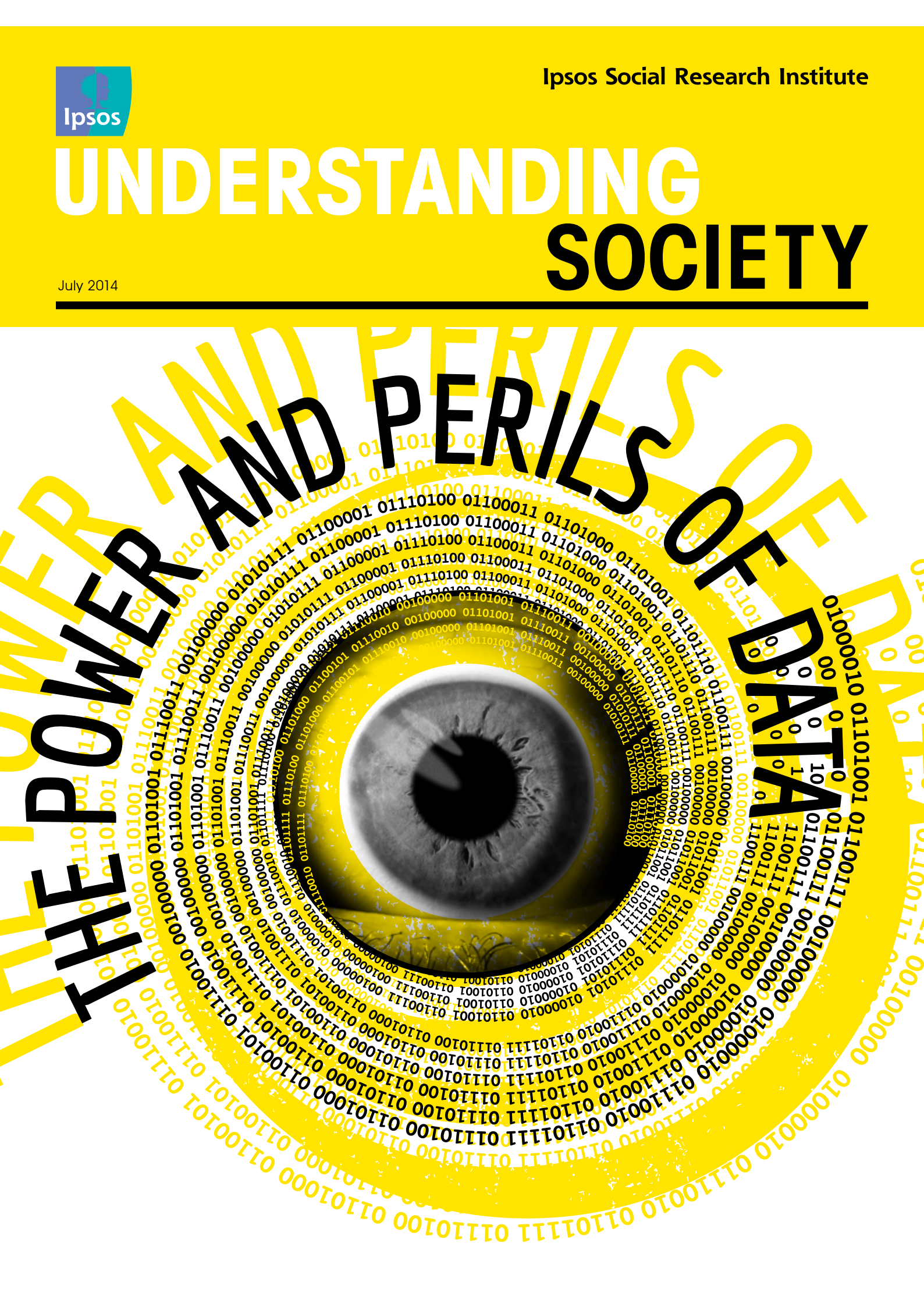
Ipsos Social Research Institute

UNDERSTANDING SOCIETY

July 2014

THE POWER AND PERILS OF DATA



The power and perils of data

Results from our Global Trends Survey,
a 20 country study of over 16,000 online adults

Is surveillance ever acceptable?

