HOUSE OF COMMONS

Science and Technology Committee

Oral evidence: The big data dilemma, HC 468

Tuesday 1 December 2015

Ordered by the House of Commons to be published on 1 December 2015.

Written evidence from witnesses:

- Royal Statistical Society
- Department for Business, Innovation and Skills, and the Department for Culture Media & Sport

Watch the meeting

Members present: Nicola Blackwood (Chair); Victoria Borwick; Jim Dowd; Chris Green; Dr Tania Mathias; Carol Monaghan; Graham Stringer; Derek Thomas; Valerie Vaz; Matt Warman

Questions 187-265

Witnesses: **Hetan Shah**, Executive Director, Royal Statistical Society, **Paul Maltby**, Director of Data, Government Digital Service, and **Gavin Starks**, Chief Executive Officer, Open Data Institute, gave evidence.

Q187 Chair: I welcome the panel to our final session on the big data dilemma—our inquiry into the opportunities and risks of big data. Thank you for coming today. Can I start by asking Mr Maltby from the Government Digital Service about his reaction to last week's spending review? The Government announced £450 million for the Government Digital Service and said that this "will continue to act as the digital, data and technology centre for government." Can you give us some idea of what that money is going to be used for, and what difference it will make in practical terms? Give us a sense of what value you will add for that nice chunk of cash.

Paul Maltby: We were pleased to see that that was the conclusion from the spending review. It continues to fund the Government Digital Service as a core function within the Cabinet Office, but it is also there to fund a number of different programmes. Essentially, GDS is now covering three broad areas around technology, digital and data, and the spending review settlement means that there is a programme to fund the common technology services—machines that work being rolled out across Whitehall and some of the technology infrastructure that underpins that. We will be able to take forward propositions around digital platform services used as common pieces within a broader digital landscape, things like the GOV.UK Verify service and GOV.UK/pay, which mean that there is a payment engine for different digital services to be used across different Departments.

From my perspective, the important part is that we are able to take forward a government data programme in a way that is new for GDS. It means we will be able to do the three elements of our data programme around making better use of data, continuing our work on open data but introducing data science more at scale across the government system. It means that we are able to start improving the Government's data infrastructure in a way that will be transformative in how we are able to use data for digital services but also for analytics; and it means we are able to support a group of common policy objectives across government for the first time.

Q188 Chair: How joined up between Departments is the data work of the Government now compared with 2010?

Paul Maltby: I came over to GDS two months ago to lead the data programme. We have newly nominated data leaders from across Government Departments at director general or director level. We appoint to a new ministerial committee and we have various external steering groups, through the work of the ODI, to keep in touch with data users. This is the first time we have truly brought together, at a comprehensive level, leadership on data across government. It is very much focused on the new data agenda, some of the technology that underpins it and new data science skills. That is not to say there has not been consistent leadership across the analytical community in government, not just since 2010 but back through time. Serious work has gone on and continues within economic statistics, operational research and other well-known analytical professions. What this means is that we can now bring together the leadership on government in one place under one programme and do that with some intent.

Q189 Chair: We should now see progress accelerate.

Paul Maltby: Absolutely.

Q190 Chair: Mr Shah, welcome. I saw on your Twitter feed that you spoke at a data ethics workshop in my constituency on Monday. I saw this intriguing quote: "If data is the new oil, then we need to expect oil spills, data disasters, data barons…and peak data." Can you explain that for the Committee?

Hetan Shah: I am not best placed to explain that, because I was tweeting what somebody else was saying and it is not necessarily fully my view. An environmental lawyer was saying that, in the same way that oil has regulatory frameworks to deal with some of the negative externalities, if we start to use the metaphor of oil for data, we also need to think of a new regulatory framework. Regulation is always lagging behind new technology and developments, which I think is precisely the point of this workshop today. It was an Alan Turing Institute workshop, and one of the things we are hoping for is that the Alan Turing Institute will take a lead in thinking through the ethics around big data. In the US there is a council of ethics on big data, and I wonder whether the UK needs something similar to take forward this agenda.

Q191 Chair: One of the questions that has perpetually come up during the course of this inquiry is where the boundary between the opportunities and the risks is. At what point do we need to be questioning the privacy issues and the discriminatory issues? Where do you think that boundary lies?

Hetan Shah: I agree absolutely that trust is the fundamental issue. We did some research last year on what we found to be a data trust deficit. Whenever you ask the UK public how much they trust any institution, their level of trust in that institution related to their data is always lower, so there is a data trust deficit. My view is that institutions that care about privacy have voiced their concerns very well, and we must make sure that any use of data takes privacy into account by design, but we must not let that stop us using data in an innovative way. You may give me the opportunity to say more about this in a moment, but data sharing within government and the use of that asset in austere times, in a period when we have few assets, is an absolute must both within government and for the wider research community and others.

Q192 Dr Mathias: Mr Maltby, what is your opinion of Government civil servants? Do you think they have the capability to take advantage of big data? We are talking about resources but also skills and culture.

Paul Maltby: That is a critical part of what the data programme we are leading is going to take on. It is not like we have just invented statistics, maths, or indeed data visualisations. You can go back to Florence Nightingale and beyond for some fabulous examples of that. We come to this agenda with a big dose of knowledge, and there is an entirely well-established and very professional cadre of people, but the world has changed and is changing utterly the way we enjoy services in our everyday lives that are powered from data analytics and the way data work. We want to bring that transformation to government.

Where are we in government at the moment? I would say that in the non-secret parts of government there are about 100 analysts we can call data scientists, or who are able to employ data science techniques. There are a couple of hundred within our data science community of interest. These are people who turn up to meet-ups occasionally—every fortnight or so—in an online community within government and are actively interested in learning on this. Clearly, in a world where we have 7,000 existing analysts across government, this is the start of a process rather than the end of it.

Q193 Dr Mathias: Did you say 100 out of 7,000?

Paul Maltby: Out of a few thousand. I do not know exactly how many there are, but it is 100 out of perhaps 2,000. What is the optimum amount, and at what pace? I do not know the exact answer to that, but certainly more and quicker. We are working closely in the data science partnership we have framed within government, so that GDS, the Cabinet Office, the Office for National Statistics and the Government Office for Science form a cadre around this as a topic within government. Colleagues in the Office for National Statistics are taking the lead on getting our existing analysts trained up, recruiting data scientists into government and thinking about how we retain them appropriately within

that. We are at the start of the journey, but there is already good stuff happening and good examples.

Q194 Dr Mathias: You do not think there are any barriers at the moment, because you are in the process of getting people on board.

Paul Maltby: I would not necessarily make those two things the same. We are getting on with it. There is a lot of stuff we can do right now, but there are barriers, such as whether analysts have access to machines that can use the tools. That has been a barrier in the work we have done with the data science accelerator programme, where we mentor bright analysts on the programme with some of our data scientists. Giving them access to a machine that works is an important factor. As I said, there is a common technology programme service across government that is now funded from the spending review, and in the Cabinet Office, DCMS and many other Departments this has already happened; in other Departments it is now on its way.

Q195 Dr Mathias: You say you need more. Can you give a ballpark figure of what you would be aiming for as regards training those analysts? What numbers do you want? Is it half the analysts?

Paul Maltby: It is very hard to say. At some level it is being able to use the full range of analytical tools available and the libraries that come through packages like Python or R. There is no reason why in theory this should not be the default toolkit for all analysts across the professions in the course of everyday work. Whether or not for every single analyst sitting there it would make a real difference here and now and it is worth the opportunity cost of training people up is a moot point. At the moment we are at the start of the journey. In some ways it is for colleagues in ONS to work out that longer-term plan, but we are just starting, so the answer is more and faster for the moment, I think.

Q196 Chair: One of the problems raised with us by industry is digital skills shortage and problems with recruitment. Is this affecting Government as well? I would be surprised if it was not.

Paul Maltby: Retention may be. A small number of data scientists within GDS have gone on to brighter and better things. That is great news because it means we are doing something right, and it means we can show a pathway through, but we need to think about it at a more structural level. We are not having much difficulty in recruiting to those posts, if I am honest. Colleagues in ONS, the Government Statistical Service and the operational research community have recently recruited 31 data scientists into government, more typically earlier rather than later in their career. Affording people later in their career is somewhat difficult, but at the moment that is not the biggest barrier.

It is about new data scientists coming in, which is great. It is a different culture and skillset and that is why we are mentoring people together, but just from my own experience in the Cabinet Office, looking at our existing great analysts, there is no reason why they will not take that opportunity when it is presented. A number of them have gone through that process and learned to use some of the tools and techniques. There is no reason why not.

The same level of curiosity and analytical rigour comes with that. It is partly skills, partly tools and partly a bit of a mindset.

Q197 Graham Stringer: How good is local government on big data?

Paul Maltby: With the Government data programme, we are learning within GDS about how we have stretched the boundary. It is not just within the Government; we are thinking about the devolved Administrations, local government and the other parts of the state that are not central Government. They are represented on the steering group of external experts we have recently established within this field.

