance
- Pathways to impact
- Ability of applicant to deliver the research

Timeline (further details in the call document)
- Outline call closes: 24 Jan
- Outline panel: 28 March
- Full call closes: 19 June
- Interview panel: 9 Oct

ii. Interaction with the Alan Turing Institute

Alan Wilson, chief executive of the Alan Turing Institute, provided a brief background on the structure and strategic priorities of the UK's national institute for data science.

Structure
The institute, which is based in the British Library, has five joint venture partner universities: Cambridge, Edinburgh, UCL, Oxford and Warwick, plus the EPSRC as a sixth partner, which represents both funding and a conduit through to the government. The institute is actively talking to other universities to join the academic partnership and exploring membership models that will draw many more into close working relationship with the institute.

People
There are currently three kinds of staff. Firstly, around 90 faculty fellows who are nominated by the core universities, generally on 40% contracts and spread among sub disciplines. Secondly, a dozen full-time research fellows with a further six due to be appointed. Thirdly, about 40 PhD students, of which 15 are new this year and have supervisors in the joint venture universities, and 25 are second or third-year students who spend a year in the institute, interacting with students from different discipline backgrounds.

Activities
Institute activities include workshops on specific research topics and the prestigious Turing lecture series featuring data science speakers from around the world.

Strategic priorities
The institute has identified four key methodological or foundational research priorities, and these emphasise its interdisciplinary focus: mathematical representations; inference and learning; systems and platforms; and understanding human behaviour. This foundation research is connected into a translational programme through six broad application areas: engineering; technology; defence and security; smart cities; financial services; health and wellbeing. The institute has developed strategic partnerships in four of the six areas, such as engineering systems connecting with the Lloyd's Register Foundation, technology connecting with Intel, defence with GCHQ, and financial services with HSBC.

Collaboration
The institute can offer networking, technical expertise, workshops, meetings and other ways of connecting. The faculty fellows directory on the website shows the expertise already represented in the institute. Contact: Donna Brown research@turing.ac.uk.

Q+A

Q: Can we suggest workshops in the Turing Institute premises in proposals?
A: You can do that anyway. We're happy to collaborate with the wide community that already represents data science and beyond. In terms of widening our universities base, I'm well aware that many universities have data science institutes and we need to work with that wider range.

Q: How do you imagine collaboration between the Turing Institute and proposals?
A: It could take many forms. I see it in the main as researchers working together. Researchers in university projects could connect to our faculty fellows with particular expertise that might complement a research programme. At the more technical end, if people need advice on specific technology, eg hardware, to make progress, then getting through to the Turing's strategic partners such as Intel via the Institute is one way. There are no constraints, we would need to see how it works out. Although set up as a national institute, we are relatively modestly funded and operating in an ecosystem in which there is a tremendous amount of work, not just in academia but also in industry and government and we need to find our most effective niche within that ecosystem. We need help in defining that niche.

2. Co-creation: Tales from the frontline

Two academics with experience of the co-creation process offered their thoughts: Des Higham with a specific case study and Robert Stevens with an overview of what works – and what doesn't.

i. Professor Des Higham, University of Strathclyde

Des Higham is a mathematician who has been working with a Leeds-based social marketing agency, Bloom. He's been focusing on understanding twitter activity through the mathematical lenses of modelling (statistical and mechanistic), inference/calibration, high performance computing and using tools such as numerical linear algebra, graph theory/network science, dynamical systems and functions of matrices.

The co-creation has worked well for both parties. Bloom has been able to raise questions, get rapid access to public domain research, and develop an interesting USP for its clients. Des's team has enjoyed tackling new questions, which have led to academic publications, having access to cutting edge data and interaction with social media experts (who can explain what the data is and validate what his team does). They have also been able to show pathway to impact (which helps with REF and EPSRC reporting) as Bloom has gained new clients, won industry prizes and increased its data analytics team as a result of the work, creating jobs for maths students.

ii. Professor Robert Stevens, University of Manchester

Robert Stevens is a professor of computer science but also a "recovering biochemist". Over the past 20 years he has worked on a variety of projects with biologists and medics looking at querying, managing and integrating their services and data. He ran through the main issues he has encountered while co-creating and offered some words of advice.

Robert emphasised that a common pitfall in interdisciplinary co-creation is mutual incomprehension. It is important to let go of any attachment to particular words in order to understand the language of the other side. However, while understanding each other's words and sublanguages is critical, it is equally necessary to understand motivations and needs. A common failing is to leap to a solution without understanding your users' problems. Developing mutual respect and understanding is crucial and takes effort.