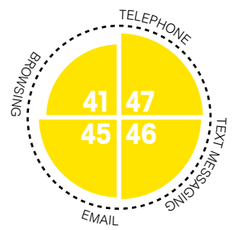
Between four in ten and five in ten find most forms of surveillance completely unacceptable. But there is a tipping point – only a quarter find it completely unacceptable to monitor other people's communications to combat terrorism.

Completely unacceptable [%]

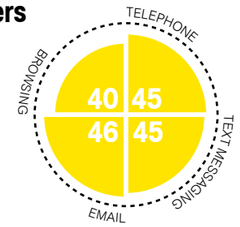
to combat crime

...if the government in COUNTRY was allowed to do the following things without their/your consent? Monitor...

your



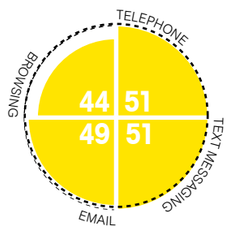
others



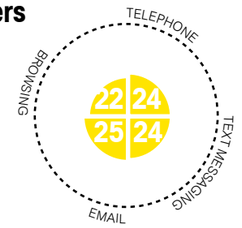
to deal with a terror threat

...if the government in COUNTRY was allowed to do the following things without their/your consent? Monitor...

your



others



Do people protect their privacy?

There is a disconnect between what people say and do. Nearly half are willing to pay for extra privacy – but less than a quarter have increased the privacy settings on their computer.

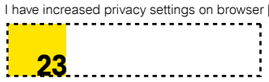
73% of those who say they are willing to pay extra to keep their details private haven't changed the security settings on their browser

Willing to pay

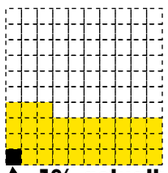
I am willing to pay extra for a service or product to keep my details private [%]

but don't change their settings

I have increased privacy settings on browser [%]



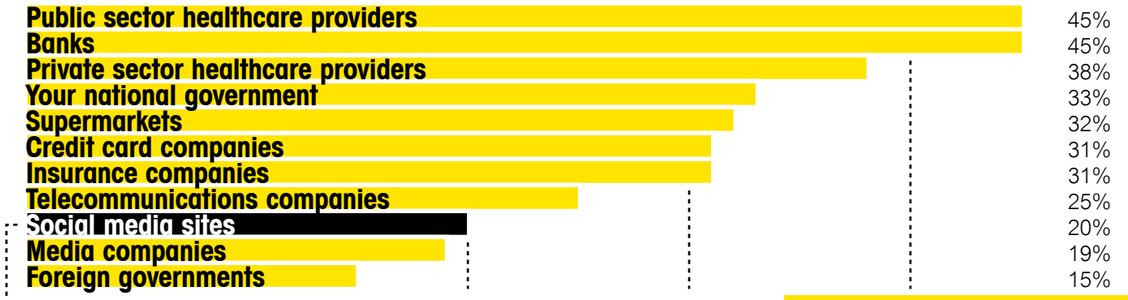
33% claim they usually read website terms and conditions



1% actually have
(according to server-side surveys)

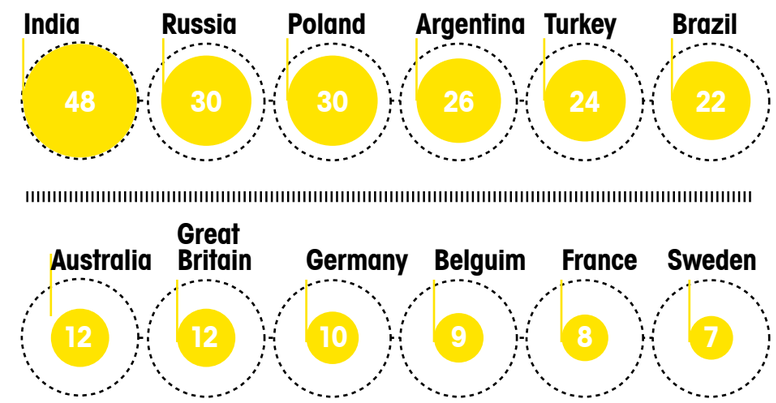
Who do we trust?