On data science, we have come across good examples, partly from our work on open data over the last few years. There are colleagues in Leeds, Manchester and Bristol. I think of things like the Hampshire hub where there is some interesting geospatial mapping work, using data science to visualise the local community. There are some great examples, but very much like central Government, this might be early phase. As a whole community, they might be at a slightly earlier stage, but it is hard to generalise across several hundred local authorities—it is not one single thing—but we see that as part of our programme and our relationship in the work as it goes forward.

Q198 Graham Stringer: You have envisaged working with local government on its data science accelerator programme. What does that entail?

Paul Maltby: With four GDS data scientists, we have managed to mentor them together on a six-month programme with 20 individuals so far. They bring a project and we provide some mentoring, skills and, if needed, a machine that can access the full range of tools. This is quite small beer compared with the broader package of development and training that is going to be needed for our existing analysts, but it is an interesting way of starting. For the moment this has been a central Government thing. Interestingly, some of the analysts most capable in this field who want to apply are not based in London. Our little team is based in Holborn. We are trying to get out of London for our next phase, so there could be an opportunity to do some stuff with local government in that field. We will have a look at that, but the bigger story is how we get some of the mainstream training and development for existing analysts in that space.

Q199 Graham Stringer: Getting out of London is always a good idea, but how hands on are you with local government? Are you deeply involved with them, or are you leaving them to their own devices?

Paul Maltby: I think it is somewhere between the two. Certainly central Government would not want to dictate an approach and tell people what to do. There are great analysts all over local government, and people have a different way of doing things in each area. This is not mandating an approach or a set of skills, but we are keen to work with people. For instance, as a more informal version on the open data side, on a Saturday a few weeks back we were at an open data camp, as they call it, in Manchester. It was largely a local authority audience, or people working at a local authority, maybe inside local government or outside it in civil society or businesses. That sort of interaction is relatively routine

among the cadre of individuals who are already engaged in it. As this new programme starts, the data steering group has representation from senior leaders in local government. It will be a question of how we see things developing. Things like data standards and data infrastructure do not easily stop at administrative borders, so we are keen to think about how we make the most of that.

Q200 Graham Stringer: How aware are you of the partnerships between some parts of local government to understand more about the employment or unemployment structure in their areas? Are you aware of that, and are you working with them?

Paul Maltby: Yes, and with colleagues in the Department for Business, Innovation and Skills. Some of their early data science work was in a similar field. This is one of the exciting opportunities. It is not just Government data or data held by Government. There are increasingly open sources of data elsewhere, or data that can be accessed and available if not necessarily open in the strict definition—LinkedIn and others are good examples. There is huge opportunity in that space to be able to understand better people's situational awareness—what is actually going on in my area, in my service, in my business—but also being able to use some of those data better to segment audiences and think about predictive analytics and tailor interventions. That is one of the very large potential gains in this field for public services broadly, whether central Government, local government or elsewhere.

Q201 Matt Warman: Moving on from that, in the world that you describe where every Department, every local council and every public body has a big data mentality built in at every level, what does GDS do?

Paul Maltby: Good question. We probably do something different. There is always a cutting edge. I have been in GDS for a couple of months, but I have been in the field of innovation within public services in government for a while. There is always a sense of how we test and bring in some of those new developments. It may be that people get great at doing visualisations. Then we will be pushing on how to bring in machine learning in the best way. They will be getting greater machine learning, so how do we bring in the best of the new artificial intelligence tools within that space? It is a very fast-moving agenda. The idea that we get to a steady state relatively quickly does not feel likely, but the hope is not to try to do all that stuff in the centre. That is not the ambition. The ambition is to be able to spread this capability, knowledge and skillset very broadly indeed. A world where Government Departments and public agencies had already fixed their data infrastructure in a way that made it interoperable and fluid, where it should be fluid so that they had fabulous data sites and capabilities integrated within the decision-making processes and services for citizens as a matter of course would be a fabulous thing, and it is something we are aiming towards.

Matt Warman: I think we will come to skills later.

Q202 Carol Monaghan: We have heard quite a bit about how Government are using the data. Perhaps I may turn to Mr Starks and ask how data are produced and shared. Could you tell us a bit about open data initiatives and how important they are to the economy?

Gavin Starks: There is a critical framing here around the language we are using. The emphasis is on the word "data" rather than "big". From our perspective, we look at data as infrastructure. The critical piece about our language—we believe there is a huge need for greater data literacy, not just in the public sector but in the private sector and across the country—is to understand that in order to make the most of data you need to understand how you can use it. That means how it is licensed. Is that licence closed, shared or open?

When we talk about an open licence we are referring to data that anyone can access, use and share. For example, a bus timetable is at one end of that spectrum. The more complex area, where we have significant questions to ask, is the shared data category. It does not matter whether the data are small, medium or big, or whether it is Government, commercial or personal. The data will probably sit at multiple points on the spectrum. We would like to see a greater set of initiatives around data literacy, building that up into a bigger public conversation. It is really important as we go forward that in the social contract between the state and citizen, and between companies and citizens, we say exactly what is open and why. We have very strong views that core data infrastructure should be open and owned by the state and maybe licensed out, and various commercial models can support that. In terms of what we see as the open remit, we should help to stimulate open innovation. The kind of work we can do to get the roles, policies and liabilities sorted out around the shared part of the data spectrum will help to unlock a huge amount of innovation and value in the country.

Q203 Carol Monaghan: Mr Shah, we are hearing about different datasets being held in different places. To what extent is variable quality of data across Departments presenting a barrier to having more open data? How has the administrative data research network been able to keep personal data anonymised when we combine different datasets?

Hetan Shah: It does not seem to me that variability of data quality is the key issue in terms of stopping the sharing of data within government and making it open. Francis Maude always used to make the argument that if you open up datasets the quality will increase, so there is a positive reinforcement between those two categories. One of the big problems is that there is a silo mentality within government, and different datasets are held and not shared across Departments.

To me, the single biggest opportunity is to move where other countries have gone—Canada, New Zealand and Ireland—in giving the statistical office a broad right to data access across Departments. At the moment, the Office for National Statistics cannot easily get hold of HMRC, BIS and DWP data. If it could, we would have more real-time access to what is going on around the country. The questions we policy makers are asking are about what is happening to productivity or how we are doing in terms of tourism. The more of those datasets you can link together, the more you can answer those questions. You would not have the privacy issues, because the ONS is interested only in aggregate data; they do not care about us as individuals.

The very interesting thing about Canada and New Zealand is that they have also mandated private sector data to be open to their statistical offices, and the private sector has said, "We are glad we are being put on a level playing field, because if I was volunteering my data to you I would be at a competitive disadvantage, but if we all have to give our telecoms or supermarket data it does not matter." As you know, mobile phone data now

tell us where people are, what they are doing and so on. It is about being able to build up that rich picture in a world where at the moment we have a census that tells us every 10 years what is happening. We have made a commitment to move away from that towards much more real-time data. That sort of legislation, in the mould that other more forward-thinking countries are taking, would be the right way forward.

I cannot speak for the administrative data research network; I do not represent them. My view is that they use a good model of safe settings to provide safeguards all the way through the system. Safe settings mean safe people—accredited researchers—and safe places, access to data only within a secure setting, and an inability to take the data out again. Researchers can link datasets within those secure hubs, do the research but not take it out again. It is also safe data, so there is the ability to anonymise or pseudonymise datasets as necessary. That is the right framework.

The fourth thing that needs to be there is penalties for misuse, not necessarily for the ADRN but across the board when thinking about how to safeguard people's data. The problem has been that we have not necessarily used those best practices across all initiatives. For example, the care.data initiative to try to share health data missed opportunities to learn from other initiatives like this which have good practice.

Q204 Carol Monaghan: We heard from a witness in a previous session who had concerns that when datasets were being combined it would be more difficult for them to remain anonymised. You seem to be saying that you do not share those concerns.

Hetan Shah: There is a multiple set; we must not think that anonymisation is the only safeguard. There is anonymisation and there is who has access. The research world has a very good track record of using data in a safe way. There is also the question of how you are allowed access. If it is in a safe setting and you cannot take the data away, that is a further safeguard. It is really important that, in an attempt to reduce the risk of disclosure to zero, we do not prevent all the useful things we could do with data.

The biggest complaint I hear from researchers is that they are not now able to get access to Government data. The Government have got a bit better over the last decade at sharing data across government, but they have got worse at sharing it with researchers. If we want researchers to be able to tell us what is happening in the country—people like the Institute for Fiscal Studies using DWP data—the more barriers that are put in their way, the fewer independent voices we have telling us about the state of our country.

Q205 Chair: We heard from the Information Commissioner and others that there has not been enough research done into anonymisation, and that actually there are methods of anonymisation that work but are not commonly used. Is this something you think the Government and industry should be focusing on, or is that just a red herring and they should be looking instead at safe settings?

Hetan Shah: In a way, it is both. I hope that the Alan Turing Institute will be doing the deep technical work on how to bring together the best privacy settings today, but for every project you need to think about what its purpose is and what safeguards are required. On each spectrum it is a mixture of what level of pseudonymisation or anonymisation you

need to protect people but also to allow the research to happen. If you cannot do much pseudonymisation, perhaps you need stronger legal penalties, stronger safe havens and fewer people who can access the data. Thinking about it as multiple spectrums of safeguards is the right way to go about it.

