HOUSE OF COMMONS

Science and Technology Committee

Oral evidence: Social media data and real time analytics, HC 245

Wednesday 18 June 2014

Ordered by the House of Commons to be published on 18 June 2014.

Written evidence from witnesses:

- Digital Science
- Alliance for Useful Evidence/Demos/David Omand
- techUK
- Consumer Data Research Centre, University of Leeds
- Universities of East London and Birmingham
- University of Cambridge

Watch the meeting

Members present: Andrew Miller (Chair); Jim Dowd; Mr David Heath; Stephen Metcalfe; Stephen Mosley; Pamela Nash; Graham Stringer; David Tredinnick

Questions 1-79

Witnesses: **Timo Hannay**, Managing Director, Digital Science, **Carl Miller**, Research Director, Centre for the Analysis of Social Media, Demos, and **Sureyya Cansoy**, Director, Tech for Business and Consumer Programmes, techUK, gave evidence.

Q1 Chair: I welcome the panel here this morning. Can I ask you, for the record, to start by introducing yourselves?

Sureyya Cansoy: Good morning. My name is Sureyya Cansoy. I am a director of tech for business and consumer programmes at techUK, the trade body for the technology industry. Thank you so much for the opportunity to talk to you this morning.

Timo Hannay: Hello. I am Timo Hannay. I am managing director of Digital Science, a division of Macmillan Science and Education, the publishing company. We make software for researchers and scientists.

Carl Miller: Good morning, everyone. I am Carl Miller, the research director for the Centre for the Analysis of Social Media—a bit of a tongue twister—at the think-tank Demos.

Q2 Chair: A number of companies—I guess some of yours—make money from the use of social media data. There is nothing wrong with that; it is perfectly respectable. Will you tell us how the business models work—how you achieve that?

Sureyya Cansoy: From our perspective, there are probably two broad ways that companies make money. One is that technology companies provide the software and tools necessary for big data analytics and social media analytics. They make the software and tools available to their customers through licensing, and they generate revenue through installing that software and training the staff in the customer organisations who will be using those tools. A smaller number of those tech companies also provide an end-to-end service delivery to customers. Instead of just providing the software and the tools, they actually provide the whole service to the customer. That is mainly how they make money.

In the broader context, companies using social media analytics ultimately make money out of gaining better insights into their customers, which increases their revenues and profitability. It is as simple as that. Some of the technology companies, in addition to providing software and tools for their customers, also use social media analytics to understand their customer base and the needs of their customers better so that they can tailor their products and services much better to their customers.

Q3 Chair: The raw data itself is not the value; it is the data plus the analytical work done on the data.

Sureyya Cansoy: Very much so. If we look at the amount of data out there, especially in the context of social media data, it is vast. You can argue that it only creates value once you analyse that data; you are able derive certain insights from the data that you can then apply to your business in order to become more efficient and effective, in order to create innovation and so on. Having said that, in order to do that analysis you need the data. It goes hand in hand, but analysis is an important element.

Carl Miller: Social media platforms mainly regard themselves as advertising companies rather than data brokers. There are two models for the platforms. They either advertise—Facebook regards itself as an advertising company—or they sell data. Twitter stands out from all the companies as having a more data-centric view of how it is going to make value in the future, but it is not making that much value at the moment. If you look at the valuation of Twitter, it is 60 times what it currently turns over every year, which is enormous, and most of that is because of the future potential that they see in the data they create and can sell.

Timo Hannay: May I comment on our business? We make a slightly unusual but hopefully interesting use of social media data. We have a range of young scientific software companies in which we have invested. We have a portfolio of second businesses in which we have invested. The most pertinent to the subject of this discussion is a company called Altmetric, which is a very young commercial organisation that was

founded a couple of years ago. Its business is gathering social media data and other data to look at what attention is being paid to the scientific literature, the scientific papers. We are trying to enable a scientific discourse online. It works by gathering data from a range of sources, so it includes social media—Twitter, Facebook, Google Plus, Sina Weibo and so on—and also the several thousand blogs that post regularly about scientific papers and scientific subjects. We use that, together with data from mainstream media, from sources such as Government policy documents, to look at what attention is being paid to scientific papers and what impact they are having.

We sell the business model to publishers; most of the top scientific publishers use our data in order to be able to point their readers and authors to what is being said about the content online. We sell data analytics to universities and other research organisations so that they can see what is being reported about their own and others' research. We also sell data and analytics to funders who want to conduct similar kinds of analysis. We are not in the business of trying to market to people through social media, or those kinds of commercial pursuits; we are trying to facilitate scientific discourse online through the use of social media data.

Q4 Chair: You will know from some of your scientific publications that, in some scientific disciplines, there is such a mass of data out there that finding reliable ways of analysing the data is quite hard. What is the risk of this analysis of big data becoming simply a big disappointment because it turns out, down the road, that you have got it wrong? How robust are your systems?

Timo Hannay: It is fair to say that these are early days. As I said, Altmetric, the business in which we have invested that is relevant here, has only existed for a couple of years. The whole area of so-called altmetrics—the field in which people are applying these new kinds of measures, particularly data that you can gather online about the attention being paid to the impact of research publications and research activities—is a new field, so we are still learning about it. It is not just a matter of commercial development; it is a matter of research and development, as we try to understand these data.

That said, it already has value. We already have 50 prominent, and in many cases large, scientific and technical publishers using our data and displaying it alongside their content. We know from surveys, and from more informal reader and author feedback, that people find this information useful. It is useful today, and it is generating revenue and commercial opportunities today, but it is still in its infancy, particularly in its use in research evaluation. You will be aware, I am sure, that research evaluation these days is conducted mainly through analysis of citations in the literature, which is valuable and will continue, but it takes a long time and measures only one kind of impact. What we are trying to do is to measure a much broader range of impacts, and to do so in real time online, through the analysis of these data.

In terms of getting it wrong, the risks are much higher on the research evaluation side, so things are moving more slowly there. We need to be cautious. We need to understand how to interpret these data, and how they can inform decisions around funding, appointments, tenure and so forth. They undoubtedly can, but we need to be cautious, because those are important decisions.

When it comes to helping people to engage in the online conversation around the literature, and helping them to discover papers that they might otherwise not have read but which are getting a lot of attention online, and drawing their attention to that fact, that has value today. That is indisputable.

Q5 Chair: Moving on to public perceptions, a couple of papers ran the headline this morning, "GCHQ are snooping on your Facebook site"—shock, horror! How aware are the public of how data is being collected about them?

Carl Miller: There really is the potential for a looming consumer crisis around the private sector use of social media data. Demos has done quite a lot of research into public attitudes in the UK, and a paper called "The Data Dialogue" was published a few years ago. It is fair to say that most people have some awareness of the fact that their social media data is being taken up, but it is impossible for them to have a specific awareness—I don't, and I work in the field—because technology, and therefore what is possible, is moving forward so quickly, but there is a growing sense of unease everywhere.

Between 60% and 90% of people questioned about specific uses of social media data are uneasy about how it is currently happening. The reason that this is the case, we think, is that people are not given enough choice about the specific deals that they can strike with data controllers about how their data is used. At the moment, you either use Twitter or you do not; you either use Facebook or you do not; but underlying that is a real diversity of views among the public about what counts as intrusion, what value they think they should get from their data and how happy they are for people to use it. Probably 30% of people are non-sharers; they do not want their data to be used at all. You have 8% who we call enthusiastic sharers, who see no real problem in people using their data however they wish. Between, there are different shades of grey—pragmatists or value seekers, who are looking for value in exchange. At the moment, they do not see that value; they do not see that they are getting the right amount of value for the value they see in their data.

This is somewhere the Government can do a lot of good, helping to create a new consumer rights regime around data, with kitemarks and clearer terms of service. If we do not see a more consumer-friendly regime being implemented, we could see significant and systematic withdrawal of consent from people about social media use. We are already seeing social media sites lean towards becoming more and more private, with Facebook profiles becoming more and more closed. For me, as a researcher, that is incredibly concerning, because the value—the innovation—is in having the data public, in having it accessible and allowing people to innovate in finding uses for it.

Sureyya Cansoy: From the techUK perspective, the technology industry is very aware of the need to reassure the public about the privacy and security of their data. What Carl said is correct; there is some awareness among individuals of how their data may be used through social media, but perhaps not full awareness. I would like to mention a specific initiative that techUK is actively involved in, which we think will make a positive contribution to the debate. You may be aware of the Information Economy Council, a body that brings together Government, industry and academia to drive the information economy in the UK. It is co-chaired by the Government and the techUK president. One of the key initiatives that the Information Economy Council is looking at, in which techUK plays a priority role, is creating a set of data principles to address how we can reassure

consumers in this new digital age without losing the opportunity to get the most out of technological innovations.

Q6 Chair: Do you mean reassure, or do you mean demonstrate that there is a commercial advantage to them?

Sureyya Cansoy: I think it is both. There is a reassuring element in it, but we are in the early days of putting it together. The work is led by the Connected Digital Economy Catapult, with strong support from techUK and others in the industry. It is very much an industry initiative, proactively saying that trust and confidence are important, and that we need to get that right. They are currently working on an initial set of principles, covering areas such as access, control, transparency and simplicity of message, which play into how aware consumers are of how their data is being used. We are currently consulting on those, and we think that they will be ready later this year, but it is important for us to show that the industry thinks it is an important issue to get right. The other important point to keep in mind is the balance between respecting people's privacy and ensuring the security of their data, and making sure that we can take full advantage of technological innovation.

If I may, I will comment briefly on the point about the news this morning. We need to be clear about the interception of social media, which is what the news was about today, and the use of social media analytics, which in the large majority of cases is based on anonymised data, to come up with insights into general behaviours and patterns—what people want from products and services and so on. We need to be quite clear that they are two different things. We should not mix security laws with privacy considerations. Both need to be legislated for properly, and laws on them need to be complied with by companies, but we think they are separate issues and considerations. Security is absolutely essential for a free society and an open economy. However, it needs to be done in such a way that there is appropriate oversight and scrutiny around it. We think that is a matter for Parliament and the Government; the companies who are members of techUK always comply with the laws of the countries where they operate.