Trust in different institutions varies considerably – social media sites and media companies are only slightly more trusted than foreign governments.



To what extent, if at all, do you personally trust the following to use the information they have about you in the right way?

Social media tends to be trusted much less in developed countries



Knowing more doesn't increase trust

Trust social media with information
know what information is held on them

don't know what information is held on them

Trust government with information
know what information is held on them

don't know what information is held on them

Scan the QR code below to download an electronic version of this report.

Editorial:
Daniel Cameron
Bobby Duffy
Aalia Khan
Gideon Skinner

Information
www.ipsos-mori.com
aalia.khan@ipsos.com

Follow
www.twitter.com/IpsosMORI

Foreword

Welcome to this international edition of the Ipsos MORI Social Research Institute's *Understanding Society.* This issue – The Power and Perils of Data – focuses on the potential for data to improve the way people live their lives, as well as exploring public views on some of the risks this greater reliance on data brings.

The last decade or so has seen a shift in the way information is collected and analysed, bringing huge opportunities to transform how we organise society, business and government. Yet this raises questions about control and what can be done by those with access.

Unique datasets taken from Ipsos MORI's Global Trends Survey show that public anxiety about data sharing and transparency is high but varies across different types of people and is dependent on the type of data in question. Public concern about privacy has always existed, but has more recently been fuelled by international events such as Wikileaks, Edward Snowden's revelations of NSA surveillance and the discovery of big companies using customer details irresponsibly.

We are therefore delighted to have an interview with Alan Rusbridger, Editor-in-Chief of The Guardian, who elaborates on these stories and the implications for individual privacy in the future. Rusbridger says that the public are right to be anxious about data use and calls for better education about the types of data being collected through the technology we own.

Public and political discourse often considers the relationship between privacy and security as a trade-off. Governments and surveillance technologists argue that in order to protect individuals from various threats, be it financial fraud or terror, individuals are required to forgo some of their rights to privacy. Ipsos is working with an EU-funded consortium to re-examine this simplistic model and investigate whether public views can help make the case for developing ways of maintaining both privacy and security as new technologies are introduced.

We are also thrilled to have the contribution of Nesta's Chief Executive, Geoff Mulgan, in this edition. Mulgan, one of Britain's prominent thought leaders in innovation, examines the future of big data and privacy. He notes that whilst we are on the cusp of a big data revolution, an acceptance of 'zero privacy' and arguments of trade-offs hinder progress and innovation. A generation of digital natives will expect to shape and control their 'digital aura' and will be less forgiving of big companies and governments 'caught' using their personal details irresponsibly or spying on them. Mulgan investigates an emerging sector which enables users to have direct control of their personal details, as well as the rising movement of people-powered data.

Moving away from the perils of data, Clifford Young and Elisa Bernd from our Washington office give us an insight into its power and its role in the 2012 US presidential election. They argue that the 2012 electoral cycle marks the rise of the use of empirical models in forecasting and the forecaster-pundit. We also have an interview with Nate Silver, who is undoubtedly the best known of this new breed of political data analysts. These types of approaches are likely to become increasingly important elsewhere, at least in elections where large volumes of polling and other data are available for analysis.

Finally, we delve further into the concept of big data and its possibilities. While the term itself has been challenged for being too vague, Ipsos' Henri Wallard scrutinises some of the main components that characterise big data approaches. He reminds us to be mindful of the 'big' mistakes that big data can lead to.

We hope you enjoy reading about our ongoing research on the power and perils of data and the implications for public policy and society. Ipsos MORI remains committed to sharing the messages from our research, in the belief that a better understanding of public attitudes will lead to better policy and practice. If you would like to discuss any of the research here, please get in touch.