Valerie Vaz: Mr Starks, before we go on to the work that you do, thank you very much for the postcard you produced. Perhaps members of the public would like one as well.

Chair: It is tremendous.

Q206 Valerie Vaz: I think it would be helpful if you set out the background to your organisation. I was trying to find out where your money comes from and to whom you are accountable. I know that you are a not-for-profit organisation, and Sir Tim Berners-Lee has given his knowledge and know-how about the internet for nothing. Could you tell me a bit about the background to your organisation?

Gavin Starks: We are three years old. We were set up partly with a grant of £10 million over five years from the UK public sector through Innovate UK. Under our legal constitution, we are independent, non-partisan and non-profit. Our board includes Sir Tim Berners-Lee, Sir Nigel Shadbolt, Baroness Lane Fox, Neelie Kroes and other non-executives. Our focus is part mission and part to create a sustainable business. Our income this year is about 50% from Government and philanthropic funding—we have a philanthropic investor, Omidyar—and through our direct income. We charge for training and we charge for membership of our networks—we now have over 1,000 paying members of a global network—and for our research and development and advisory services. We also run start-up programmes, and there are various competition funds around them as well. We do a lot of work both in the UK and internationally, and we have been growing, roughly doubling every nine months, since we started.

Q207 Valerie Vaz: In terms of governance, who does the UK Technology Strategy Board report to? Is it one particular Government Department or a number?

Gavin Starks: We report to the Innovate UK monitoring officer. They do not have a seat on our board, but we provide them with reports on a regular basis.

Q208 Valerie Vaz: Who do Innovate UK report to?

Gavin Starks: They report to BIS.

Q209 Valerie Vaz: It is mainly BIS as opposed to the Cabinet Office, but you collaborate with different Government Departments.

Gavin Starks: We have very strong links with Government, both with Paul's office and GDS as a whole. We also have very strong links with other Departments. For example, our head of policy is seconded right now to DEFRA to help them create their data-driven strategy. This is where we see a huge need for an increase in data literacy across government, the public sector and the whole of the commercial sector as well. This is not

just about data science; that is the easy bit. The much harder bit is around all the processes, policies, standards and so on that we are helping to co-design and co-create with our partners in the commercial and public sectors.

Q210 Valerie Vaz: In terms of the work programme, who does that? Is it the Government who set it up and push you to do it, or do you do it and bid for Government work and then you push it? In terms of open data, how does that go?

Gavin Starks: We bid for Government work and for commercial work, but usually with our partners. Looking across the range of our work, we do a lot of work with different countries around the world helping them to set policies, but not directly bought by those countries. Similarly, in the UK we work directly with DEFRA. It is a secondment arrangement in that instance, but we charge for training and advisory work. We also have various points of interaction where there is no fee exchange at all. For example, we work on some of the data steering groups with the Treasury. I am co-chairing a working group there on creating an open banking standard for the UK. We hope to take that forward into a standard next year.

Our function is to help to bridge the views of the private sector, broader civil society and the public sector, convene those opinions and facilitate the conversation that leads to the best outcome for all the different actors involved. In terms of governance, we are mission-driven to help bring together those voices. We are not following the agenda of a particular Government, company or civil society organisation.

Q211 Valerie Vaz: In terms of accountability for public money and the end result, who sets that? I am conscious that sometimes you can have projects that do not have an end date.

Gavin Starks: Our remit was to try to match funds within the five-year period. Our income this year is match funding. We have a turnover of about £2.5 million this year on top of our grant funding. That grant expires effectively at the end of next year, so we have been building a sustainable business underneath to take it forward long term.

Q212 Valerie Vaz: Who owns that information? Is it just public because it is a public-private partnership?

Gavin Starks: Owns which information?

Q213 Valerie Vaz: All the research that you do.

Gavin Starks: We publish everything under open licence.

Q214 Valerie Vaz: Everybody has access to it.

Gavin Starks: Everybody has access. All of our materials, whether that is our creative outputs, our reports, the research and development, the tools, the techniques or the standards, are licensed openly for anyone to use for any purpose.

Q215 Valerie Vaz: Can you briefly touch on the national information infrastructure and the data steering group and explain how well it is going, and what they are actually doing?

Gavin Starks: From our perspective, a huge amount of time and energy is being invested in working out what our data infrastructure is and to start thinking about data as infrastructure. That is a big mindset shift. We should really be thinking about data as infrastructure in the same way as we think about roads as infrastructure. Roads help us navigate to places; data help us navigate to decisions. Those decisions need to be made by everyone. There is a lot of work to be done to work out what we would classify as data infrastructure for the country: for example, our geo-spatial information. DEFRA has just released its dataset called LiDAR, which is very detailed environmental mapping. That has helped local businesses, citizens and Government make better decisions about their built environment.

The first question is: what problems are we looking to solve? What datasets help to support them? When we look at the data infrastructure piece, we need to ask questions not just about geospatial but, for example, about the banking sector. Building on some of the earlier points, there are well-established structures and organisations like the FSA and ICO that have defined roles about how data can be used. We then need to take them forward and ask where they fit on the data spectrum, with data infrastructure as the core. A simple thing would be where all the ATMs in the country are. That should be open data, but currently it is not. That is because companies think about this information by default. Similarly, we have had a pattern with trading funds, for example, where things are considered to be closed and open. That has gone through a huge evolution. There is a huge amount more work to be done, but we have seen some very good progress, with the Met Office, for example.

Q216 Victoria Borwick: One of the obvious impacts of open data recently has been TfL releasing their data, which has enabled a whole lot of wonderful people to write various apps. When they first released their data there was a lot of concern that people were going to take over control of the trains. People totally misunderstood. How can we use that example of where data have been opened and the end result is that lots of people have come up with apps, free to the public, and it has not meant a calamity, to try to communicate why open data are useful? In my previous experience, I saw endless committees sitting round debating how dangerous all this stuff is—"Gracious me, what's going to happen? The world will come to an end." Actually, you can show how things can be controlled, and there is a way of explaining to the world that we are not trying to destroy society; we are trying to enhance our way of living. Can you make some comment on that? It follows on very well from what my colleague asked.

Gavin Starks: It is a fantastic example. The journey TfL have been on is illustrative of the journey everyone goes on, which is initially, "We have the data. We should build all the things around it; we should own the relationship with everyone." There is a better solution. A data supply chain has emerged. There is a company called Transport API that aggregates TfL data and 70% of the transport data of the country. They have thousands of developers building solutions from that. They power the screens at Heathrow.

Q217 Victoria Borwick: Hundreds of human beings.

Gavin Starks: It is a very interesting example. The principles of open data enabled them to exist and a further commercial ecosystem to exist, as well as providing more open data back to the system for the benefit of everyone. This is a process of open innovation. It is a different way of thinking and it is encouraging people to embrace network thinking. We are in a network age. Very much thanks to Sir Tim, over the last 26 years we have had 1 billion websites connected with one another. When we look forward from this point we see more people embrace that form of thinking. Everybody starts with the belief that in some way the sky will fall. It very rarely does. There are some great examples. TfL now have a report that indicates a return of more than 50 times on investment from being open instead of being closed. So there are huge economic benefits to society and social benefits in the mix.

Q218 Victoria Borwick: Paul, I saw you nodding. Do you want to add to that?

Paul Maltby: It is a great point. If you said to the public, "Who uses open data?" presumably most people would say, "I've no idea what you're on about. I'm busy. Thank you very much," but ubiquitously, in London and around the country, businesses, some of which have gone through the ODI process—some have not—have incubated and are performing in that way in a world market. Once you have that excellence and capacity, transport data in Berlin might be very similar to the transport data in London, as long as it is made available. It is a great example of what this really means. There are various examples where you see open data fuelling decision making. Sometimes it is an app on a phone, but often these days it is not; it is data being pulled in every day by larger businesses to help their decision making. Where is the supermarket located? What sort of chocolate biscuits is it selling? Those sorts of things are powered by it.

Q219 Victoria Borwick: Which road is least congested?

Paul Maltby: Indeed. Sometimes we talk about things that sound somewhere between theoretical and science fiction. The best example of this stuff is by getting on and doing it. We are not talking about the far-flung future but about a present reality. To go back to the data science work in government, we did not start with a theoretical big strategy document saying, "This is a wonderful thing. We should probably do something." We hired some people and got on with it. There was plenty of thinking that we needed to do along the way, not least, as was mentioned earlier, to make sure we had a forward-looking, forward-leaning but also ethical approach to how we deal with data science. These are powerful tools and we need to think about the application of them. Thinking and policy work certainly need to go underneath that, but the best way to explain what we need is by getting on with it, showing some of these things and using them.

Hetan Shah: Drawing a distinction between personal and non-personal data is really important, because in the transport sector there is so much we can do with non-personal data, and we must not let the debates around personal data stop us from that. The other distinction is the aggregate and the individual level data. Even if it is personal, if you are interested in it as an aggregate, we have always done that: for example, with censuses and so on. There is no problem with that. It is when you get to individual personal data that it

becomes more complicated. You may be aware as a Committee that there is an EU data protection draft regulation that could be very negative for medical research in this country; it is a side-effect of post-Snowden, Google and concerns around data.