Timo Hannay: I want to make a couple of points to add to what my co-panellists said. I completely agree with what Carl said earlier about the fact that users of social media need to be better informed, and I agree that even those of us who are closely involved in it sometimes do not understand all the implications of what is going on. Equally, we as users of that kind of data, and the companies that are gathering it, need to understand what user and consumer expectations are, and therefore what best practice is. I shall give a simple example.

We take a datafeed from Twitter and use that to see what attention is being paid to the online scientific literature. If someone were to delete a tweet, Twitter's system would immediately tell us and it would get deleted from our system as well. That seems to me to be consistent with what I would expect as a user of Twitter and, in that particular regard, it constitutes best practice. There is a useful role to be played in understanding what user and consumer expectations are, and in setting best practice standards on how we should handle the data.

I reiterate briefly that it is very context-specific. When we talk about the use of social media data, often the case that we have in mind is using discussions I may have with my

friends online to sell me something. That is very different from some uses, and I consider our use of it to be an example; people who are engaged in online discussions about scientific literature and scientific results want to engage in a global discussion, and we are trying to make that information more visible. There are what I would hopefully characterise as relatively uncontroversial uses of this kind of data and slightly more controversial uses, and we need to distinguish the two.

Carl Miller: I have two quick points. The first is that there is a really baffling incoherence between people's stated views when you do surveys and what they actually do. People say that they are concerned about supermarket loyalty cards, yet they sign up to them in their millions. It is the same with social media. People say that they are concerned about it, yet more and more people continue to sign up to these sites and use them. This is something not very well known about the detail of people's everyday lives; we don't know how concerned they are, outside the explicit polling type of setting.

Point two is that the most concerning part of the whole data uses industry is third-party data use. This is the bit where people say—again, pinch of salt—that they are most concerned not about the social media platforms using it but how it gets taken to those more mysterious and, to them, shadowy worlds, where people take their data and sell it, resell it and sell it again, packaging it up and selling it again. This is the bit that they know least about and the bit that they are most concerned about, because they cannot actually judge what the implications of this industry are for their privacy.

It is clear when you know that someone is reading your tweets and may be analysing them, but what about when a third-party sophisticated data broker using technology that you cannot understand is packaging your social media data with loads of other kinds of data about you, perhaps to create a detailed profile that will then be sold to advertisers or political parties? It is that third-party use where there is the least amount of public awareness and the highest amount of public concern.

Timo Hannay: And none of us fully understands the implications of that—even the people who are experts.

Carl Miller: We are struggling with it.

Timo Hannay: We don't know where it is going to lead.

Chair: I have warned some of my colleagues about the things that they sometimes tweet.

Q7 Stephen Metcalfe: I want to go back for a moment to the value question about all this data. Some wild figures are being thrown around. One is that the UK will benefit by £216 billion by 2017 and that it would support 58,000 jobs, but that is about £3.7 million per job. Is that a realistic figure, or is it lots of doubling up, with the same data going round and round?

Sureyya Cansoy: I believe that that is from a CEBR study, the first study on big data in the UK and its economic value. It sounds realistic, because it is not only talking about it from our perspective. It not only talks about the data analytics tools of the technology industry and the revenues that they get from selling those tools and services to their customers; it also talks about how those tools then help the customers to achieve efficiencies in their

businesses, and about achieving better innovation and creating new businesses. I have not looked at the numbers in detail, but when you think about the impact that big data analytics will have, it is not in one specific part of the industry but across the economy. With that in mind, it is quite possible. I think that number is cumulative; it is between 2012 and 2017, a four or five-year period, which is about £40 billion per annum, give or take.

Timo Hannay: I cannot speak to those figures specifically; I can only speak anecdotally about our businesses and our opportunities. There are big opportunities. If we look in the area of scientific information and research evaluation, the established business is worth hundreds of millions of pounds worldwide in that one area. It is going to be revolutionised by the use of online data, including social media data but not limited to that. Through businesses like Altmetric, the UK is already taking a leading role. We are ahead of the curve compared with our competitors and colleagues on the other side of the Atlantic and elsewhere in Europe.

There is a huge opportunity. The businesses that we run are early stage but growing rapidly, so we will have three times more staff at Altmetric at the end of the year than we had at the beginning, with about five times the revenue. It is from a low base, but it holds huge promise. It is important, and it is a source of huge economic value, quite apart from the fact that the usage that we are making of the data is itself trying to support scientific endeavour, which of course supports economic growth and other forms of social and intellectual progress.

Carl Miller: The main value that we see in real-time social media analysis is in social research. For us, the growth of all this data in social media is the datafication of social life. It is the first time that all the things that happen everywhere in society have come together in a way that is inherently amenable to analysis and collection.

The commercial social research market is valued at about £3 billion to £7 billion a year, depending on how you value it. Real-time social media analysis, and even just social media analysis, has had strikingly little impact on this market up to now. Ten of the largest providers in that market account for about 6% of the total value. Social media research is a very small part of their business, so what we have seen is a growth underneath that, which is social media only—analytics companies, dashboards and so on. These have had a great impact on marketing and advertising in a few specific fields, but little wider impact throughout the whole of the business area of social research. That points to an important disjuncture, which is happening currently: social media is not having an even impact across either the private economy or across the whole of society.

At the moment, a few fields are completely redrawing their business models on the basis of the exposure of this data, but within civic society and especially within Government we have seen very little innovation indeed. There is so much more value that can be got from Government and the civil service using all of this incredibly valuable data, because of the many things that they do. Every Department has a stake in understanding what people think, and that is not currently happening. There is much more that the Government can do to support innovations in the public sector, helping to take the skills out of these concentrated little hubs of expertise and technological capability and moving them to companies, civic society, charities and people working for the social good, to allow them to get value out of this too.

Q8 Stephen Metcalfe: Despite what the Government are saying about adopting big data as one of their A-grade technologies, you are saying that it is not happening.

Carl Miller: It is not being used by the public sector. We have seen a visible increase in support for this within academia. In the "Eight Great Technologies" paper, I notice that social media is not mentioned as a distinct discipline. That is slightly worrying, because in my eyes it is a distinct discipline. It does not just come under the umbrella of big data; it is a new academic field and a new discipline within research, and, in my view, it should be a new profession in the civil service. It does not yet have that coherent body of expertise, so we are not seeing it emerge as a distinct and coherent community, a distinct and coherent set of skills or a body of technologies. At the moment, a sprinkling of practice is happening in disparate areas across the economy.

Q9 Stephen Metcalfe: You do not share Mr Hannay's optimism about Britain being well placed to make much of this.

Carl Miller: I think that we are. My experience in the EU, where we certainly have a strong higher education sector and lively tech innovation hubs, is that we are considered leaders. But the UK can take a step forward to bring together all the expertise that happens into something more coherent and more disciplinary, more recognisable as social media science or a social media analytics discipline. That is what we are lacking, in my view.

Timo Hannay: There is absolutely a skills gap. It is not limited to the public sector, which you are more expert on than I am. Today, we constantly have unfilled technical positions; we need people who can bridge the divide between computer programming, data analytics and, in our case, scientific research and research evaluation, and that is a challenge. Admittedly, we are in a particular scientific software niche, but the changes that we are talking about apply to scientific research generally. We need researchers who are more expert in data analytics and computer programming. It is changing every domain of society, so I completely agree with what Carl has been saying. There is a skills gap; we see it and feel it, and it is evident in other areas too.

Stephen Metcalfe: I think we are going to come back to that.

Sureyya Cansoy: I support that view. Talking to techUK member companies that provide these tools, the anecdotes that they tell us show that they believe the take-up of such tools is relatively strong in the private sector, particularly in sectors such as retail, retail banking and transport, but they think that take-up by the public sector is low. The UK Government have come up with a number of encouraging initiatives around open data and big data. However, I echo what Carl said; so far, in our view, they have not really used the full potential of social media analytics, or big data analytics broadly speaking, for the benefit of the public sector in delivering better services, and more effective and efficient services, to people.

I assume that we will come back to the skills question, but we also agree with the comments on the skills gap. I will leave it at that until we come back to it.

Q10 Stephen Metcalfe: On the public sector issue, do you see any improvement?

Carl Miller: There are courageous civil servants in most Departments who are trying to innovate in this area. They believe in evidence-based policymaking; they see all this new evidence and want to learn how to use it, but there is not the technology or skills within the civil service. I genuinely believe that they are not being supported in how to innovate by senior leadership in some Departments. In my view, the civil service has to react to this. There need to be defined experts.

Q11 Stephen Metcalfe: I understand that.

Timo Hannay: Could I comment briefly on the subject of Government? In some ways this is only tangentially related to social media, but we use Government information and we welcome open Government data. It is incredibly valuable, and a driver of innovation and progress, but Government can learn things from some of the social media companies in the way they release data. For example, we index Government policy documents from the NHS, the WHO, the IPCC and so forth, but it is a lot harder to deal with than our datafeed from Twitter. There are some things that the Government can learn from those kinds of organisations and the way they make data available to organisations like us.

Q12 Stephen Metcalfe: I have one final question, Chair, because I am very conscious that time marches on. From what I gather, the one measure that would improve the UK's prospects of making the most of this would be to get the public sector more engaged with using this technology.

Sureyya Cansoy: Yes.

Q13 Stephen Metcalfe: Does it make any difference to the fact that this data is being held by American-based companies as opposed to UK-based companies? Is that a barrier to us making the most of it?

Sureyya Cansoy: I do not believe so. The social media world is very much a global one. The data can be obtained in a number of ways, but I guess that there are two common ones. One is using the analytic tools and software available to analyse publicly available data. Because some of the social media data is publicly available, it can be analysed anonymously and collectively to come up with insights about trends and so on, and not necessarily about individuals. Another way that it can be done is by purchasing anonymised datasets from the companies in question. Having spoken to our members, I think that they make a lot of use of openly and publicly available data, and to a degree they purchase some of the anonymised data, but I believe that the cost may be a barrier for some businesses.

Timo Hannay: I don't think there is a particular issue with whether the companies are American or British, or from anywhere else. For us, the issue is more around the degree of competition and choice. For us, as data consumers and users, the question is: can we access this data from a number of sources, do we have choices, and even more important

for the consumer and users is do they have a choice? That, for me, is much more of an issue than the nationality of the business.