Bobby Duffy
Director
Ipsos Social Research Institute

@BobbyIpsosMORI

Data privacy: Where do people draw the line?

Bobby Duffy, London

Data privacy feels like a very modern problem. The media is full of stories about government surveillance and what happens to the data we all provide to private companies. But of course it's far from a new concern, and closely related discussions on threats to data privacy go back decades, at least since modern databases were developed.

Yet superlatives abound in current discussions. Various think tanks,[1] consultants and trade groups[2] have told us that the latent demand for privacy has never been greater, that privacy is the next key consumer rights issue[3], and that our personal data are more valuable than gold or oil.

Our surveys also tell us the greatest sin a company can commit in the eyes of the public is losing their personal data, followed by selling that data, even if it's anonymised – much worse than exploiting foreign workers or even charging more than competitors.

There is no one public opinion on data privacy

Our new 20 country study[4] released at an event with King's College London this year adds more international depth to this picture. It confirms some of what we knew, and in particular that there is no one public opinion on data privacy – in three separate ways.

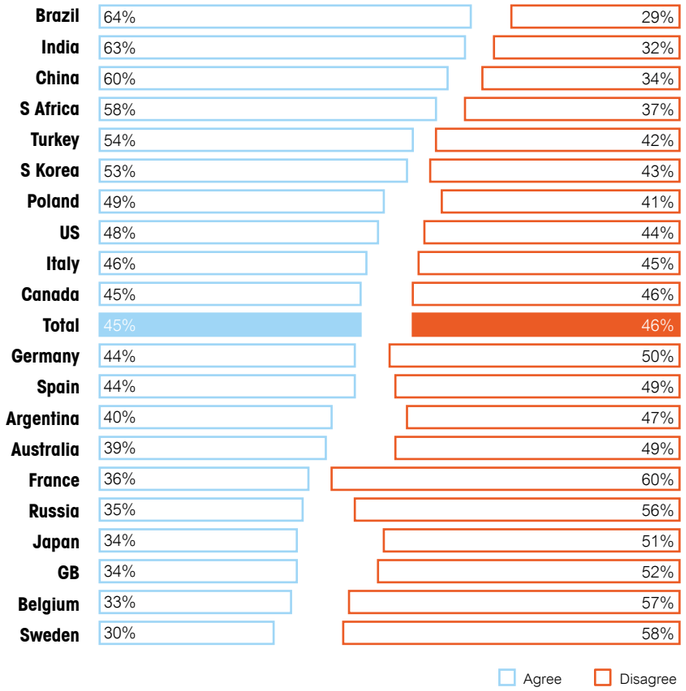
First, there is just variety in how concerned different people are. Indeed there is remarkable consistency in how populations segment across time and in different developed nations: we tend to find around 10% are 'privacy unconcerned', 60% are pragmatists, where concern depends on the circumstances, and 30% are 'privacy fundamentalists'.[5]

But, second, this gives a false sense of certainty in opinion. In practice, it's not just the pragmatists who change their views depending on the circumstances – experiments show that large proportions of the two groups at either end of the spectrum can be shifted, depending on what they are offered or how they are reassured.[6]

And, third, stated concern about data privacy and how people actually behave are barely nodding acquaintances. We can see the massive disconnect between what people say and how we know they act in this new survey. For example, nearly half of people across the 20 countries say they are willing to pay for increased levels of privacy for their data. But at same time, in the same survey, only a quarter of the same people say they have taken basic steps to increase the privacy settings on their browser. This means that three quarters of those who say they would pay for additional privacy haven't changed a simple setting on their computer.

Figure ONE.

Q: To what extent do you agree or disagree...
'I am willing to pay extra for a service or product to keep my details private'?



The latent demand for privacy has never been greater, privacy is the next key consumer rights issue, and our personal data more valuable than gold or oil.