Q220 Victoria Borwick: Unintended consequences.

Hetan Shah: A small citizens movement has sprung up and is saying on Twitter, "I want my data to be used for future medical research, social research and so on." Even in the case of personal individual data we must not take the view that everybody wants to lock down their data. It is a public good.

Gavin Starks: One more example is that over the last 18 months we have run a challenge series, in conjunction with Nesta, which has been very focused on social challenges. It has been quite a long process to nurture and incubate some of those ideas, but seven companies have emerged. PwC has just written a report estimating that they will generate a five to 10 times return on investment. I am delighted with that, but I am particularly delighted because they started off by trying to address social challenges around housing, healthcare or energy. We see that in the case of some of our larger partners like Arup and Syngenta in agriculture; they are starting to publish their own open data because it helps the entire market move forward, and that increases their business.

Q221 Derek Thomas: In the private sector, I understand why the data a private business might have would potentially be commercially sensitive and why it would resist exposing its intellectual asset to competitors. How do we encourage them to open up their data to Government and be confident that they will get a benefit in return?

Gavin Starks: There is a very simple answer from a business point of view: show me the money. In the work we have been doing to develop new business models, at one level it is quite a simple transition, which we have already seen in software. Software has transition from product to service, and we are seeing data transformed from product to service. Within the service provision you can have multiple pricing tiers, including free and open for different use cases. We have people like Thomson Reuters. To build on the Syngenta use case, they opened up data around pesticide usage and its compatibility with different types of crops. That increased their reputation. It also helped organisations they were partnering with and their customers deploy their products better, so they increased their sales. The argument is that we are going to keep producing use cases and benefits to show why this brings you more value. A very simple way to think about it is that, if you have a warehouse filled with goods and they are not moving anywhere, they are not gaining any value. If you have data sitting in a warehouse and it is not connected to anything, it is not gaining any value. In a digital economy, the more connections you have to a piece of data the more valuable it is, so it is a completely counter-intuitive way of thinking about it from the perspective of most businesses. They all start with, "This is our data," exactly like the TfL example, but going on that journey they create more value in the long term by taking an open innovation approach.

Hetan Shah: It is worth distinguishing data we might want the private sector to make open to everybody. ATMs are a good example. There are also water company boundaries. There are all sorts of things that only the private sector holds, but because there are

different organisations there is no single point. Most of those things are not commercially sensitive. There is a whole set of things that are not commercially sensitive and the public would benefit if they were open. A more complicated area is where it is commercially sensitive but the state would benefit from having that kind of oversight. Take prices as an example. How we calculate inflation is a real headache. At the moment people still have to add up the goods. We all do online shopping. Being able to scrape that data from supermarket websites would be really helpful, but at the moment they are not keen on that, so we need a relationship between the statistical office and those organisations. This is one of the areas where regulation would help, because it then creates a level playing field.

Gavin Starks: The work we are doing with the open banking working group on behalf of the Treasury is to create two things. One is to create what is called an open API that enables easier sharing of your personal information—looking at liability, privacy and so on—and the other piece is an open data API on banking products, like mortgages and loans. That helps to create a more transparent marketplace. In all the principles about open, one of the core commercial arguments from our point of view is that they reduce transactional friction. If you reduce friction in an economy, you have a stronger economy.

Paul Maltby: The conversation we have just been having about what is happening about Government and data is a nice analogy. Going back to the point in time—actually it is still often the case—when data was used because of a particular service or need within a particular team in a particular Department, it did its job and was never used again; it was perhaps used for some statistical purpose but largely it was redundant. What we have seen over the last Parliament is a very rapid move towards open data—the UK has been the best in the world in many ways on this measure—some of which can be made freely available for industry, academia, civil society, campaigners, people who want to criticise what we are doing and all those other good uses of data. What we are getting to now, and it is a big feature of the data programme, is sorting out the bit in between—the shared data you might not want to make fully publicly open but that can serve purposes that are much more than sitting in one particular service and being used for one particular topic by one particular team. It is a matter of putting in place the technical infrastructure and technology that allows that data to move appropriately to the right place. We do not want everything to move everywhere. This is about appropriate movement or access to data, but also shifting the mindset within public services, because this is data on which public service innovation for the next decade or so will be built, in the same way that in the digital economy there has been very, very rapid acceleration, innovation and growth for those who have made their services and, indeed their data, more freely available within a platform environment. It is not exactly the same, but it is a useful way to look at how Government thinking about data is shifting, from a very old locked-down place to one where it is locked down but some of it is made open, and there is a more sophisticated version where in many ways Gavin's diagram of closed-shared-open starts to get at the heart of what it means.

Hetan Shah: May I make one other point about public services?

Q222 Derek Thomas: May I ask you the next question because it may be linked to that? I understand that you have called for the ONS to have access to private data similar to

Statistics Canada. My question, which you may have been attempting to answer, is: would that actually improve public services, or is it simply to improve Government statistics?

Hetan Shah: There is a link between the two. If you want to know where to locate new transport hubs, you would be able to see where people are based and what journeys they are taking. All of this would feed into public services. You might want better tourism statistics to help you develop services at local authority level; you might want internal migration statistics within the UK. All these things could be improved and would help with more efficient and better allocation of resources. There is an absolute link between the two.

The other very important point I want to make about public service delivery is that, as we now have a multiplicity of providers delivering public services, one thing that has been missed is holding all those providers to the same data standards. Private schools and hospitals are not held to the same data standards in terms of reporting back to the state as public sector standards. In a sense there is a procurement issue. Similarly, having talked to some local authority leaders, they tell me that academies are no longer providing them with data. How can you run a local authority without some of that data? It is fine to have a multiplicity of providers, but the data must come back.

Q223 Matt Warman: Estonia is always held up as the model in this area. It has its own league tables and so on. Essentially, it has a digital identity card. We could have endless debates about whether a digital identity card would be a good idea, but how do you see the relationship between that digital identity card and the GOV.UK Verify scheme? Where will that end up? There is a lot of consumer enthusiasm for some of those services, but a lot of concern on the civil libertarian side of the argument. Where will Verify end up?

Paul Maltby: We work very closely with our Estonian colleagues, nearly daily, certainly on a weekly basis, and there is a lot of sharing both ways. There is, however, a very fundamental difference between GOV.UK Verify and the Estonian system. As you rightly said, the Estonian system is built on a common citizen identification register or number, and with that they link across all the state's services. They can easily translate from one service to another and have an overview of the population at macro level in what is a very small country—indeed, it would be a small city in the UK. Parliament spoke very clearly on this matter not that long ago and said it did not want a national identity card. Indeed, it is not so much the card as the common identifier underneath that.

The Verify service is fundamental and incredibly important. It would save a huge amount of money on identity and reduce some of the security risks over the next period on data. It enables those within government providing services to have trusted knowledge that you are indeed you. Instead of a process where we have amalgamated all the Government data on you and, therefore, we have some mega-database behind that, it is not that; one bit is a verification service provided in the market. Indeed, those of you who have gone through some of the processes like self-assessment tax might have seen it. You can choose from a number of different market providers—Experian and many others—and they will verify identity. You go through a series of questions, and once that organisation is happy to the standards we have agreed between Government and the provider, they will provide the service to you. That is the fundamental difference between those approaches.

As we go through the next few years, particularly as the spending review money will be in place, a big part of what we do—we are already doing it—will be to think about canonical registers as a fundamental building block for the Government's data infrastructure. That often means, essentially, lists of things provided as a service. Instead of every team in every Department having databases with their own list of companies—it might just be what people have entered on a web form; it may be the Department's own list of businesses—the idea is that, for example, there is already an organisation such as Companies House providing information on limited companies that can be pooled through an API into a dataset or database. The idea of registers as a fundamental platform to enable a data infrastructure and then the data economy outside is something we are currently building with Government Departments and starting to show, but the question of a personal register, or common personal identifier, is not part of the thinking or planning. If Parliament changes its mind about that at some point in time we would respond to that, but it is not required to build this sort of service.

Gavin Starks: To build on that, the fundamental difference is that the Estonian system is centralised; the UK system is federated. I think this echoes what we see happening across the private sector and in some of the development work we are doing at the moment with the Treasury. Federated systems reduce friction in the digital economy. More actors come in and provide small roles. The TfL example is similar. It is a more federated approach, which applies to identity as much as to banking. I think we will see that across all sectors. That pushes us into more of a ledger or registry conversation in trying to think about data infrastructure, but all these phrases are quite new, so we really need to invest first in data literacy.

Q224 Matt Warman: Could that federated model ever get us to the point that Estonia is at, in terms of having an individual way of verification? Effectively, you would plug into another service.