Carl Miller: For researchers and academic tech innovators, there is an issue around collective bargaining, as they often do not have the budgets to allow them to acquire data in the way that private companies do. The Economic and Social Research Council was going to put out a particular call on social media that would allow that kind of stuff to happen—to allow the UK higher education community to come together collectively to bargain with the social media companies. Disappointingly, that has not come about yet; lots of universities would like a little more help in acquiring some of this data.

Q14 Graham Stringer: Will the right to be forgotten affect the way that you do business? Will it affect your companies?

Sureyya Cansoy: I shall respond from the tech industry perspective as briefly as I can. In terms of the whole debate around the right to be forgotten, there are two angles. There is the recent European Court of Justice Google ruling, requiring the search engine to remove links to certain data if requested by an individual. The second one is in the proposed EU data protection regulation, where I believe there is a new clause that deals with the right to be forgotten.

We understand that it is an issue for consumers, particularly in the social media context, where younger people may want to remove certain data as they get older or move into professional life and so on. However, I have been looking at the European Court of Justice ruling on Google, and we think that there are a number of practical challenges with it, and we do not think it is the right mechanism to deal with what it is trying to address. In the interests of time I shall not go into detail, but the fact that it only removes the search link to the data and not the data itself is a major drawback; and another is the fact that somebody needs to make a decision about whether the information is in the public interest and should be retained, or whether the link should be removed. There are a number of other challenges around, but in the interests of time those are some of the key ones.

Q15 Graham Stringer: I may not understand the arguments, but I understand the issues involved. What I really want to know is whether it will affect your businesses at all.

Timo Hannay: We do not see it as particularly affecting our business. We want to be able to abide by any reasonable requests that users may make to remove things. I gave the example earlier that if you delete a tweet it gets deleted from our system automatically. That is a simple example of the same sort of thing. However, removing all information from the internet is difficult, needless to say, so we have to be pragmatic about it. What we will see over inevitably a long period of time, I hope, is that attitudes to this kind of information—things that people may post that they regret later, for example—will be considered more normal, because most of us will have fallen victim to it at one time or another, so the right for it to be forgotten may ultimately be overtaken by the right for it to be forgiven. In practice, it is difficult to completely remove all information from a global online network. That is a pragmatic point.

Q16 Graham Stringer: You talk regularly—it is the basis of your business—about having access to lots of social media data. Doesn't that data belong to the websites themselves—to Twitter and Facebook?

Timo Hannay: Yes, so we license it from them.

Carl Miller: It still belongs to the user, technically, so it is the users' data that they have licensed to social media platforms, essentially to be able to use how they want. Basically, we acquire it from them but it is still owned by them.

Sureyya Cansoy: My understanding is that, in the majority of cases, the data provided are large sets of anonymised data. Once data is anonymised at that scale, it is no longer subject to data protection rules and constraints, because a person is not identifiable by the data. I believe that the Information Commissioner's Office made a useful submission to your Committee, highlighting some of those issues. In terms of personal data, yes, of course it belongs to the individual; however, anonymised data provided at that scale is different from personal data.

Timo Hannay: To be clear, the data that we and others access does include information about, for example, your Twitter ID. It is the same information that you would find on the public Twitter website, but it is available in a machine-readable format that we can use and analyse. We can therefore link to users' Twitter accounts, for instance when users have commented on a particular research article. It is pseudonymous in the sense that you have no information other than their identity and their self-description on the system—you do not necessarily know who they are—but it is not completely anonymous in the sense that it is aggregated and you cannot tell one user from another.

Carl Miller: I have one quick point about data availability. There is a massive inconsistency between platforms about what data is available. Twitter is the only social media platform that currently has its own "ology"—its own body of study about it—and that is because it has made its data much more available than nearly any other platform. Realistically, if you are looking at one of the services that bring in all the data that is conceivably available and make it accessible to people to use, 60% to 70% of the entire ecosystem will be Twitter. Facebook is much more difficult to acquire data from; there are many more privacy implications with it. It is not actually part of Facebook's business model to make its data available in the same way. It is an advertising company fundamentally, not a data science company.

Timo Hannay: Of course, we make use of the data under licence terms with those companies, so they will put certain restrictions on what we may and may not do with the data. Among other things, that will help them fulfil their obligations to their users under the terms of use on their website.

Sureyya Cansoy: In addition to complying with the necessary laws and regulations, an increasing number of companies are concerned about ethics questions. It is not only "Can I actually do this under the law?", but "Should I do this? Is it a fair treatment of the data?" Most companies involved in the analytics of such data are taking a much stronger stance on the ethics question as well as on compliance.

Carl Miller: The industry has failed, as yet, to create a set of best practice guidelines.

Sureyya Cansoy: We are hoping that some of the work that we are involved in through the data principles will address that, as I outlined earlier.

Q17 Graham Stringer: You touched on this in your earlier answers, but do you have concerns that the data you are collecting has not had informed consent from the individuals concerned?

Timo Hannay: I am somewhat concerned about that. As I say, the use to which we put the data would, I hope, be considered by most people as uncontroversial, but it would certainly give us greater comfort if there was as much transparency as possible in how the data were being used, and that people were aware of that.

Q18 Graham Stringer: Do you think that that would come from the original contract with the website?

Timo Hannay: Possibly, yes. I do not have a strong opinion about the mechanics of how it should happen, partly because there is such diversity of data sources and services that it differs according to which website or service we are talking about. It would inevitably have to involve them, because they are the ones with the relationship with the user.

Carl Miller: There was a wonderful statistic that if you read all the terms and conditions on the internet you would spend a month every year on it. I would love to see a kitemarking regime, where an independent body could authenticate whether or not the terms and conditions were clear. That is a basic responsibility for a platform. For us to have confidence that informed consent has been given between the creator of the data and the platform, we need to know that the terms and conditions are clear enough for that person to understand what is happening. The big concern is that technology is moving so quickly that informed consent, in the sense of knowing all the things that it is possible to understand by that data, is changing so rapidly that I do not know how it can be publicly achieved.

Q19 Chair: It also needs to be consistent with the law of the country where the person resides.

Carl Miller: Right. It is incredibly tangled. Legal jurisdiction is meeting rapid technological change and globalised information architectures. It is a really tricky problem, and informed consent in those conditions is something that is far from being clearly secured. Having said that, with some platforms, like Twitter, it is so unapologetically open; the first line is something like, "If you tweet, you are asking Twitter to make your tweet as public as possible. You understand that your tweet can go around the world." Clear statements like that are helpful in informing people's reasonable expectation about what can happen and what is possible with their tweets. The other reason that Twitter has its own "ology" is that, ethically, it is clearly the most straightforward source of data to use.

Mr Heath: I wonder whether anyone has ever met anybody who had read all the terms and conditions before clicking on "I agree."

Chair: Or who understood them.

Q20 Mr Heath: They will not understand them if they have not read them.

I think you are suggesting, Mr Miller, that Twitter is open about the purposes that it sets out for itself. I wonder whether other companies gather information at an early stage with a view to retention and sharing, or whether that is a secondary consideration—that the information is there and they then see whether it is useful. Do you see the difference between the two? I just wonder whether companies get unnecessary information from their clients, with a view to seeing who it is going to be useful to.

Carl Miller: We are certainly seeing a subtle pivot towards more data-centric business models. That is because every time someone innovates a new way of using Facebook posts or tweets or anything like that, they make that data more valuable. Social media began basically as an advertising industry; it began as a way of bringing people to a place where you could advertise to them in ways that were highly defined and specific. Now, most social media platforms are looking at ways of leveraging their data that would be much more valuable than that, to gather and sell. The big data mentality is that if you have data, you want as much as possible. The basic premise is that data is valuable. For lots of companies, it is not just a platform; they are gathering as much data as they can get hold of, because they know that it is going to be valuable in future.

Timo Hannay: Certainly that is the incentive. The simple answer to your question is that I do not know. By definition, if they are not making that data visible—at least not yet—we do not know what data they are gathering. The concern is, first, that it is opaque and, secondly, that the incentive is to gather as much data as possible precisely for the reasons that you explained.

Q21 Mr Heath: When they gather a lot of information, the concept of anonymity becomes a bit of a nonsense, doesn't it, given the technology that we have for putting a number of disparate bits of data together? I was going to mention earlier the Information Commissioner's view on anonymity; I just do not believe in anonymity.

Sureyya Cansoy: We had some further discussions with our member companies better to understand this point. We think that, overall, anonymity strengthens the privacy of individuals, as they are no longer personally identifiable. Of course, there is the possibility that, by merging more than one dataset, you may be able to re-identify some of those individuals. We have seen some examples, which we cited in our submission to you.

Having had further discussions with technology companies who understand this really well, their take on it is: "What is the motivation behind merging those different datasets to re-identify people?" If the motivation is just to show that, academically, it is possible to do it, then, yes, it is possible; but evidence of it happening on the ground is very limited, which is supported by the Information Commissioner's Office. However, a greater risk is if somebody is after personal data or identity for criminal purposes. According to our member companies, it would be much quicker and faster to use cyber-attacks and hacking to obtain the information than to combine different sets of anonymised data. That tells us that perhaps it is much more important to focus on the security of the datasets being held. That is the view that we got from our companies.

Q22 Mr Heath: I am sorry, Mr Miller; I interrupted you earlier.

Carl Miller: There are two quick points. The datasets that we gather, as Demos and CASM, are large and aggregated. I do not care about individuals; I care about broad social trends, so I am not trying to identify anyone. I care about how people react to a policy announcement or a speech and stuff like that.

De-anonymisation is a technology; there are ways in which you can attack anonymity, and they are being sketched out in the EU data protection regulations at the moment. It is always going to be a changing field; there will always be an arms race between anonymisation techniques and de-anonymisation techniques, to try to find the find the right strategies to anonymise when you need to that are both not too onerous on the company and also relatively impervious to it. There have been a few mini-scandals, where academic papers using so-called anonymised data were published, yet within hours people managed to work out who the people were. It is a constantly changing landscape.

Timo Hannay: A rule of thumb is that you should assume that any data out there publicly, even if it has been anonymised, will potentially become de-anonymised. You never know what technologies or other information will become available. It is dangerous to assume that so-called anonymous data will remain anonymous, so I agree with the premise of your question.

Q23 Mr Heath: Mr Miller, you mentioned EU legislation. We have the Data Protection Act. Is it fit for purpose? Does it get the balance right at the moment in terms of the burdens that it places on the industry and the protection of the individual? Will the proposed EU legislation pose difficulties or is it moving in the right direction?