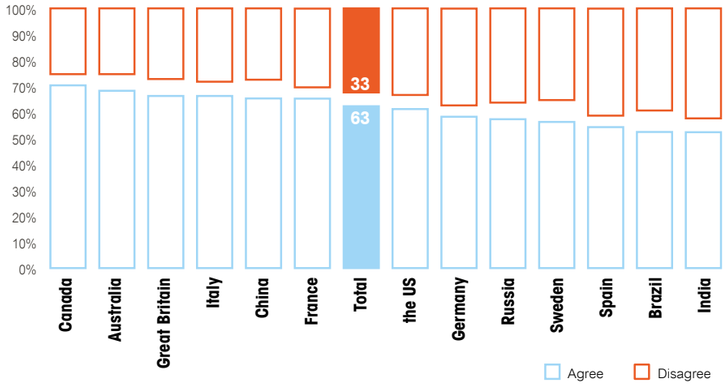
Do you really read terms and conditions?

We also asked the softest question possible on whether people read terms and conditions or user agreements on websites - giving people the best chance to come clean and admit they don't always. But still a third of people insisted they do always read them. This ranged from a quarter of (marginally more honest) Canadians to over four in ten in Spain, Brazil and India. The evidence suggests a lot of people are kidding themselves or us: server-side surveys show that barely 1% actually do read them. And this is no surprise, when some service agreements are over 30,000 words, longer than Hamlet.[7]

There are many examples of people being tested on how much attention they pay to what they're signing up for – and failing. The best of these is probably Gamestation's famous clause which they included in their terms on April 1st 2010 and which gave the computer game retailer the 'non-transferable option to claim now and forever more your immortal soul'. 88% of people signed up.

Figure TWO.

Q: To what extent do you agree or disagree...
'I often don't bother fully reading terms and conditions on a website before accepting them.'?



Different countries are at different stages on data privacy

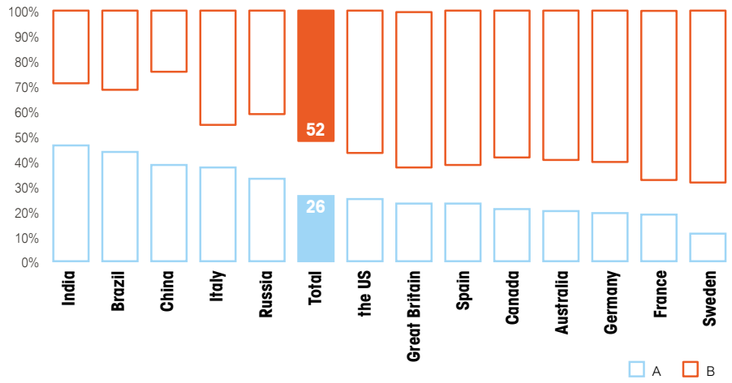
The international aspect to our new study also highlights the different stages the public are at around the world, and gives some clues to likely futures. In particular, people in more developed markets are more focused on privacy, at least in stated attitudes – and less likely to want to give it up for greater personalisation. This seems to be mostly a function of exposure to personalised services and recommendations, which people often find more irritating than helpful. People in the West are also more focused on anonymity: promising this does little to opinion in countries like Brazil, India and China, but it shifts views in the West.

The survey also shows the real problems that some industries have developed in a short space of time. Social media and media companies are only slightly more trusted with our data than foreign governments. And it's clear from this and previous work that revelations about the scale of security service surveillance does set some of the tone - but it's not as direct an effect on opinion of data privacy as we might expect from the significant media coverage. More important are the everyday interactions where people are surprised by how their online activities can be tracked so that recommendations follow them around the web. Indeed it's that sense of connection across spheres that unsettles people the most – the world and our data may be linked in ways that we couldn't imagine a few years ago, but that worries many of us, as we like to keep different aspects of our lives separate.

And this provides a hint of the tricky future ahead. In particular, more exposure to and understanding of what can be done with our data doesn't make us any more comfortable – in fact, for many of us it increases our concern.

Figure THREE.

Q: Which comes closest to your own opinion...
A - I am happy sharing information about online activities so that I get personalised services/relevant recommendations.
B - I would rather keep information and online activities private even if I do not get personalised services and relevant recommendations.



How should government and businesses respond?