Paul Maltby: The way that could happen within the different approaches is that the data are held at individual rather than state level, and with your individual knowledge you are able to pull in from the different systems, which themselves are not necessarily linked together but are linked together at the level of you. That is talking about what could happen in the different directions in the medium to long-term future. It is not a current reality in the infrastructure that we have. It is swings and roundabouts. The Estonian system and the systems of other countries have a single identifier, which makes a lot of the data architecture around personal services much easier in a way, but it comes with the trade-off that there is a big system, and the Government knows and has all these things linked up. In the end, that may not be the most elegant solution.

Q225 Matt Warman: Does that mean that there is in a sense a possibility of the benefits of a one-identifier system without the concerns that would legitimately arise from Government holding all that information, or do we have to choose between one and the other?

Paul Maltby: You do not have to choose between one and the other, but I do not think that is a straightforward choice facing us right here, right now.

Q226 Matt Warman: Ultimately, if Verify goes to its logical conclusion, does that speed up the services we have or open up a whole load of new services?

Paul Maltby: We are talking about Government as a platform within these digital services. Essentially, that means having some common building blocks within digital systems. Instead of having a big proprietary system built from top to bottom, you are talking about smaller parts that are more loosely joined. Of course, there is no reason why a service like Verify should not also have customers outside Government as well as inside. There is no reason why another Government, or indeed any private sector company, could not also use that. It might not use that system, if that made sense. These are designed as open platforms for many teams rather than a small number of them. That is the purpose of them.

Q227 Chris Green: Mr Shah, you published in 2014 a data manifesto. How well do you think the Government are doing against the action points you identified? To put this in a similar context, in 2010 the Conservative manifesto had an invitation to join the Government of Britain. There was a lot of talk of the post-bureaucratic age and ideas about open access and information being made available to people. How well do you think we are doing at the moment?

Hetan Shah: How long have you got? I will try to give a very quick view. I have the data manifesto here. It has 10 points, so I will not take you through all of them. In effect, we were saying that evidence and data need to be put at the heart of Government decision making. There have been some positive moves around that. The What Works centres are developing academic ideas and building links between policy makers and so on. This Committee is starting to use the evidence check that has been developed by Sense about Science and the Institute for Government, so the Select Committee is playing an important role as well.

I have talked about data sharing. Institutions for trust is a very important area. It feels to me as though the Information Commissioner does not necessarily have the resources it needs to provide the trust in the landscape, so that seems to be one of the missing areas. On science and research, this Committee played an important role in getting the good budget outcome we have seen. The investment in the Turing Institute has been positive.

Q228 Chris Green: You said earlier that researchers now have less access to Government data, so in some areas we are going backwards.

Hetan Shah: There are some areas where things do not feel as good—that's right—but I think there is a wider mentality within government which is open to how data can help. There are specific units within the Cabinet Office and the stats authority—the Office for National Statistics—which take this agenda seriously, but within Departments it is patchier. The question is how you cascade some of what is happening at central levels through Departments to help that.

A question was asked earlier about whether we had the skills and capabilities within government. One of the possibilities is around the education system. We have seen the Government introduce a new core mathematics curriculum for those 16-year-olds who do not take A-level maths, which is 80% or so of people, so just building our quantitative

skills as a nation will help. As to other opportunities, all of you will be aware of the Nurse review of research councils. That will now create a stronger Research Councils UK, a top-tier council. One of the problems for data statistics and data science is that it fell between research councils, as it were, so there is now an opportunity for the new Research UK to take a more strategic approach to the skills agenda for the country.

Q229 Chris Green: One area for information is clinical trials data of pharmaceutical companies. Are you seeing as much progress as we ought to have had in that area?

Hetan Shah: We support the campaign that Sense about Science runs called AllTrials, which is all about trying to get pharmaceutical companies to register their trials. There have been some major steps forward in that area, but there is still a lot of work to be done.

Q230 Chris Green: Because we have to have all the data, as opposed to just what they feel comfortable releasing.

Hetan Shah: Precisely, but in the space of two or three years we have seen quite a major change in the mentality around that. Like some of the cultural changes Gavin has been talking about, pharmaceutical companies are recognising that they cannot get away with what they were doing before.

Q231 Chris Green: It is going to happen; they just have to change their culture.

Hetan Shah: They have to change their culture, but institutions like this Committee and civil society organisations need to keep up the pressure.

Q232 Chris Green: Can Government change their culture? Can we have a situation where data, when it is released, is released to the Minister at the same time it is released to society as a whole? There must be a few drawbacks.

Hetan Shah: That is something we campaign for every time there is a general election. Unfortunately, it has not made it into government yet. On the issue of pre-release access, we believe that in this world everybody should now have access to the data at the same time upon release, so you do not get the dark arts of spinning we have seen happen in the past, which reduces trust in data.

Q233 Chris Green: Trust in democracy and politicians is key. If the data was released at the same time to everyone it would be useful for business, charities and other organisations. It would also be quite handy for the media. The media can set their agenda and get the ball rolling. Ministers responding initially would say, "I don't know; I've only just seen the data." They will have to spend a great deal of time getting up to speed with the data. In that sense, should we be trusting the media far more than we do? It is the balance of trust, and trust in politicians; if you gave that power to the media, which is effectively what you would be doing, would that lead to a problematic area as well?

Hetan Shah: I think that over time conventions will develop. Everybody would know that the data had come out fresh and nobody had looked at it. You would give the Minister more time to make up their mind, but you would also be giving the media and the people they would be turning to—it might be the Institute for Fiscal Studies or the King's Fund—the ability to analyse the data at the same time. You would have varying interpretations of what the data was telling you, not just the central diktat of the Minister.

Q234 Chris Green: But that will depend ultimately on a more data-literate population as a whole, and that is going to take quite a long time to achieve. All these things have to work through the system and in conjunction with one another, but should any particular Department lead the way in releasing data first, perhaps in less sensitive areas?

Hetan Shah: We would say that applies across the board. You were saying that it needs an increase in data literacy across the population, and that is right. One of the very interesting things we have seen is the emergence of data journalism—journalists who are now taking data seriously. There is also the rise of fact-checking organisations like Full Fact who reliably check what is said by the media or by politicians and say, "This is the data. This is what we can tell is the truth, or isn't," and so on. As we see a plethora of civil society mechanisms to hold things to account, as numbers gain currency in public discourse, that is one way of creating more trust in the numbers we are seeing.

Q235 Chris Green: In recent years, statistics show that crime is down by 30% in Britain. A great deal of crime data is released. Has any work been done with the data available to say that the reduction in crime is due to one thing or another? Can we say of this 30% that the data is being used effectively and these are the changes in policing that have achieved these outcomes?

Hetan Shah: I am not an expert in this area, but my understanding is that it is massively contested. Some have argued that it is due to technological developments, like cars which are much more difficult to break into and so on, but the drop in crime is a phenomenon we have seen across the western world. The data will help us answer those questions and tell us things that are absolutely out of court and are still live theories, but my understanding is that in social research it is not a settled matter.

Gavin Starks: Bring us more data and we will be able to work out whether the causes and effects are correlation or causation. On your point about transparency in government, we are in the process of helping DEFRA. There is a maturity model that can be applied across different Departments to gauge where they are in giving better data to Ministers. The evidence upon which people are making decisions is critical. To give an example, which is perhaps tangential, we have just helped Burkina Faso to hold its first open democratic election. Instead of a latency of days between votes and the collection of them, the data was projected live into public squares, so the results were instant. That is quite a substantial change for the country, but those are the kinds of impacts we see for every Department and every company currently.

Chris Green: It sounds like an excellent example.

Chair: I thank the panel for their evidence today. It has been fascinating and optimistic. Thank you for the time you have taken. I am afraid we have to go on to the next panel as our witnesses are waiting outside.

Examination of Witnesses

Witnesses: **Mr Edward Vaizey MP,** Minister of State for Culture and the Digital Economy, Department for Culture, Media and Sport and the Department for Business, Innovation and Skills, **Rebecca Endean OBE,** Director, Research Base, Department for Business, Innovation and Skills, **Baroness Shields,** Parliamentary Under-Secretary of State for Internet Safety and Security, Department for Culture, Media and Sport, and **Peter Knight,** Deputy Director, Research and Development Directorate, Department of Health, gave evidence.

Chair: Welcome, Ministers and civil servants. Thank you very much for your time. We are very grateful to you for taking it. We are going to have a vote any minute. This is our final session on the big data dilemma. We have called it that because our evidence has been very clear that there are huge opportunities associated with big data, but also risks that are creating barriers to realising those opportunities. In the evidence from the Department for Business, Innovation and Skills, the Department was rightly proud that since 2011 the Government had invested over £500 million in developing the UK's big data capability. While we are gone, perhaps the Ministers can think about how they are going to explain what they have achieved with that money. I am now going to adjourn the Committee while we vote. We will recommence when we are quorate.

Sitting suspended for Divisions in the House.

Q236 Chair: I now resume the session. Thank you so much for waiting. You are very patient. I understand that we are also expecting a vote in the Lords. It is very unreasonable of the House to continue having votes while we are holding this very important session, but we must accept that that occasionally happens. I gave you quite a lot of time to think about my first question. What exactly have you been doing with that £500 million since 2011? Who would like to start?

Rebecca Endean: I am happy to take this one.