Sureyya Cansoy: I shall try to answer that from the techUK perspective. Having spoken to several technology companies, we believe that the UK data protection framework is broadly fit for purpose. Having said that, we need to consider the implications of the new digital age on laws that came into force before then. There are a few suggestions from member companies on how that can be addressed, for example, in terms of guidance rather than additional legislation, which we touched on in our submission to the Committee.

In terms of the proposed EU data protection regulation, we completely support the need to have a robust, flexible and fit-for-purpose data protection regime. I cannot stress enough how important this is for the technology industry. We need to assure the customers of our members of the privacy and security of their data. We also understand, as I said, that the new digital age means that rules that date back 20 or 25 years need to be looked at again. However, we think that the proposed EU data protection regulation is not the right mechanism. As it stands, we think that it would be very damaging to businesses across industry—the tech industry definitely, but also a number of others. We believe that it would hit jobs and growth. An estimate by Deloitte suggests that the EU could lose as many as 2.8 million jobs as a result.

Q24 Mr Heath: Is that figure just plucked out of the air, or does it have some substance?

Sureyya Cansoy: I do not know the basis of the calculation. We can look into it, but the gist of it is that it will affect companies significantly. It is very evident. We have been speaking to large and small companies, British and international, and anecdotally I can give you an example.

One successful small British cloud company, which started around three years ago, grew from two to about 50. Discussing the implications of data protection on their business, they told us that if the proposed EU data protection regulations had been on the table when they were setting up their business, they would have thought hard before taking the risk. It is possible that they might have decided not to go ahead, to the degree that they thought that it would be very burdensome for businesses. It is an area where we are doing a lot of work, so if you would like further details we would be happy to provide them.

One good thing is that the Ministry of Justice understands that the current draft is not what we need. We are working closely with them to make sure that whatever regime we end up with is one that works for individuals in protecting their privacy but also allows technology companies to innovate, as well as allowing other industries to operate within those premises.

Timo Hannay: We do not find the current UK data protection framework problematic in any particular way. I am not familiar enough with the EU proposals to comment on them specifically, but I would say that for us the confidence and trust of the users who ultimately provide this information is paramount, while trying to avoid anything that is too heavy-handed and bureaucratic. The heavy hammer of legislation, although an important way to deal with this issue, is not the only way. I refer to some of the things that we discussed earlier on understanding user expectations, and identifying and encouraging best practice. Those things are just as important as legislation.

Carl Miller: The major concern that I have with the regulation is that it is hard to see how that body of abstract legal provisions will play out technologically and in practice—how it would change the architecture of social media platforms, what new technologies would be required, what judgments it would be asking social media sites or data controllers to make. There is a lot of uncertainty, which is of course not what you want when you are trying to discuss a law. Broadly, I agree with the other two panellists that there is a simmering concern that it is going to be burdensome and hostile to growth.

It is important for a clearer public case to be made about the social good of big data and social media research—all the things that it can do for people, such as delivering better health, and more agile and responsive government. That requires it to be used in areas where it is not used at the moment, and for it to be used in areas that are concerned with delivering public goods. Inevitably, it is going to come down to people balancing individual privacy with other public goods that they think are important, but at the moment I don't think people see the value of having their data being sucked up in huge quantities and crunched. Basically, they think that that is for private profit for large and sometimes quite opaque big data analytics firms, rather than for them.

Q25 Pamela Nash: I want to look at the reliability of data, and how that impacts on how the data is used. We have had evidence of concerns about data being cherry-picked by researchers to suit their own ends and the results that they wish to have. Do you share that

concern? Are tools available yet, or being developed, to help us to spot when it is happening and to tackle the problem?

Carl Miller: There is the question of inadvertent and deliberate manipulation or challenges to reliability. The larger concern that I have is not the deliberate manipulation of data, but the fact that methods across the whole of the industry are young and weak and often cannot do the job. It is not just the fault of the Government that these methods are not being used within the public service; the methodologies that exist right now cannot usually satisfy the evidentiary requirements of public policy makers.

Yes, I have concerns. The first concern is that black box techniques are used. With technologies that can crunch and understand big data, there is very little clarity for many people, including users, about how they work and what they are actually doing. With most of the services out there, you can plug in some words on one side, you cannot really tell how the calculations happen, and you might get a graph coming out on the other side. Exposing those black box techniques so that everyone can understand how that analysis happens is important, and I think it is possible. Despite the arcane and often complex mathematics and technologies that are involved, it is possible for people to understand how big data analysis works.

In terms of deliberate manipulation, we are seeing a fair amount of automated production of content on social media, so it is likely that one in 10 Twitter accounts is fake. You can go out and buy Twitter followers if you want, and I have been tempted to do it many times. You can buy millions of Twitter followers to make yourself—

Q26 Pamela Nash: I was going to ask about that, but it is different from cherry-picking the information that is available; we are talking about manipulating data that is online.

Carl Miller: On the cherry-picking point, the difficulty is that, as a researcher, you cannot publish the datasets you have studied, because it is against the terms and conditions of the social media sites. In science, you obviously analyse a dataset and then make the dataset available so that people can see how you selected the data. We cannot do that. I have never encountered deliberate cherry-picking in research, but it is certainly possible and it is hard to verify.

You have automated data being created, and many states are stepping into this game now. We have seen a number of calls, made openly and issued by national Governments, for technology to allow analysts to behave, for example, as if they were 1,000 or 10,000 Twitter users. As we see Twitter and digital fora becoming more politically and socially significant, we are going to see increased interest in being able to manipulate the content. That said, there But there are corrective mechanisms.

There are three important layers where people are working on how to identify and correct this kind of stuff. You have technological responses; the same technology for national language processing, which is used to generate automatic data like language on social media, is also used to detect it. You then have methodological responses; a lot of researchers are now triangulating data, and this is something that they are increasingly going to do. If I want to present to the Committee social media research on attitudes, I will do so alongside conventional social research, as a way of trying to compare all these new

and unfamiliar methods with the more mature, older, more conventional and trustworthy ones.

Then you have use. This is a really important one. Decision makers, whether in the private or public sector, are going to have to grow more familiar with the idea of using outcomes that are uncertain. A lot of the big data technologies that are being used are inherently probabilistic; they are predicated, for instance, on Bayesian mathematics, which basically gives you a result that has a confidence score attached to it. We are going to see an increasing epistemological shift, as it were, away from outcomes with evidence that is certain and clear, towards evidence that is shrouded and surrounded by a cloud of caveats and confidence scores. We are going to have to become more comfortable with uncertainty, as well.

Timo Hannay: We, too, are more concerned about the gaming of systems than about the cherry-picking of data. Regarding cherry-picking, we try to gather data from as wide a range of sources as possible, and we try to detect every mention of scientific research that we possibly can, but we do not get everything. Sometimes our users and customers tell us about things that we have missed and we try to add them back in. We are trying to be as comprehensive as possible, and certainly not to cherry-pick in any way.

When you are using this kind of information, and directing people to relevant material or research, particularly if you are using it in some kind of research evaluation, which will come in the future, there are concerns that people can game the system. If they think that by tweeting or writing on Facebook about their own research they will get more readers and that ultimately their funders and employers will see it, there is an incentive to try to game the system.

We also have to bear in mind that current systems, particularly around citation analysis, are gamed already, so people self-cite; publications self-cite. The answer is to use a wide range of different sources and measures, and that is what we are developing. We do not just use social media; we use mainstream media and policy documents, and we are moving on to patents, for example, to look at the impact of research. Within social media, we do not use only one source; we use as many sources as we can get hold of, which reduces the chances of gaming. In addition, we use algorithms to detect gaming. If you keep tweeting about one particular journal, we surmise that you are probably the publisher or editor of that journal, and we will downgrade those tweets accordingly.

Q27 Jim Dowd: I want to go back to what was said earlier about skills, but before doing so I want to look at the terms and conditions issue that arose earlier. I was leafing through my Sky television a couple of weeks ago, updating software of some kind or another, and I went on to the page headed "Terms and Conditions." It helpfully said up front that it was 110 pages. That immediately deflected me from pursuing it any further. Okay, we live in inordinately litigious times, and, as you said, Chair, there are all kinds of jurisdictions where these things have to apply, but it occurred to me that although it is not a deliberate attempt to mislead, it does mislead, because effectively people do not engage with it at all. That gives the company—in this case Sky, but also Twitter and Facebook and the rest—the opportunity to do whatever they like with the consumer and the information they derive from them. Is that your feeling?

Carl Miller: Some social media sites have taken some positive steps. If you go to Twitter's terms of service, there is more of an effort to state them in plain English, but you are right: there has been little incentive for many platform providers to do anything other than issue 100-page documents, because everyone clicks "Yes." That pressure, that incentive, has to come from outside the company. That is why, in addition to EU data protection regulation, which is very-rights based, a consumer protection regime is important.

It is not particularly clear what role the Government have to play in that, other than supporting it, but we certainly ought to have kitemark terms and conditions, where companies are incentivised to put in plain English in a few pages what the implications of people putting their data on those platforms really is. I do not see why it cannot be done.

Timo Hannay: I completely agree. It is a huge issue. It is an issue for me as an individual in my personal capacity, as well as in my professional capacity. I agree with Carl; we should be identifying and rewarding best practice. That is the way to do it.

Sureyya Cansoy: We agree with that. One of the key things that we are looking at as part of the data principles is the importance of simplicity and transparency, so that people understand what they are signing up to. We are completely behind that.

Q28 Jim Dowd: Mr Hannay, you mentioned the skills gap and numerous vacancies in various regards; and, Mr Miller, you said that this data-mining, as it effectively is now, is developing in an almost random fashion, with no discipline and no strategy. The Government have come forward with the data capability strategy. I am not sure how familiar you are with it, but do any or all of you feel that it is of any value?

Sureyya Cansoy: First, having the right skills is absolutely essential to maximise the benefits from social media analytics. There are some interesting numbers about how big that gap is. For example, one number from e-skills is that they expect big data job vacancies to grow by 23% annually by 2017, in three years' time.

Q29 Jim Dowd: What is the base for that? You say 23% growth, but what kind of numbers do you have?