Robert has learned that solutions in search of problems tend not to work. A difficulty is the assumption that computer scientists are "the people who build databases" and so effort needs to go into identifying a problem that will help the user but also spark fundamental computer science questions – achieving both at the same time is hard work and will involve compromise on both sides.

3. End users: Introducing Tesco and the FSA

i. Ben Dias, Tesco

Ben Dias is lead data scientist at Tesco and he was clear about what he would be looking for in a co-creation partnership: novelty – and the more fundamentally new the better. Tesco already employs data scientists and sponsors PhDs, so if an idea is in the literature it is easy for the organisation to access and implement it. Tesco is looking for game-changers, the ideas that will change the industry in the future.

These are likely – but not exclusively - to be in the three broad areas Ben's team works in: operational research (optimising parts of the business to reduce costs); machine learning (forecasting, predicting); and modelling and interpretation. Scalability is very important.

Tesco is interested in the call for:

- Competitive advantage.

- Building internal capability.
- Recruitment - to build profile in the data science community as there is a shortage of data scientists. Ben's team started as three people, three years ago, and is now 15, with several new vacancies coming up all the time.

In return, Tesco is offering huge amounts of data (the team works across the group so, as well as Clubcard data, there is a lot of other operational data such as supply chain data as well as data from other businesses in the Tesco group). Ben also emphasised that Tesco can offer interesting problems and genuine collaboration: "We wouldn't just give you a problem and some data. We work agile and that's all about collaboration, communication, rapidly doing stuff and changing if we need to do so."

Q+A

Q: What about issues of privacy and consent – how do you categorise and anonymise your datasets?
A: We have PII data relating to Clubcard type data and, in that case, we anonymise and aggregate it to avoid any user being identified. We take privacy very seriously. Sometimes we create a dummy/synthetic dataset at the start to test on and then explore further. In some cases we would have to test the algorithm on your behalf and then give the results to you.

Q: What about reproducible research – the EPSRC is keen on being able to replicate results, which means releasing the data?
A: One way we've got around that is with collaborators who have access to other public data sets (eg public health data). We can publish using that data and then also use the algorithm on our data without publishing it.

Q: What if you applied an anonymisation algorithm to that data, could you release it?
A: No, we're not ready to do that yet.

Q: How does your team fit into the structure of the organisation?
A: We work across the group, although we report into the chief customer officer at the moment. We are usually easily able to prioritise our projects. But if we have competing priorities between projects it can be escalated up to the Tesco Executive if required.

Q: Can you give an example of a current project?
A: Clustering customers, segmentation. We're very mindful that customer segmentation is not a fixed thing. One of our PhDs is looking at how we can dynamically segment people based on the business question we're asking. But we don't want you to come up with just another clustering algorithm, it should be something completely different. For us, data science is about changing how decisions are made in the business based on data. We task ourselves with answering the questions that the business is not even asking yet – the unknown unknowns. Our challenge is that anything you do has to have a "so what?" This is what you have to do to change the business. Another example is change point analysis. We have time-series data that, for example, we use to forecast sales, but what happens if something like the credit crunch happens? The current models that any data scientist builds look at the past and then predicts the future – but how can you get the data to adjust those predictions when the market changes significantly?

Q: When you make your predictions based on customer behaviour do you then do trials?
A: Yes, we do trials online and also in stores. Some stores are designated trial stores but we're also doing a/b testing where we find two stores that are similar and change one and then see the difference with the other.

ii. Sian Thomas, Food Standards Agency (FSA)

Sian Thomas is head of information management at the FSA, which is the government department in the UK responsible for food safety and food standards. With regard to data, it has a data-driven, agile approach but suffers from a shortage of data scientists and would welcome the expertise this call might bring.

Two examples of where the FSA would benefit from co-creation are:

1. Regulating our future. Regulation is complicated and has remained the same for a long time. At a recent parliamentary reception the chair of the agency board explained what the agency is trying to achieve: "Create a new blueprint. We will move away from a one size fits all approach, to tailored and proportionate regulation that reflects relative risk, reinforces accountability, and delivers more for public health.'" The scale and scope of this challenge is huge, and data is a key component at the centre of any successful outcome.

2. Developing a surveillance strategy for the 21st century. While the FSA is the competent authority, the majority of the current sampling programme is delivered through its local authority partners. The agency is in the process of considering how to deliver a more data-driven approach. This initiative was discussed at the agency board meeting on 23 November and will be the subject of a workshop on 29 November.