Mr Vaizey: I think we want to make some opening statements.

Chair: We do not have a huge amount of time, so let's start.

Rebecca Endean: We have spent it on a range of things. I can deal with this very quickly because we are short of time. It has been invested mainly in infrastructure in the research and science base. We have spent it on high-performance computing, in particular ARCHER and the high-performance computers in Hartree; better networks for the academic network, including JANET, which connects up all the universities; better

software, including the new IBM collaboration in Hartree, which brings things like Dr Watson technology into the UK; better data, storage and curation for researchers; and, finally, things like the Alan Turing Institute, which is about improving data science and making sure we get world-class scientists working on some of the key problems that face us today.

Q237 Chair: Minister Vaizey, you seem keen to contribute. The evidence from the Department for Business, Innovation and Skills states that fully exploited data could be worth about £216 billion to our economy, and other evidence says that it would create about 58,000 jobs and mean about a 3% uplift in productivity. Clearly, that is very exciting and we should be looking at how we do that, but other witnesses have told us that most companies estimate they are using only about 12% of their data at the moment and we need to look for ways to improve that. What is your strategy for doing that?

Mr Vaizey: May I start by saying that it is a great opportunity to be able to speak in front of this Committee and to give evidence with my fellow Minister and colleague Baroness Shields? You have called your inquiry the big data dilemma. I think you are quite right to balance the opportunities that big data brings with some of the risks, but what I would add is that big data is not going away. It is the Government's role to be on the front foot. I am sure we will cover a lot of ground in questions, but, broadly speaking, it is Government's role to be talking about big data and putting the UK front and centre of big data policy, to invest in it and to work with business.

As you say, Chair, some of the calculations show that there could be a significant uplift in productivity and value. Those figures can change in real time; they may go up or down, but I think everyone would agree that if we can make progress on how companies, both large and small, use big data—not crack big data, because we will be looking at this issue constantly—and if Government can play its part in opening up data, we will see some companies benefit significantly. You said that companies use on average only 12% of their data. One of the issues we are facing is that there is, as it were, a wave of data arriving by the terabyte, and processing it is very difficult and a big challenge. People forget that there are huge datasets out there that people do not have the time, skills or opportunity to analyse.

To echo what Rebecca said earlier, the Government's role is to invest in big projects that can help companies analyse big data or invest in skills, because we need data scientists who can help companies crack those big datasets, and also work with companies to tell them about the opportunities and give them a route map to engage in big data. I am sure Joanna will want to add to some of those comments.

Q238 Chair: I will target a question to Baroness Shields in a second. To clarify one point, you said you want to make progress in the way companies use data. What would progress look like in specific terms?

Mr Vaizey: We do not have a specific figure, plucking numbers from the air, for how many companies we want to see using big data by any given date, but, for example, the digital catapult, which probably really got going last year, engages with a number of companies. We want to see that engagement increased. They have made the use of data

very much front and central to their first work project in establishing themselves, but we do not have a specific figure for the number of companies we want to see engaging in this way, unless Rebecca has thoughts on what success would look like.

Rebecca Endean: One of the key aims of the digital catapult is to help small and medium-size businesses especially have a platform with that. There is a whole range of assets across the research base, including Hartree and various universities, actively engaging with business to try to help them use their data and share the assets they have available to make it easier for them to do that.

Q239 Chair: Baroness Shields, one of the big issues running as a thread through the inquiry is what has commonly been called the trust deficit. We heard evidence from the digital catapult that 60% of consumers are uncomfortable sharing their data; 80% of consumers think that organisations gather data simply for economic gain; and the ICO says two thirds of people feel they have lost control of their data. It is your job to run internet safety for the Government, so what do you think we should be doing to try to address the issue of trust, because it seems to be the key barrier in realising some of the undeniable opportunities that are there?

Baroness Shields: You are right to point out the big questions that are generated when everything we do in life is catalogued from devices that are literally becoming an appendage to our lives and minds. It is the business model of many companies to take that data and understand, almost down to your intentions, what you want to do and serve you up the best possible solution around that. With that come lots of challenges.

Before we go further into that, I want to point out a study that we did last year called Tech Nation, just to illustrate for the purposes of the Committee the real depth of data science, analytics and machine learning that comes about when you make use of data. The fact is that the UK has 16 centres of excellence across the country. We hear a lot about the northern powerhouse, but we literally have 16 centres in which we have expertise in these areas, in general associated with some of our great universities, but also as a result of working with the digital catapult, Tech City and the organic growth of those businesses supported by Government. Tech City is a low-cost initiative by Government. It has a budget of £2 million. It does not sound much, but the impact of that initiative, putting a spotlight on technology in this country, has led to enormous investment and innovation around this area. We have those 16 thriving centres.

On data security, there is a chasm in terms of what people feel about trusting data. If you talk to teenagers, they do not care; they have given up privacy and decided that they are happy to share absolutely everything in their lives and have it catalogued. We have some responsibility to look out for their interests, especially in terms of their rights online, making sure that their rights are served and that there are different rules for young people, because they are not necessarily always making the grown-up decisions we make. On the internet we treat everybody the same. It does not matter if you are seven years old, 25 or 65. Everyone has the same applications and services. That is a real issue.

Q240 Chair: The week before last, the Committee went to Oxford as part of Parliament Week. We held a session on big data with 90 sixth-formers who, quite frankly, asked better

questions than our Committee did of our experts. One of the questions that arose in terms of the opportunities of big data, but also privacy questions, was where the creepy line between opportunity and privacy is. Where is the creepy line?

Baroness Shields: Society and culture determine where the creepy line is. Tim Cook had a quote. In the summer he had a bit of a rant about technology companies in Silicon Valley—Google, Facebook and companies like that—who literally data-mine every single action we make on the internet. It was quite interesting to hear his perspective. He said that hoovering up all this is far more intrusive than anything Government do, but people do not really see it that way; they do not worry about a commercial company having access to all their data and information, whereas when you talk about bulk collection of data for surveillance purposes everyone is up in arms. He said that on a daily basis we give up much more information than we realise. I think the creepy line is moving further and further out; we accept more and more of our privacy being open.

Q241 Chair: Should we be leaving that to society, or should Government be taking action, legislative or otherwise?

Baroness Shields: It is hard to say. It depends. If it is a matter of national security, of course it has to be a Government issue; when it relates to protecting your financial security, absolutely. When it comes to personal issues and what you want to share, it is a little more difficult. It is really up to the individual. With the terms and conditions on every product we use, we always say "Accept." No one reads what it is being used for. We have been conditioned to say, "Yeah, yeah, thanks; I'd like to use the application," but the reality is that if you look deep inside it there is probably not a single person in this room who would be comfortable with what is being given up in that case.

Q242 Jim Dowd: First, I apologise to you, Chair, and the panel, because I will have to leave shortly. I have a brief question for Baroness Shields. The attitude towards data, even devices, is very much generational, in so far as the younger generation are much more comfortable; older people tend not to be. Is that a social development, in so far as they have been used to it from a younger age, or do you think that as they become older their attitude will become more defensive?

Baroness Shields: I am not sure that the train has not left the station. They are much more accepting of this as a new reality, because you get a lot from giving up data and information; you get a better experience. If you are shopping over the weekend on Black Friday or Cyber Monday and decide not to buy something, you will notice that on the next site you go to that particular product will show up and remind you that maybe you want to buy it. In some sense it is good; it reminds you that the item in the cart you left at the last site you visited is still available if you want to purchase it. In some sense, that gives you an illustration of how much is out there, and that you get a better service. In general, companies can use that data to deliver better services on a daily basis, whether it is train times or something else. We give up a bit but we get a lot as well.

Q243 Jim Dowd: You think people will become more relaxed over time.

Baroness Shields: I think so. There are certain things that you definitely do not want to share, but often my worry in the work I do is about trying to educate young people as to what they are giving up and making sure they are comfortable with that decision and that they are old enough to make it.

Q244 Carol Monaghan: Some of the companies and organisations that have been referred to—the digital catapult and TechUK—have already spoken to us in previous sessions of the inquiry. These are organisations whose remit is to promote business and entrepreneurship in the field of big data. What is the extent of the Government's role in promoting entrepreneurship in big data?

Mr Vaizey: We mentioned the catapult. The catapult is looking at all sorts of applications of data. The digital catapult is particularly focused on small businesses. It encourages small businesses and is designed to be a space for them to try out new business models and have the resources available to do that, which only a big company might have. That is the first issue. There are big companies as well. Rebecca referred earlier to the collaboration we are doing with IBM in terms of their health technology. They can work with the big data projects that the Government are funding. There are hundreds of millions of pounds going into these projects, and even big companies could not necessarily match that kind of research, so it is important that we collaborate with business.

Given the conversation we have just had about the risks out there, my creepy line would be the difference between identifying an individual as opposed to anonymous data. For me, the big opportunity with big data is huge datasets that allow applications to be put in place that will make a positive difference to citizens' lives. It is linked to the internet of things and machine-to-machine technology. If you can monitor traffic and link that dataset with air quality, you can provide a better experience for people living in a city, but nobody wants traffic monitoring data to be mined somehow to track where you have been in your car. That is the kind of thing we have to guard against, but it is important for us to work with business. Pure science research is very important, and where business comes in is in the real-world applications.