Sureyya Cansoy: We are talking about annual growth of the actual big data job vacancies available—how many big data jobs are advertised every year and the percentage growth. That is what it reflects. I can provide further information afterwards on the number.

Another number that we have seen suggests that 57% of recruiters dealing with big data vacancies say that it is difficult to find people for the jobs they are looking to hire for, and anecdotal evidence from techUK member technology companies that we have spoken to suggests that there are some talented and skilled people out there, but they seem to prefer to go to start-up companies and are not necessarily willing to work in established organisations or for Governments. With that in mind, it is quite clear that there is definitely an issue that needs to be addressed.

Another point is that the skills issue, in terms of digital skills in the UK, is broader than just data analytics or social media analytic skills, and it is a top priority for both techUK

and the Information Economy Council. There are currently 1.1 million digital skills people in the UK economy; techUK and e-skills predict that we will need another half a million by 2020 if we are to maximise the benefit from the information economy. That is quite a substantial gap across the board.

In terms of the data capability strategy itself, it was produced by the Government jointly with the Information Economy Council, and the industry and academic institutions had the opportunity to contribute. We think that it addresses the right areas. It starts by asking what skills are needed by the industry, where are the gaps and where is action needed, and what needs to happen in schools, higher education, apprenticeships in terms of promoting the reputation of the industry and so on. It talks about industry, Government and skills bodies all having a role to play in getting it right.

We think that the raw actions are broadly right. With anything like this, the proof of the pudding will be in the implementation of those actions. One of the things that we would like to highlight is that the sector skills council for the technology industry in the UK should have an important role to play in making sure that some of those actions are implemented. There is also the importance of continuity beyond 2015, should there be a change in Government, for example, because this is not a short-term issue and actions should not be short term.

Timo Hannay: I shall speak to our on-the-ground experience, which is necessarily anecdotal but hopefully informative. We certainly find that London is a great place to set up these kinds of businesses. A great pool of talent is attracted here.

Q30 Jim Dowd: May I stop you on that very point? How much of it is indigenous or locally generated, and how much is attracted from abroad?

Timo Hannay: The Digital Science office in London has about 70 or 80 members of staff, and we reckon that there are about 12 nationalities, and probably about 20 languages spoken. They come from far and wide, and that is incredibly important. The fact that London attracts people from around the world was very significant for us in setting up here.

That said, demand outstrips supply, and we expect demand to go up because of some of the macro-trends that we have been discussing. We are based in King's Cross. Google are building their new UK headquarters just down the road. They are going to be employing a lot of technical people; we hope that will attract people, but also that it will provide competition for the best staff. Altmetric, which I mentioned earlier, currently has 11 staff but seven vacancies, of which five are technical—essentially developer and data analytics-type roles. That gives you some indication of the difficulty that we are having in getting the high calibre of person that we are looking for.

Carl Miller: Within data science there is a shortage, but that is understandable; there has been an explosion of this new discipline, and suddenly incentives have changed. We are seeing secondary schools with the new computing syllabus, and both undergraduate and postgraduate courses are changing—quite rapidly for the higher education sector—to meet these new demands.

The concern I have is that at the moment, in terms of skills, social media analysis is being treated like a branch of big data, or data science, but strictly speaking that is not true at all. In my view, social media analysis requires a hybrid skills set. This is both a social and a computational science. It requires people who understand both culture and people, and also all the new ways in which we can handle data that is unprecedentedly dynamic and large. At the moment, my concern is that if you simply see social media as part of the "eight great technologies"—if you simply see social media research as part of the UK data capability strategy—we are going to get very good at counting things from social media, but we are not going to know very well what we are counting.

Within the field of research, which is feeding all the economic activity that is going on, we are seeing the numbers game galloping ahead of the squishier, softer, slower social scientific work that happens underneath that, and that is going to be an important brake on growth. That is the reason why it is not being used more broadly across the economy.

Q31 Jim Dowd: There is no real attempt to grow specialism itself; it is just an assumption that we will get talent from elsewhere, which can be adapted.

Carl Miller: There is no consistent attempt or incentive to break down the disciplinary boundaries that currently stand in the way of getting the hybridised skills sets required to do the everyday job of getting insight from social media. It is not something that computer scientists can just sit down at a computer and build algorithms for. Social media is a new cultural space, and it plays by different linguistic rules.

To understand behaviour on social media—to do things like predicting what is going to happen in future, to model and to understand attitudes—is a new and distinct discipline, and it forces the meshing together of both social and computational sciences. My worry, and the reason it has not yet been applied beyond the advertising, marketing and retailing industries, is that if you are a policy maker you get a smorgasbord of metrics—raw metrics, numbers. You will get some numbers about how many people were positive or negative, and you will get some numbers about how many shares were happening, but in the social sciences the serving up of raw data is the beginning of the story, not the end. No social scientist would ever accept the presentation of data as an adequate description of what is going on. We have to throw these disciplines closer together. That is a challenge for the higher education sector above all, but I am concerned about this being treated basically as a small branch of data science. In my view, it is not that.

Sureyya Cansoy: We think that different skills are needed in the technology industry, which produces the software and tools to enable the analytics, and perhaps different skills are needed in the organisations that use the software and tools to come to the sort of insights that they need for their business or organisation. I agree with Carl that we need a mix not only of technical skills but, depending on the business that you are in, perhaps a knowledge of the business that you operate in and the kind of insights you are trying to get with that analysis.

Chair: Thank you very much for such a comprehensive set of answers. I am sure that there will be other things that you can feed in as the inquiry goes on. Thank you.

Examination of Witnesses

Witnesses: **Professor John Preston**, Professor of Education, University of East London, **Professor Mick Yates**, Visiting Professor, Consumer Data Research Centre, University of Leeds, and **Dr Ella McPherson**, Research Fellow, University of Cambridge, gave evidence.

Q32 Chair: Good morning, and thank you very much for coming. I think that all three of you were listening to at least part of the earlier exchanges, so you have the gist of our direction of travel. First, I ask you to introduce yourselves for the record.

Professor Yates: I am a professor at Leeds University, and I work on the ESRC-funded consumer data centre. In a previous life I was at Dunnhumby, which you may or may not know is the company behind the Tesco club card.

Dr McPherson: I am Ella McPherson. I am an ESRC fellow at the University of Cambridge in the department of sociology.

Professor Preston: I am John Preston. I am professor of education at the University of East London. My area of specialism is disaster education and public response in disasters.

Q33 Chair: We are talking this morning about social media, and one of the areas that we are particularly interested in is how Governments and governance could be improved by smart use of social media. Is it possible to see some gains?

Professor Yates: First, there is a distinction that was not made by the first group between social media and big data, in the sense that social media, generally speaking, is unstructured—the photograph that you put on Facebook, the colours of your shirts and so forth—whereas most big data historically is structured datasets, such as till records, census records and so on. You have an opportunity to take the structured and the unstructured.

In a commercial sense, if you knew what colour shirts you liked, you could make a better promotional offer, based on your Facebook pictures. In the social science sense, you can do the same thing: attitudes. What are people's attitudes to different kinds of food? Where do they go? How do they visit and what methods of transportation do they use? That will come out of social media, but it would not necessarily come out of structured datasets. If you put those two things together, you have a rich way of thinking about consumer behaviour, which goes well beyond business.

Dr McPherson: Our perspective on this is that social media data is a complement to other sources of data that the Government are already using. We think that it should be used in addition to, not as a replacement for, other methods of inquiry such as interviews and surveys.

Professor Preston: In terms of my research, which is on how social media could be used in a disaster or emergency, I would agree that it is not a panacea for Government, but you have to be careful in the ways you use it. It does not work, for example, in warning and informing the public of a crisis, where old media such as radio and television, are much more effective. It does not work in terms of helping people to organise themselves in terms of a large-scale event like the evacuation of a city; there we found that social media could disrupt the evacuation and cause congestion. Where it does work is in terms of

helping policy makers in Government understand how a situation unfolds, to look at how populations are responding during a disaster or emergency and during the recovery, to put resources where they are best used. You have to be careful how you use social media in an emergency situation, but it is of use.

Q34 Chair: Leaving aside disasters, those are interesting comments on how other situations unfold, evolve and develop. Are there examples, outside the ones that Professor Preston gave, where social media could be a useful tool in planning Government responses?

Dr McPherson: One area where social media has proven to be very useful is in understanding what is happening in closed societies or countries under conflict. You were talking before about issues to do with data manipulation. That is always a problem, but in scenarios where previously it had been difficult to get information out of situations, social media is proving to be a way to access it and to get some idea of what is going on in those scenarios.

Professor Yates: You could also use social media to understand how ideas flow through systems; people swap ideas, change ideas, improve ideas. As a way of gauging innovation in a system, you can look at what is happening and what people are saying. In theory, you could also use it to stimulate innovation if you were part of the process of communication itself.

Q35 Chair: I ask this question as somebody who is over 65, who has a Twitter account and Facebook account and who is "LinkedIn." Are there problems interpreting the data about some demographic groups within the population?

Professor Preston: Certain demographics use social media a lot less, and even if people count themselves as social media users, we distinguish between heavy users of social media and those who might have put down only one or two tweets in their lives. It is not only an age thing; it is also socio-economic status and ethnicity in terms of the users of Twitter or other social media, so you have to be careful when you interpret the results. That is why it does not work very effectively as a method—

Q36 Chair: Surely that is true of any data source.

Professor Preston: Not necessarily. If you look at the cohort studies data, for example, that the Economic and Social Research Council fund, it is designed to be representative of the population. The millennium cohort study, for example, is designed in such a way as to be representative, whereas Twitter is based on users who want to tweet, or not.

Dr McPherson: That means that it is important to understand who the user base is, and to have research into that. It is always important, as I said, to combine it with other sources of data.

Professor Yates: It is blindingly obvious, but people use social media deliberately to share, and they choose what to share. That tends to be a younger person's thing, but it also applies to the 65s. That is the difference with some of the more structured datasets, and it

is where social media is completely different. People have taken the decision to share information about themselves, positively, with the rest of the world.

Q37 Stephen Mosley: I am interested in whether the UK Government can use social media data and real-time analysis to react to threats and to things that are happening within the UK. Do you think that the Government are properly equipped to use social media information and real-time analysis to react to emergency events?