Both business and government have decisions to make. How open should they be about what happens to our data, and how far do they chase the benefits of knowing more about us?

It seems blindingly obvious that honesty will be the best policy. Many studies (including our own) have shown transparency is key to trust, and trust is related to all sorts of other good outcomes. But things are not so straightforward in this case.

More exposure to and understanding of what can be done with our data doesn't make us any more comfortable – in fact, for many of us it increases our concern.

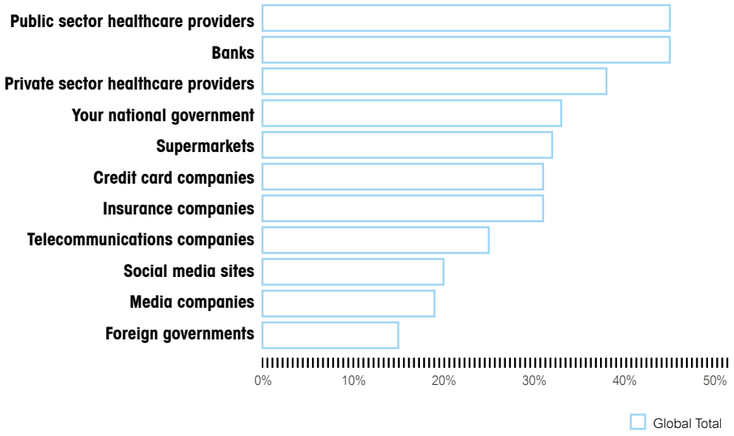
First, we shouldn't kid ourselves that openness will automatically lead to trust. Technological capabilities have raced so far ahead of public knowledge that the implications of greater transparency are unpredictable – many will get more worried as they find out more.

And while there is great concern, there is still little action by most individuals to protect their privacy or control their data. There are therefore weak incentives for private companies and governments to really push the issue. Being a first mover, highlighting concerns that people don't really have at the front of their mind and encouraging them to restrict access to their valuable data, is a dubious advantage.

Of course, while business and governments have similar incentives, they have very different interests – and a key dynamic will be the extent to which one shifts attention to the other. We're already seeing how legislative bodies are acting to protect citizens by enforcing people's 'right to be forgotten' online. Internet and other companies are at the same time championing their role in protecting our data from government intrusion.

Figure FOUR.

Q: To what extent, if at all, do you personally trust the following to use the information they have about you in the right way?



But as we have seen there is also such variety in views across situations, individuals and countries that there is no one 'right' response - it needs careful tailoring.

Companies need to be clear about where on the spectrum they stand, but be flexible in how they act and communicate – and careful not to be too clever.

For example, there is talk in the academic and think tank worlds about how the combination of more sophisticated analysis and insights from behavioural economics could allow us to 'personalise privacy', defaulting people to what we think they'd prefer. But many will find it particularly creepy that decisions on privacy (of all subjects) can be made for us by connecting our data and predicting what we want.