Q245 Carol Monaghan: Do the Government have a specific role in promoting business around big data? Are there boundaries regarding what the Government should and should not be doing?

Mr Vaizey: Rebecca may want to come in with specific examples of how we work with business, but the Government have a threefold role. One is to talk about big data. That may sound a bit facile, but, as Joanna was saying in respect of Tech City, the issues that Government say are important to business do gain traction and attention. Organisations like the catapult have a role in getting out to businesses and showing them how they can use big data. For example, in my own constituency I have the satellite applications catapult. I always dine out on the example they gave me when I first visited them; they engage with supermarkets because supermarkets use satellite imagery to monitor their car parks. The catapults can go out to business, evangelise and say, "You don't realise how you could be using big data to improve your business performance." Related to that, Government have to get further down the chain in providing the skillset: for example, in terms of some of the changes we have made to the school curriculum and also to further

and higher education. We need to make sure that businesses have available to them people with the right kind of skills to take advantage of this revolution. I do not know whether Rebecca has any examples.

Rebecca Endean: For small businesses, particularly in relation to catapults, there is a clear market failure about helping them get at the technology and taking products to market. We have fantastic assets in the science and research base across all our universities, and fostering really good collaboration between universities and the use of some of these assets, which are basically there only on a national, or even international, scale—for example, Hartree super computing facilities—is clearly a role for Government.

Baroness Shields: To add to what Rebecca and Ed were saying, the Government have great convening power. What we have been able to do in these 16 centres of excellence across the country is bring universities, companies and start-ups together. I was at the Imperial College data science lab, which we launched just last month. You have small businesses engaged in that. At Imperial West there are 60 or 70 small start-ups. All of them are working together with big companies on data visualisations. I saw a Bitcoin visualisation that showed malware interjecting illicit transactions into the universe of Bitcoining. Small companies were reviewing that data and determining what they could do with that and how they could improve their Bitcoin businesses. You see all of this coming together.

At Edinburgh University there is collaboration with a lot of e-commerce companies and start-ups. They are all coming together with the university. That is really important. You have the Government, universities, the Alan Turing Institute and the Open Data Institute. During the floods two years ago, the Open Data Institute and the Government Digital Service hosted hackathons with universities and start-ups. People over the course of a weekend created real applications that helped people in need, whether it was moving livestock to higher ground or food and supplies to the right places. All that comes together in the collaboration. We are quite good at that in the UK—*[Interruption.]*—in spite of this interference we have.

Chair: We have moved on from the creepy line to the creepy sound of the wind outside.

Q246 Graham Stringer: The care.data initiative was delayed. Can you tell us what lessons have been learned from that, and how the scheme that is going to be implemented by 2018 will differ from what went before?

Peter Knight: Thank you for inviting me to give evidence. It is for NHS England to comment on the details of lessons learned, because they are the people operating and running the project. What I can do is give you an update on where it is at this point in time. It is currently in what is called a pathfinder phase. There are four clinical commissioning groups across the country looking to pilot what the material and the consultation around it will be used for. I guess the lesson learned is to make sure you get your material right with the population before you go forward.

The second thing is that Dame Fiona Caldicott is doing a review of both the consent question and the question around what cyber-security standards we need to operate across the healthcare system. Dame Fiona is taking oversight of what care.data does before it

starts to extract data.[1] Until she is satisfied that those two things have been completed, care.data will not be taking data out of the systems at the moment. NHS England will have more detail on that, and certainly I will get them to write to you about the lessons learned.

Q247 Valerie Vaz: Minister, I understand that you have to leave early. It would have been quite nice to talk to you about lots of different things.

Mr Vaizey: I am happy to stay. Nothing is more important than this Committee. The all-party writers group will have to wait.

Q248 Valerie Vaz: Could you tell us a bit about the digital transformation plan and how we are getting on with it?

Mr Vaizey: The digital transformation plan was announced in the summer Budget. The plan is to produce a report, I hope by the end of January—don't hold me to that—which will look at drawing together all the digital initiatives the Government are currently undertaking. It will look at all the landscape you are covering in terms of organisations like the Alan Turing Institute, the Open Data Institute and others, as well as Innovate UK, but it will also look at each area where Government operate, whether it is schools, transport or the Home Office, and how we can start to embed digital. The work is being undertaken by officials in the digital economy unit in my Department, but it is important that we work very closely with the Cabinet Office and the Government Digital Service. For me, the great prize of the plan will be essentially embedding digital and fundamentally changing the way the Government do business, but also the kind of experience citizens have in a whole range of different areas.

Q249 Valerie Vaz: Will you be looking at privacy issues and regulation as well?

Mr Vaizey: Privacy will form an important part of that. What is clear from all the questions so far in this evidence session is that privacy and security sit right at the heart of everything we do, because it is not just keeping data anonymous; it is also keeping data secure and not allowing it to leak. Cyber-security and keeping data secure is a very important aspect.

Q250 Valerie Vaz: Will that work look at the skill shortage in data analysis?

Mr Vaizey: It will certainly look at the skill shortage. We are investing a lot in skills. It is important to say that every developed country is facing a skill shortage. The US is facing one, even though we laud their leadership in technology; other European countries are as well. We are fundamentally changing not just the school curriculum but also further education and university curriculums, embedding a much closer relationship with business; you need that relationship with technology companies because technology changes so quickly. But there is no magic bullet. It will take time. We are working very hard to address some of these issues.

[1] The witness later clarified that, Dame Fiona is setting the criteria for NHS England on Information Governance for care.data to achieve before it starts to extract data.

Q251 Valerie Vaz: You are straddling two Departments. Which Department has the responsibility for it?

Mr Vaizey: DCMS has responsibility.

Q252 Matt Warman: I want to talk a little about how EU data protection regulation meshes with UK data protection regulation. Where are we at the moment? This has been going on for ever.

Mr Vaizey: First, can I note this seminal moment in my parliamentary career? This is the first conversation I have had with Mr Warman that is not about broadband. That is an important point to mark. I think we should return for the anniversary next year. He has led from the front on that.

On data protection, we are expecting to complete the trialogues by the end of the year. An important point within your question, Mr Warman, is that DCMS has taken over responsibility for data protection, which from my perspective was a great prize. I do not want to be accused of empire building, but the key for me was to link the DCMS digital technology agenda with data protection because they are, as it were, both sides of the same coin. The trialogues will end this year, and we expect first reading to commence in 2016. It has been a long and very tortuous issue. One of the ways we have made progress is by parking some of the more contentious issues.

Q253 Matt Warman: Does that mean we have not signed up to the draft as it is and we are still negotiating on bits and pieces?

Mr Vaizey: We are still negotiating on bits and pieces, but the trialogue is also something we have to take account of. The European Parliament will have strong views as well and we have to steer a sensible middle course.

Q254 Matt Warman: What are the sticking points?

Mr Vaizey: The general principles are about the level of burden on business. We do not want to place too many onerous reporting requirements on business. We want to make sure we get that balance absolutely right.

Q255 Matt Warman: Do you think our current safeguards are sufficient, regardless of the European regulations?

Mr Vaizey: We can live with them as they are. We do not want the data protection negotiations to drag on for too long. They have already dragged on too long. At the same time, we also want the best result. I do not think it would be sensible to have any kind of interim measures between our current regulations and the future regulation because that would be confusing for business. We want to get on and get the regulation approved and then deal with the equally important aspect of working with business on implementation. It is very important, even when a regulation is approved, that you do not simply sit back

and say, "There you are. Get on with it." You have to have a conversation and dialogue with business about the best way of implementing the regulation.

Q256 Matt Warman: There is still the issue of who owns an individual's data. We trialled that here with midata, and the principle is in the European regulations. How did those trials go, and where do you anticipate us ending up?

Mr Vaizey: We support the principle of getting companies to release useful data. I think we have made a lot of progress with it. We have made some progress with the personal current account market; we have also made some progress with the energy market. Rebecca is nodding her head, so I have obviously read my brief properly. Do you want to add anything?

Rebecca Endean: No, I do not. It's a perfect reading of the brief.

Mr Vaizey: Do you want to say anything else about the data protection regulations?

Rebecca Endean: The main thing is to make sure it is proportionate and risk-based. In particular, as you know, on the research and science side we are very conscious of the need to protect personal, individual, social and medical data, but at the same time they have massive potential to improve things, like stratified medicine. Peter, I do not know whether you want to say a quick word about how we want to see the EU data protection regulation move on that.

Peter Knight: From a health research point of view, it is absolutely vital that we get a balanced, proportionate regulation, particularly when we are talking about the reuse of data. That is an area where there have been lots of different views across the European sector. It is hugely important that reuse is balanced with privacy and the rights of an individual. Research value is huge. In medical research, we learn a huge amount from the data we collect and use for research in that space. It would put Europe in a difficult position if the secondary uses part was put into a difficult position. Colleagues in DCMS have been fantastic at taking the line forward very clearly and holding the ground on where the line should be. Certainly, the medical research fraternity have lobbied very strongly in this area.