Professor Preston: In our analysis, we looked at three cities in terms of local government: London, Birmingham and Carlisle. They each had different strategies in their use and interpretation of social media. Birmingham was very much ahead of the curve in using Twitter, Facebook and YouTube; London, being the seat of Government, was quite top-down; and Carlisle said that they did not want to use social media that much—it was very much face-to-face contact and radio communications that they wanted to use. It depends on the city's orientation as much as anything. Birmingham has changed recently. The resilience team has changed, so it depends upon officials who have an interest in doing this for it to be a success.

Policy makers find that they lack the tools, or the confidence to use the tools, to do this in real time; for example, during an emergency. There are tools that enable you to analyse sentiment—how the population is feeling emotionally in a crisis—and that can be incredibly useful. Are people fearful? Are they anxious and what are they anxious about? In real time, policy makers would not really want to use that tool because they would be scared of the implications; they would be scared of making the wrong decision on the basis of social media data.

Q38 Chair: Before we move on, may I test that a little further? Have you undertaken any research around communities close to COMAH sites that are used to emergency planning processes?

Professor Preston: In a more recent project that I am doing on population response to infrastructure failure, populations close to COMAH sites are used to using social media and looking at social media in terms of what is going on, whereas populations that are not near COMAH sites are not as used to it.

Q39 Chair: They are more aware of the likely effect.

Professor Preston: They are more aware of the effect and of the alerting protocols.

Dr McPherson: May I add something about its use in crisis or emergency situations? I am concerned about using social media to establish events—to establish what has gone on. The point that we made in our written submission was that using data sources requires verification. You need to do a technical verification of social media information that relies on particular technologies, and that requires knowledge and understanding of how social media data can be manipulated. It takes a human element, because at the end of the day establishing the truth is a subjective exercise. Moving from social media data telling you

what is going on in an event to reacting in real time I don't think is possible, because you need to put verification in the middle and that takes time.

Professor Preston: There are two verification methods for social media in real time. One of them is to use the users themselves. In the Australian floods, people were using the hashtag "mythbusters." "Mythbusters" is a television programme about busting myths, conspiracies and so on, and people were using the hashtag to say what was real and what was not. That became a trending hashtag, with people using it to ask whether the flood was really happening in an area. People actually made use of that hashtag. Another one is the use of real-world data, using social media in conjunction with CCTV data or other kinds of situation awareness that the emergency services might have. But I agree that it takes a while to establish what is true and what is not. One of the things that we fear in emergencies is that people, or even terrorists or subversive bodies, might set up slop-bucket accounts to tweet that something is happening when it is not occurring, in order to make the situation worse.

Q40 Stephen Mosley: That was the area I wanted to move on to next. Is there any evidence, in the situation that you just highlighted, of third parties trying to encourage other people to do things?

Professor Preston: One of the biggest examples is the Syrian Electronic Army, as they classify themselves. No one knows who they really are, but they tweeted that the White House had been destroyed and no one knew where Barack Obama was, which caused billions to be wiped off the stock market in the United States. People are using it for disruptive activities at the moment.

Dr McPherson: There is an absolutely huge amount of evidence emerging about this in Syria, but in other cases as well. I was looking recently at a video of someone being shot by a water cannon. It was being circulated in Latin America, with the title "This is happening in Colombia," but there was the same video in Mexico. The problem that this raises is that of verification. Even if they are ultimately disproven, it requires a lot of resources for groups to debunk them. It takes time, and it takes people away from other work that they could be doing, so it is quite disruptive.

Q41 Stephen Mosley: Looking at the UK, have there been any cases of that? I am on social media quite a bit, and see these stories circulating. I guess there is no checking, so how do you debunk these myths?

Professor Preston: Twitter can be self-correcting. In our project, we had an analysis of a plane crash in Cork in Ireland. Social media got there first, in terms of how many casualties there were, before other media. It tends to be self-correcting so, first, people were tweeting that there were 20 casualties, then someone said that there were no casualties, but both of those were debunked. They iterated to the right—unusually, before the old media got there and the BBC broadcast how many casualties there were. People act almost like scientists sometimes. It is not just rumours; people say, "Is this correct or not? I know someone who works at the airport."

Dr McPherson: There was some great research done by *The Guardian* in conjunction with the LSE about the rumours circulating during the London riots. They showed timelines of tweets supporting the rumour and then debunking it, and how that worked, so it is self-correcting.

You asked how you might verify. This is an emerging area, and I have been looking at it in terms of human rights organisations as well as journalists, and guidelines have been in development in recent months. The fundamental approach that verification professionals take to this information is to assume that it is false and your job is to prove that it is true, and if you can't prove that it's true you do not use it. I remember some quote that it is less about the snazzy technology than it is about good old gumshoe detective work. You might try first to find the source—the person who posted it—and speak to them directly and ask them details about it, just as you would with any other bit of information. It then needs cross-referencing with other databases. Technology can help; for example, you can use satellite images to verify whether large-scale bombing has taken place. You can also look at weather databases and ask the person, "What is the sun doing right now where you are?" and cross-reference to see whether they are actually there, but it is a complicated process.

Professor Preston: It is also important to remember that old media can be hacked. The emergency alert system in the United States has been hacked several times in the last few years, broadcasting erroneous messages at state level. It is not just new media that can be subverted.

Q42 Stephen Mosley: Lastly, do you think that social media platforms should be obligated to share information with Governments when there are concerns about security issues?

Professor Preston: In disasters and emergencies, they are willing to do so readily anyway. Most users have an idea that they are sharing data, and they would like it to be used for altruistic purposes. If there is a disaster or an emergency, people are willing to do so. With the Boston marathon bombing, for example, people were willing to share their data and images with the authorities. People are less willing to do so if they think that the security services are spying on them for other reasons.

Q43 Stephen Mosley: Or if social media is being used to encourage that trouble.

Professor Preston: Yes, absolutely.

Q44 Jim Dowd: I want to look briefly at how real-time analysis can be used in non-emergency circumstances. First, I want to test what you said, Professor Preston, about Twitter being self-correcting. Surely it is not self-correcting at all. It undermines the authenticity and the believability of Twitter the more conflicting messages you get. If everybody is tweeting—unless they are doing it maliciously, as Stephen suggested—"I believe this to be true," what are people going to do? I think that they will lose confidence in the whole system.

Professor Preston: People do not behave like that. They look for corroboration. If you saw a tweet saying that there was a fire at King's Cross station, people would say, "Well

you're saying this, but what evidence is there"? And then they would look for someone who had tweeted—

Q45 Jim Dowd: In other words, they do not trust it.

Professor Preston: Not necessarily initially; there might be a bit of overshooting, but eventually it comes to be corrected over a period of time. That happens more rapidly as social media progresses.

Professor Yates: In effect, you could argue that they do trust it, because they are looking for sources of corroboration from the very same system, and they assume that the system will provide the correct answer, so generally speaking—

Q46 Jim Dowd: Are you suggesting that that is a false assumption?

Professor Yates: That is the way that people actually use it.

Dr McPherson: My evidence has shown that it is not at the system level that trust is taking place; it is looking at every individual user and deciding whether or not to trust that user. You look at the information but you also look at the source, and you do not trust the information unless you trust the source.

Professor Preston: As a small example of that, we did an exercise with the Olympic Delivery Authority when the Westfield shopping centre opened. We were looking at different kinds of data. Someone was tweeting that one of the gates was shut and you could not get in that way, and people were saying, "Maybe someone is trying to get in more quickly," but once they had tweeted a photograph, there was some corroboration and people started to believe that the gate was actually shut.

Q47 Jim Dowd: I accept that; over time, if the same message is repeated, it becomes the norm.

Professor Preston: No information is correct at first shot. If we heard that there was an emergency in London or elsewhere, the broadcast media would probably get it wrong.

Q48 Jim Dowd: A lot of social media users are the least likely to believe the Government or any public authority.

Professor Preston: Maybe, yes.

Q49 Jim Dowd: Can we go back to non-emergency uses of real-time analysis? Are there any circumstances where it will be of benefit, other than in emergencies?

Professor Yates: It is not a real-time analysis of social media, but the US amber alert system on missing children is quite a good system. You can get real-time alerts sent to people's phones through tweets or whatever, in whatever way people have signed up to the

system, and they provide information to the highway patrol. It is a small slice of use, but it is a clever use and it is well accepted in the US.

It is a way of communicating as well. It is not just, "I've got a disaster. Let me understand the disaster," but what you would like to communicate to people. To what extent are the Government prepared to open their communication channels to talk to people? I go back to the question of trust. I am not sure that we know each other, so I am not sure that I trust you, but maybe after an hour I might, because you have been saying sensible things that corroborate—

Q50 Jim Dowd: But I am a Member of Parliament.

Professor Yates: Sincere apologies. My general point is that we as human beings trust each other over time by understanding what is being said and the context for what is being said. In fact, social media provides you with lots of different personalities. On Twitter, for example, I am reasonably well regarded—believe it or not—for photography, but I am probably not well regarded on politics. People know the context of my personality in that social media, and that is how people use it. Government could do the same thing. What is the personality that you want to project, in certain circumstances and situations and in certain media, to communicate with the public at large? It is a positive opportunity rather than a negative one, if that makes sense.

Professor Preston: Another example, away from disasters, is looking at flu or the spread of infections. You can use social media to work out where outbreaks are and when there will be a demand for antibiotics and so on, and use that to predict where the demand will be on GPs and pharmaceutical services.

Q51 Jim Dowd: Can it be used to replace or supplement the traditional census?

Professor Yates: You can supplement it but not replace it, for all the reasons that everybody has given so far. The census will provide a broad look at everything, in a structured and detailed way, and you cannot replace that with social media, as to a certain extent it self-selects. Social media can create user attitudes—even colours, if you like; different things that are not available in traditional census mechanisms—and it can be extremely valuable. It is one of the things that we will eventually do in the centre at Leeds.

Q52 Jim Dowd: You are clear that it could not replace the census.

Professor Yates: It could not replace it, no.

Q53 Jim Dowd: The civil service has traditionally not been the most responsive or adaptive of institutions—for good reason, I hasten to add; I do not say this as a criticism. How amendable do you think it is? Do you have any evidence that they will be able to utilise this either at a national or local level?