Openness is a central to the FSA's values. It is currently working on an ambitious open data programme, which will result in 95% of its data being open by end of the year. Details of its Information Asset Register and list of datasets are available. The FSA likes to work in the open and so partners would not only be able to access most of the data in the public domain and share it but also talk about it as broadly as possible. If something works with the FSA's data then it may also be possible to look to extend into other departments or organisations.

Q+A

Q: How do we use food data really badly?
A: While we have always collected data, it is not put to effective use. For example, meat data is collected every day for every animal slaughtered but it sits in databases and is rarely used. It's a system built at a time when we wrote the information on a piece of paper and over the years it's been digitised but never really transformed to a digital system. It may be that we are unaware of the alternatives or the wider value in this type of data.

Q: Do you have supply chain data?
A: The legal requirement is that food businesses keep data for one step up and one step down. There is no obligation to make FSA aware (which also means that the FSA does not have the ancillary cost of storage) but we would like to be able to access this type of data without having to hold it in a central repository.

Q: Do you have different datasets where the same organisations feature but it is hard to match them up?
A: Yes. Local authority datasets will be in three places and when I try to match up addresses of premises the best I can do is about 60%. They should match but they don't.

Q: How does the UK compare with other nations?
A: Most of the rules come from Europe so they are very similar but implementation is different.

Q: Food surveillance – do you have surveillance data in your institution for analysis?

A: Most of this is already published. We might have surveys of a particular microbe or contaminant. All of those datasets are already published on our website. They use the current surveillance system, which we're looking to improve.

4. Panel discussion

A lively panel discussion involving all of the speakers covered the skills required to be a good data scientist – a mathematics background was seen as essential – and the "people pipeline problem" of recruitment when there is a shortage of data scientists. The democratisation of data science and the enabling of citizen scientists was seen as potentially helpful, with the caution that "a little bit of knowledge can be a dangerous thing". The opportunities arising out of data science provoked a range of responses, from the chance to have an impact and make a difference to people's lives through to the potential of data fusion and the new models that could arise as we are required to think across disparate datasets, forcing mathematics and statistics to interact with each other in novel ways. Data science was acknowledged to be changing the rules of the game in many fields, not least business where it can be particularly disruptive.

5. FAQ: Questions and answers relating to the call

These are recorded in a separate document.

6. Attendees

Charith	Abhayaratne	Sheffield
Plamen	Angelov	Lancaster University
Atta	Badii	Reading
Christian	Beck	Queen Mary, University of London
Quentin	Berthet	Cambridge University
Liam	Blackwell	EPSRC
Rita	Borgo	King's College London
Jacek	Brodzki	Southampton
Donna	Brown	Alan Turing Institute
William	Browne	Bristol
Yue	Cao	Northumbria University
Rui	Carvalho	Durham University
Alison	Cleary	Strathclyde
Colm	Connaughton	Warwick
Felix	Cuadrado	Queen Mary University of London
Beatriz	de la Iglesia	East Anglia
Varuna	De Silva	Loughborough University
Ben	Dias	Tesco plc
Tom	Diethe	Bristol
Miriam	Dowle	EPSRC
Andrew	Duncan	Sussex
Clare	Dyer-Smith	Cambridge University
Michele	Erat	EPSRC
Catherine	Godbold	EPSRC
Scott	Hale	Oxford
Edwin	Hancock	York
Yulan	He	Aston University
Philippa	Hemmings	EPSRC
Des	Higham	Strathclyde University
Jeanine	Houwing-Duistermaat	Leeds
Thomas	Howard	Newcastle University
Henrik Jeldtoft	Jensen	Imperial College London
Mark	Jones	Swansea University
Eiman	Kanjo	Nottingham Trent University
Benedict	Leimkuhler	Edinburgh
Weiru	Liu	Queen's University Belfast
Panos	Louvieris	Brunel University London
Robert	Maskell	Intel
Hannah	Maytum	Lancaster University
Christophe	Mues	Southampton
Mirco	Musolesi	UCL

Hao	Ni	UCL
Harald	Oberhauser	Oxford
Iadh	Ounis	Glasgow
Stefanos	Papanicolopulos	Edinburgh
Norman	Poh	Surrey
Thibaut	Possompes	EDF Energy
Alessandro	Provetti	Birkbeck
Magnus	Rattray	Manchester
Jeremy	Singer	Glasgow
Elizabeth	Sklar	King's College London
Andrew	Smith	Nottingham
Patrick	Spens	PwC
Robert	Stevens	Manchester University
Sian	Thomas	Food Standards Agency
Hui	Wang	Ulster University
Tillman	Weyde	City University London
Roger	Whitaker	Cardiff University
Hywel	Williams	Exeter
Alan	Wilson	Alan Turing Institute
Kay	Yeong	Dyson