Q257 Matt Warman: You touched on the issue of malicious re-identification of individuals. I guess we will have the opportunity to insert that when we put European regulations into UK law. Is there a plan to do that, or is it not an issue from your point of view?

Peter Knight: The key point is that we have more than one area of law to rely on in the reuse of data. We always tend to go for the privacy positioning one, but actually we have contract law and employment law as well, where malicious use of data is occurring. We need to make sure that we use the law appropriately across all three of those different areas, and use the appropriate law for the intent.

Q258 Chris Green: Minister, the Data Protection Act provides for civil rather than criminal penalties. Do you think there is a case for introducing criminal sanctions for data protection breaches?

Mr Vaizey: To pick up what Peter said just now, there are criminal penalties under section 55 of the Data Protection Act for the active obtaining of data unlawfully with intent to disclose it. As I understand it, the regulations will give member states an option to introduce criminal penalties where civil penalties currently exist. I have to say in all honesty that I have an open mind about that. To turn something into a criminal offence is always a big step, so we would need to look very carefully at it. As far as I am aware, unless others on the panel want to correct me, we have not come to a conclusion within government on whether or not that would be an appropriate step. We want to get the regulations approved and then return to it.

Q259 Chris Green: Are there any particular reasons you would highlight for not doing so?

Mr Vaizey: The fact that it is a criminal offence maliciously to access data with the intent, effectively, to misuse that data covers what perhaps the ordinary person in the street would regard as a criminal act. I would not want criminal legislation inadvertently to catch people who have been negligent, however much they might be condemned for their negligent behaviour in allowing your data to become available. We would have to think very hard, if we were to introduce criminal penalties, about what kind of behaviour they would catch. I stress that I am not ruling out introducing criminal penalties; I am just saying there should be a serious threshold to cross before we decide to go down that route.

Q260 Chair: We had evidence from the Information Commissioner on this issue. He has called for a criminal offence to be introduced. He feels that the penalties are not sufficient. His view is that section 55 would not hold up for cases of re-identification where big datasets had been anonymised but then re-identified through jigsaw techniques, or something like that. Is any work going on at the moment in DCMS or BIS properly to test the legislation against the very fast progress going on with data at the moment?

Mr Vaizey: I will write to the Committee on that point. We have recently taken over data protection policy within government, which is a very good thing, as I said earlier. I am not aware if the Ministry of Justice, when they had that portfolio, were undertaking that work. It is important to listen to what the Information Commissioner has to say. The Information Commissioner levies pretty hefty fines on companies, but I am all too well aware that you can spin out the civil process and potentially get away with not paying a fine if your company goes bust and so on. If the Information Commissioner is saying that we should take a serious look at introducing a criminal penalty and if he can make a convincing case that section 55 does not cover what is in effect—echoing your point, Madam Chairman—a new technique because of the rapid way technology can change, we would be very sympathetic to what he has to say.

Q261 Chair: I heard what you said about coming to a balanced and proportionate conclusion on the EU data directive. Is there a Government position on explicit consent versus unambiguous consent? There are different European positions on that, as I understand it.

Mr Vaizey: We are concerned about the meaning of explicit consent, and what kind of hoops business would have to go through to introduce it, as opposed to deemed consent or

unambiguous consent. We have a similar issue with the e-privacy directive and, in common parlance, how many clicks you would have to go through to show that you have given consent. The current position, which we will all be familiar with when we use websites, is suitable, relatively friction-free and gives you the chance to give consent. We would want to be very clear as to what the Commission meant in practical terms by explicit consent.

Q262 Derek Thomas: One of the concerns about big data is that computers may well decide about human life rather than a human being. We might decide that's not such a bad thing, but when decisions affect human lives—for example, computer-driven credit assessments—are we concerned about that? Is that a problem? Should Government intervene in this area?

Mr Vaizey: There is a philosophical element to your question, Mr Thomas. We know that in many areas of our lives we come across the terrible phase, "Computer says no," car insurance being an extremely good example. All of us in this room know we are brilliant drivers, but given our demographic or perhaps our children's demographic, the insurance company, or more accurately the computer, thinks something different.

There is a series of different regulators across business who I think should play a role in this. That draws out a very important point: big data and new technology are for everyone and no one should assume it is somebody else's job. How companies use an algorithm to assess whether you are eligible for insurance and so on should be as much a matter for the insurance regulator to discuss with that business sector as it should be for, say, the ICO or Government Departments charged with the whole big data agenda. From my perspective, my agenda as a Minister working with Lady Shields is to drive research and take-up of big data to help the UK economy and UK business, mindful of the privacy and security concerns that exist with this kind of technology, but I would take how business uses big data in terms of interacting with its customers almost on a case-by-case basis.

Q263 Derek Thomas: Would you say that the upcoming EU regulation provides better protection?

Mr Vaizey: Yes. By definition, I think it will provide better protection and, hopefully, a uniform system of data regulation across the European Union, which will be a very big prize. It is a long overdue update to take into account the fact that companies can now use much more digital data.

Q264 Chair: You have been very patient. I am very grateful to you for taking the time out. As a final question, we heard on the previous panel from Hetan Shah, head of the Royal Statistical Society. They have given us evidence that, while the Government are now better at sharing data between Departments, they are less good at sharing it externally. The key recommendation they made in this inquiry is that the Government should open up statistical data-to-data bodies like the Royal Statistical Society and the ONS. They pointed to Canada, which has opened up public data as well as private data, and they believe this is a very good way to develop policy and encourage good use of big data. What is the ministerial response to that proposal?

Mr Vaizey: I am surprised. I would have thought it was almost the other way round. We have made 20,000 publicly held datasets openly available for reuse. I think we came top of the 83 countries surveyed by the World Wide Web Consortium. I think the problem is Government Departments sharing data between Government Departments. We have set up the Government data taskforce with the chief scientist and others to try to get Government Departments to take big data seriously, to see the opportunity and also to share it. Mindful of the ethical concerns Peter talked about earlier surrounding things like care.data, this provides massive opportunities. I think we need to look at potential future legislation to allow that sharing to be made easier between Government Departments. Rebecca, you may have a view. I hope it is not different from mine.

Rebecca Endean: It is not. When we talk about physical data and stuff like data-backed themes, we have been very good. Where there is an issue, which is not surprising, it is the personal data we hold on individuals in areas of health, higher education and education. Where I used to work—the Ministry of Justice—we hold great datasets in government, but we have to hold them very securely because they include very sensitive data. We are trying to explore ways of making that data available to academe in a way that is safe and in accordance with the law, and also bears in mind the important ethical and privacy issues academics take very seriously. There is also quite a lot of work on growing metadata so you can work out what the data is telling you. A lot of people across government, including the Cabinet Secretary, are very keen to make this happen and progress is being made. Some of the data protection and privacy issues related to these datasets are very important and we need to treat them very carefully.

Q265 Chair: Clearly, privacy issues are important; we are spending a lot of time talking about them, but are the Government or different Departments too risk averse? Are some more risk averse than others? Is there a cultural problem that needs to be overcome, or is it merely a technical or legislative one?

Rebecca Endean: There is varying progress. If I could put in a plea for my old Department, the Ministry of Justice, they developed a very interesting and novel way of helping charities work out who is and is not reoffending, by allowing charities to send their data to the Ministry of Justice. The Ministry of Justice did the matching and analysis and sent back the results. That was hugely successful. There are various data labs like that growing up across government. In addition, there are ways of working with the ESRC and others to try to get big datasets out. There is always more you can do. Sorry, this is my favourite subject so I'm ranting. I will pass over to Peter so he can have a word.

Peter Knight: There is also an issue about compute over, as opposed to taking data away; you can run your algorithm and get your result back and use it in situ. Health is working a lot to look at that area, because of the privacy issues and the confidential personal information held there. The concept of being able to compute over and get a result out gives you a balance between safety and utility.

Mr Vaizey: The answer is yes, but you can understand why.

Baroness Shields: I am curious about whether that comment related to commercial opportunities or to better services for citizens, because they are different. If it is commercial data we are doing really well, but when it comes to utilising personal data it is

very difficult to make sure that it is anonymous and you are protecting privacy. There are so many new techniques to use machine learning to repurpose the data and construct who that individual was, backwards from the point of it being anonymous. We have to be really careful with that. I would err on the side of a little more caution when it comes to personal data anonymised for commercial purposes. I would be more than happy to suggest that we give more to other Departments so that people can get a better service from Government, but we have to tread carefully. We have pushed the boundaries more than any other country, and that is really positive and commendable. I am sure there is more we can do, but if it is for commercial purposes we are doing pretty well.

Chair: We have come to the end of our session. Thank you for the time you have taken, and for your patience in waiting for us while we had two votes on the Immigration Bill—shocking—and I thank members for their brief questions during the end of this session. That brings us to the end of the big data dilemma inquiry. The conclusion we have probably come to in this session is that we must not be so occupied with managing the risks of big data that we manage out the opportunities, but clearly I cannot predetermine what the Committee will conclude in our report. We may come back to you with further questions as we prepare our report. Thank you for your time.