Professor Yates: Can I address that by stepping back? The biggest barrier to the use of big data and social media analytics in general is the understanding and management of what it

is capable of. We have a skills shortage, and I would like to come back to that later, but we also have an understanding shortage, even in industry, about how you can use analytics and data to make different kinds of decisions. In that sense, the civil service is no different from industry at large. The system needs an education job at principal level, and on bringing in the right people to do that. It is an "experiment and see" mentality, perhaps, but it is the same problem that, frankly, you will find in most big companies.

Q54 Pamela Nash: We had evidence from the University of Cambridge touching on some of the themes that you discussed with Mr Dowd, saying that real-time analysis is incompatible with verification. That is common sense, and we did not need to ask the University of Cambridge for that. Does that rule out its use by the Government for civil policy purposes? How do we get around that?

Dr McPherson: It does not rule it out at all, but what needs to be in place is a rigorous verification system. There is a new publication called "Verification Handbook," which speaks to a lot of experts in this area about the techniques that they use. First and foremost, it says that whatever verification system you are going to use should not be put in place when you have an emergency and want to use the information. It should be systematically thought out ahead of time. It is something on which participants can be trained. It has ethical aspects as well, which should be included in that training. I don't think it rules it out; it just means that verification needs to occur, and that verification should be based on a plan.

Q55 Pamela Nash: Can you tell us more about what you mean by that? These are situations where I imagine that it would not be easy to plan ahead unless, as in another inquiry we have done, you have good horizon scanners where you could use Twitter or other social media to analyse an unfolding situation. How do you plan ahead to use that information?

Dr McPherson: There is a methodological level of planning, which is understanding how verification works, and it can be done with social media; but there is also how this sort of data would fit into existing Government bodies and their existing sources of data. You can imagine that the police have all kinds of sources of information, including on-the-ground contacts and networks. You would then get social media data, and try to verify it. You would plug it into your existing on-the-ground networks, with whom you can have quick telephone or face-to-face contact. These will be people that you trust in that community, and you can ask them to verify it for you. It is a system of triangulation that, at its basis, should be used for any source of information, but I want to highlight that it needs to occur in this situation because social media is particularly vulnerable to certain types of manipulation.

Professor Yates: There is also a distinction between individual verification and mass verification. With mass verification, you have the same techniques as used with other data sources. You are looking at a vast quantity of information, and trying to figure out the truth from it. That is a bit different from identifying whether we said a certain thing ourselves at a certain time or a certain moment.

From a broader sense, social media can be extremely useful because it is very fast. You can look at the way that some companies use data. I am not sure if you are familiar with

how Google do translation, but they favour quantity of data over accuracy, and that is how they get better results. They have literally digitised every document that has been translated and figured out the context. If I hold up a candle and say, "This is light," does it mean that it is illuminating or that it is not heavy? It will depend on the context in which I hold up the candle. The software deals with that by crunching vast amounts of data to get to a very fast response, which is why we all get instant translation on our phones. The same kind of techniques can be applied to social media.

Professor Preston: There are other uses outside verification. Whether the data is verifiable or not is one issue, but in an emergency situation people go through different stages. We found that at the first stage they use social media to seek information. They then go through an emotional stage and then an opinion-sharing stage.

If you know that as a policy maker, you can tailor your types of message to those different stages. Which stage are the public in at the moment? If they are seeking and sharing information, you want to put information out there, but if they are responding emotionally then you want to put something out about being concerned or being compassionate about the victims of the emergency. If they are sharing opinions, you want to deal with that side of things.

Another facet, away from verification, is that some people on our project came up with a tool to detect outliers on social media during an emergency. If people are saying things very different from anyone else, it might be interesting for you to pick up on it, even if it is not verifiable. They might have access to information that others do not, or they might be an outlying group with an interest in the crisis.

Dr McPherson: What you said reminds me that whether or not it is true does not have a bearing on whether it has an effect. It does have a bearing, but it may not have a bearing on whether or not it has an effect because it depends on what other people who are looking on social media believe. Even if you eventually debunk some kind of rumour, if it is circulating and people are believing it and reacting to it, that is another area where the Government might be concerned about responding.

Q56 Pamela Nash: In terms of the processes that you have described, Professor, and the nuances that you mentioned about the different stages of people using Twitter, are the UK Government doing that already? If not, how do we get ready to make full use of it?

Professor Preston: In my conversations with civil servants, they are certainly interested in dealing with it, and emergency managers are interested too.

Pamela Nash: That was a very political answer.

Professor Preston: Yes. They are interested in doing it, but they do not feel that a tool is available that Government could necessarily trust. In terms of sentiment analysis or analysis of emotions during any sort of political situation, they would like a tool that is kitemarked—one that Government could trust—but it is not there at the moment, so that makes them wary of doing it, whereas at the local level, you see people experimenting with all sorts of different things in local authorities and local government.

Q57 Pamela Nash: Is there any incentive for anyone to produce that tool?

Professor Preston: There is a commercial imperative for someone to produce a verifiable tool that Government could use and trust. In the States, the Federal Emergency Management Agency is very big on engaging the private sector in producing these sorts of things, with Government contracts and so on.

Q58 Mr Heath: Professor Preston, you mentioned observations in a flu epidemic, which reminded me that when I was a Minister at DEFRA we were dealing with Chalara fraxinea, ash dieback, and one of the things that we had then was an app and an ashtag hashtag, which we were rather proud of.

Chair: How much did you pay for that?

Mr Heath: I wonder whether there is greater scope for observational science, using citizens as observers, in collecting information about natural phenomena such as epidemics. If so, won't we need better protocols for interpreting the information that comes in? Obviously it is partial and some of it is incorrect, but it nevertheless has some value.

Professor Preston: There is great scope for doing that. In my example about Birmingham and how Birmingham Resilience organised things, they depended on people at a local level to say whether events were happening and to tweet and use hashtags and so on. There is huge scope for that. Although the data may have observational errors, most people are minded enough, in terms of helping the scientific effort, not to skew the data. I cannot think of any example of a scientific effort that has involved Twitter, like spotting bird species and so on, where people have deliberately skewed the data.

Q59 Chair: Does that conform to the Google rule about collecting volume? Every now and then, the pinpoint data might be inaccurate but the general pattern is correct. For example, David's constituency was badly affected by the recent floods, and an individual Twitter message might have been inaccurate, in the same way as an individual physical measurement might not have told the full picture, but the collective pattern does.

Professor Preston: That is right. You can get sampling errors.

Q60 Mr Heath: I want to talk about the collection of personal information, and the disparity between the preparedness of the individual to give personal information to companies, whether they are social media sites or private or industrial, and their lack of preparedness to do the same for the Government. This may all be down to George Orwell, or there may be something else.

Professor Yates: It goes back to what we said before. In social media, people make a conscious choice to share specific information for a specific purpose—some of it silly, maybe, but some of it for a useful purpose. They have already decided to give that, and they know that it is going to be used in some way, shape or form, by other people on the system or by the company itself. That is a bit different from being compelled by law to give information about your tax return, your expenses or whatever. There is a different

sense for the human being about how you deal with that information. In one sense you are doing it freely, but in another sense you are being compelled. If you can get people to do these things freely, that would be incredibly helpful. There would then be no sense that they were being imposed upon.

Q61 Mr Heath: Quite so. We freely give information to our doctor about our symptoms and our personal details, but there was marked reluctance by some for that information to be shared for purely benign research purposes in the NHS database.

Professor Yates: I completely agree, but that is a failure of education. It goes back to the point of to what degree we are educating the population at large about the value of using data to take better decisions on behalf of all of us, in many different ways. We do not do that very much. Even in my business school, we teach analytics almost as a statistical activity, whereas we should be teaching it as a contextual activity about how to use the insight to get to a different place. Although it might be hard to convince the population that it is something they should truly understand, we have to start that process. Had we had that education in place, I contend that getting informed consent would have been a lot easier.

Q62 Mr Heath: That is helpful. Thank you.

Professor Preston: There are platforms available where people can share information for disasters. There is one called Ushahidi, where people are very willing to share their information, but it depends on the level of government you are talking about. People might be wary of sharing information with central Government, but if it is about helping in an emergency at city or local authority level, people might be very willing to share information.

Q63 Mr Heath: We talked with the previous panel—I think you were all listening—about the application of our own data protection laws, but also the potential extension of the EU legislation. Are we going to inhibit our Government's ability to use data effectively by perhaps increasing the protections available under data protection laws?

Professor Yates: That is a difficult one. I have to state up front that I am more in favour of the US style of using data to innovate rather than some of the rights-based approaches, although obviously there is a balance between the two.

I was at a conference recently where a gentleman called Steven Finlay, in a new book, talked about rules on how to deal with personal data. He said that it depends on the data itself. How immutable is the data? Your DNA cannot be changed and it should be completely protected. To what extent does it benefit you as an individual or society at large? To what degree should we protect it? How will the data impact me as an individual? If it is going to impact me a great deal, I really should be protected.

There must be some sort of balancing act with the needs of the individual. Just to paint it as, "I must protect the individual's information" is too broad; to what extent does the individual really need it to be protected, and can we protect it while also allowing

companies and Governments to use the information in a sensible way for the benefit of people at large? I am not sure that any legislation deals with that.

Q64 Mr Heath: That suggests that you think that personal information is in common ownership rather than the ownership of the person to whom it refers.

Professor Yates: No, I tend to think of it as personal. I am sorry if I suggested otherwise. Your DNA is very much your own information. Can it be used by doctors to help you? Of course it can, so to what degree does it belong to the doctor or you? I am not really sure, frankly.

Professor Preston: Doing research is a legal and ethical minefield, and data protection legislation is obviously a central part of that. Anything that could make innovation or academic research easier in this area would be welcome.

Q65 Mr Heath: It would probably do the reverse.

Professor Preston: That is right, yes.

Professor Yates: I have an observation on that which might be helpful. I was thinking in preparation for this session that, in the legal world, there is the Creative Commons approach to you releasing your IP on the web, with certain restrictions on it. The idea of the kitemark was suggested, but I wonder whether, from an individual point of view, we could have some way of saying, "I am prepared to share this," with some kind of licence, "but I am not prepared to share that," with another kind of licence, so that the individual could positively take a decision about which data they want to share and in what context. That would be very helpful.

Q66 Chair: We have talked about licensing regimes. Taking the simple example of the club card, the trouble is that if you want the points you have to sign up to the terms and conditions; if you want to be on Twitter, you have to sign up to the terms and conditions. Companies, under current legislation in this country and elsewhere, have all the aces.

Professor Yates: They do, in a way. That is absolutely true.

Q67 Stephen Metcalfe: Following on from that point, I want to look at users' expectations of how their data will be used. We talked earlier about end-user licence agreements. To show how helpful they were, Jim told us that there were 110 pages. Are they designed to be as accessible as they appear, so that you can hide all sorts of stuff in them? Is there a way of making it easier for people to know the key issues that they are signing up to—the bit that you would not expect to be in there, as opposed to what you would expect?

Professor Yates: I think that there is. It is almost like what are the bullet points of the agreements. I completely agree. If you take the example of a loyalty programme, whatever the terms and conditions and the detail, people are making a conscious exchange. They are allowing the company to use their information on what they have bought or the planes that they have flown on in return for something—some kind of discount, some kind of service

improvement or whatever. That is why people sign up for those programmes. People understand the idea of exchanging their data for a benefit. If those simple things could be explained in those 110-page agreements—"What is the benefit of Twitter sharing my information?", "Why would Facebook share my information?"—it would be helpful.

Professor Preston: Socially, people treat social media a bit like they treat the pub. They feel that if they go into a pub and have a private conversation, it does not belong to the pub; it is their conversation. They interpret Twitter or Facebook in the same way—as a place to have a conversation. People need to know what they are signing up to, but because they are US companies, the advice would be exhaustive and legalistic, and their lawyers will have advised them that the terms and conditions should cover any situation that could potentially occur.

Dr McPherson: Even if you know and understand the terms and conditions, there is a difference between the theory of what you are signing up to and the practice of how you use it. In my own Twitter use, for example, I think of my followers when I tweet; I am not thinking that it could be used to draw conclusions about other things, or that it might end up travelling the world or in a newspaper. I read an interesting piece recently that advocated considering Twitter users as what in journalistic ethics is referred to as "inexperienced sources," who would not necessarily know what the personal implications are for them of their information going broadly public, and emphasising caution.

Professor Yates: Another side to this is that when you log on to a website and you sign in with Facebook or Twitter, not everybody realises that you are allowing the other site that is using that login to access your Facebook or Twitter information. You get a thing that says "sharing your profile," but that is through the API, so you can access profile information from Facebook or Twitter without paying Facebook or Twitter, just by saying, "I sign up for this." It is not a barrier to understanding that individual.

Usually, there isn't informed consent. If you sign up and share your Facebook profile and photographs or whatever, they then become part of the other person's system. In the mobile context, that is called a token. That token is long lasting. It is there for a long time; it does not just disappear.

Q68 Stephen Metcalfe: As is often the case, I suppose it is getting people to catch up with the speed at which the technology is developing and to get social attitudes to change. As politicians, we would want to put everything that we ever write, say or tweet in the context of how we would feel if it was on the front page of the local or national newspaper. If you do that, you protect yourself. With the pub analogy, if you are having a private conversation but someone is standing next to you and you are aware that they might be listening, you probably would not think that the pub is recording it for security purposes.

We have to help people understand how the world around them is changing, but do we need to do that on their behalf, from a Government point of view, to protect them—to create systems that avoid them making these mistakes, or do we need to educate people about the freedoms and choices that they have now?

Professor Yates: Education, education, education. To me, that is the most important thing. You talked quite a bit about data scientists in the first session. Unfortunately, data science

in most countries tends to get defined as a technical activity. As it was originally defined by IBM, it was a combination of analytics and artistry, so that you understand the business or social context of what you are doing. There is a great data science course in America at the Illinois Institute of Technology. We have one at the University of Bedfordshire, but even my institution does not have a course where we train all these people across the board on how to use analytics and insight to take public decisions. Education has to be the answer rather than legislation.

Q69 Jim Dowd: Following on directly from that, you said earlier, Professor Yates, that you wanted to come back to skills and training in these matters. I asked the first group about this. It is a new science or discipline—whether it is science or guesswork, I am not quite sure. How are we adapting to it? How is the provision of courses for professionals in this area?

Professor Yates: The data capability strategy that was mentioned before is a good document. It is robust and it points things in the right direction, although largely from a technical point of view; it does not necessarily address the social, business or contextual understanding that we also need. If you look at the data, you get the insight, but the visualisation and impact of the data is more important than the data itself. How do we train people in that? We do not have enough programmes that do that. We have some in the DTCs—the doctoral training centres; there are 21 in the UK. The White Rose in Yorkshire is doing a big data programme to address some of these issues. We need more of those programmes. That one is aimed at doctoral students, not at students in grammar school, high school or comprehensive school, so we probably need that kind of programme too. More education at different levels needs to be addressed.

Don't get me wrong: this country is probably one of the leaders in some of this analytics work, largely because of what companies like Dunnhumby have done in the retail space. We know how to do the work; the shortage tends to be finding people who can put it into context, not for doing the analytics itself. That is where we need to focus.

Professor Preston: In terms of skills, it is a very interdisciplinary area. The benefit of doing our project on social media in emergencies was that it brought people together. I am from education, and there were economists, physicists, mathematicians, computer scientists and people involved in language processing.

It is partly a reorganisation issue—how we get people from different disciplines to work on this complex interdisciplinary problem. It is not necessarily a matter of new resources; it is a matter of how you organise the resources that you already have. As my colleagues said, we already have great potential in the United Kingdom in terms of social and physical sciences.

Q70 Jim Dowd: One of the broader problems that the Government have in highly technical and advanced areas—in the field of IT, for example—is that they cannot compete in attracting skilled labour into their service because it is so commercially valuable. Is this not likely to replicate itself in the case of data analysis and big data? It is of such enormous value out there that the Government could not possibly compete on labour rates and so on.

Professor Preston: There is a possibility of strategic partnerships, working with universities and other innovators and with private sector innovators to make these kinds of products, or whatever you want to do. It is not necessarily a matter of the Government funding it all themselves. I don't think that works. If you look at the Department of Homeland Security and FEMA, their approach is very much to have science centres in the universities to innovate in this area. That is one approach.

Q71 Jim Dowd: I am not sure about the record of the Department of Homeland Security in building their new headquarters. There is not much faith in what they are up to.

Professor Yates: The UK has a great reputation on open data; it is one of the world's leaders, so the Government have a good reputation in this field. It is not a bad reputation; it is a question of building on it with the kind of partnerships that John is talking about. That would be helpful.

Q72 Chair: Professor Preston, on your observation about the multidisciplinary nature of the skills needed, do the research councils get it now? Is there a joined-up approach within the research councils to help fund projects that fit into this category?

Professor Preston: Yes. Our project was jointly funded by the Engineering and Physical Sciences Research Council and the Economic and Social Research Council. It brought together those two bits of expertise in one project. Those sorts of activities help. The research councils are getting it in terms of big data.

Professor Yates: The project that we had was funded by the ESRC. We would not have been awarded a bid unless we had put a multidisciplinary bid together inside our own institution. It was almost forcing us to think in an interdisciplinary way, which was a good thing.

Q73 Stephen Mosley: Talking about skills in the UK, do we have the infrastructure in place to take advantage of the market?

Professor Preston: We are trying an approach of concentrating on a few universities or providers to deliver those skills. Whether that approach will work in practice, I am not sure.

Q74 Stephen Mosley: I didn't mean the skills. Do we have the physical infrastructure—the data centres and the like?

Professor Preston: I could not answer that, sorry.

Professor Yates: Generally speaking, we are pretty good on that, although I would like higher-speed internet in my village.

Stephen Mosley: Get your postcode on the record while you're at it.

Professor Yates: Sorry. That was a flippant comment. To be serious, generally speaking, we are pretty good at building decent infrastructure, and some of the initiatives that the Government have taken in this space have been quite good. Commercial interests demand that you do not just have British infrastructure, you have global infrastructure; systems like Hadoop and so on are cloud-based, so you do not need the physical infrastructure in a particular jurisdiction or country.

Q75 Stephen Mosley: With the increase in regulation from the UK or Europe, do you think that this sector could be a flash in the pan that will evaporate when regulations choke the volume of data that is being collected?

Professor Yates: No, very simply. It is an unstoppable train, given the amount of data that we are collecting on everything. Soon we will have an internet of things, where our fridge will talk to our car and remind us to go to the supermarket. It is an unstoppable train.

Professor Preston: Companies like Google operate almost above the regulatory environment in Europe. They will say that if they cannot do it in Europe, they will take academics to the US and work with them there.

Q76 Jim Dowd: What if Google runs the world? What will happen then?

Professor Yates: Perhaps they have already taken over the world.

Q77 Stephen Mosley: Do you see the threat from regulation to be for home-grown UK and European-based companies rather than global companies? Would you draw that distinction?

Professor Preston: It could threaten what UK and European companies are doing if we are not careful, whereas the US has a looser regulatory environment for companies in this area.

Professor Yates: That is generally true; it is a looser environment that is more focused on driving innovation than on protecting the rights of the individual. That is an obvious distinction. At the same time, we should be quite pleased with what the UK does in this field. We are good at analytics and social science, we have some really good universities and we are quite good at being interdisciplinary. As I said, a lot of these technologies are cloud-based, so they do not need physically to be in the next city. They are being used in different ways, and the more that we can encourage innovation to use those systems, the better. We are going to be fine.

Q78 Chair: Do you agree with that, Dr McPherson? At least one Cambridge academic who speaks a lot on this sector expresses great concern that we will one day produce a less benign Government, and that all our data will be misused by the state. But you have academic freedom here.

Dr McPherson: Personally, I am concerned with the ethical aspects of using this data, with a lot of topics that were brought up by the previous panel about the inability to be truly anonymous and the inability to future-proof. I know that there are a lot of cases

where we are talking about using the data for innovation and so on, and for very positive reasons, but I am thinking of the more vulnerable people—I am thinking of my human rights work—those for whom exposing their identity has tremendous repercussions. That is the area that concerns me.

Q79 Chair: You accept Professor Yates's comment that it is an unstoppable train.

Dr McPherson: Yes, for sure.

Professor Yates: By the way, I agree with the ethical concerns. Don't get me wrong; I was just addressing innovation issues, not the ethical concerns.

Dr McPherson: I changed the subject.

Chair: We'll leave it there; it is a fascinating point to stop at. Thank you very much for your attendance this morning.