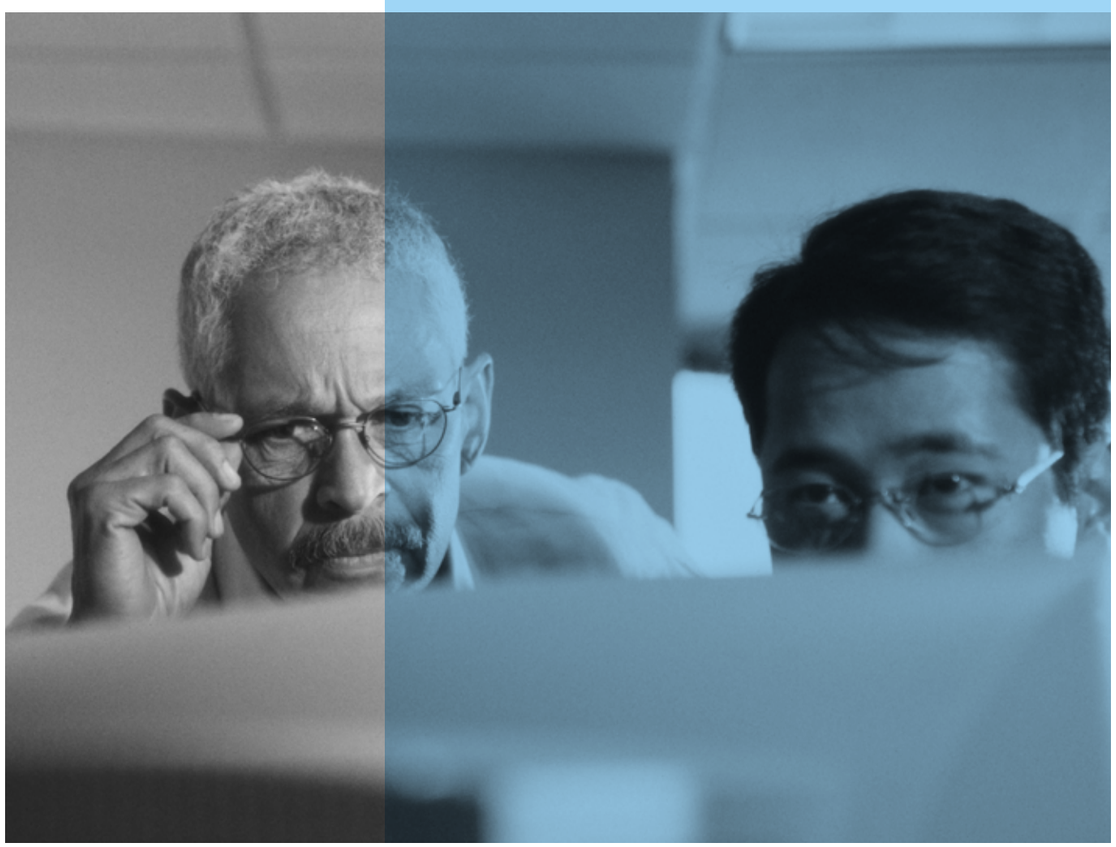
There seem to be two factors that will unsettle this uneasy balance – events and legislation, which will be closely linked. It is difficult to see many events gaining sufficient public attention to topple governments or companies – but as we've already seen, they are leading to regulatory change.

But getting that legislative balance right is going to be tough.

While there is great concern, there is still little action by most individuals to protect their privacy or control their data.

In the end, as Geoff Mulgan reminds us later in this edition, what we're trading our privacy against is not so much personalised services or recommendations as all the wonders of an internet provided (largely) for free: 'Surveillance is the business model of the internet' as one technology commentator describes it. Governments and legislators need to be mindful that where the public draw the line will depend on what they are asked to give up.

Companies need to be clear about where on the spectrum they stand, but be flexible in how they act and communicate – and careful not to be too clever.



No place to hide

An Interview with The Guardian's Editor-in-Chief, Alan Rusbridger

Ben Page, London

The Guardian and the Washington Post were recently awarded the Pulitzer Prize for Public Service for their coverage of the National Security Agency's surveillance activities based on the leaks of Edward Snowden.[8] The Pulitzer committee praised the Guardian for its 'revelation of widespread secret surveillance by the National Security Agency, helping through aggressive reporting to spark a debate about the relationship between the government and the public over issues of security and privacy'.[9]

BP: In a relatively young digital age, how do we marry the instinct to both protect our data and embrace the advantages of using it? Is the concern for privacy a growing trend or a temporary blip?

Oh no, I think this is going to be one of the big issues of the 21st century, simply because it is so self-evidently problematic. That doesn't mean that, in the end, people will not settle for the benefits over the disadvantages and make the compromises when necessary. But I think that the potential for these things to be quite toxic is always there.

I suppose I think of myself on the utopian side of the argument - I'm sort of like the Clay Shirkys of this world instinctively, more than the Morazovs. Clay Shirky is the guy who wrote the sunny books about the internet[10], and Morazov wrote the gloomy books about the dangers and the threats[11]. I've always thought of myself on the sunnier side of the argument, mostly because I like technology and I like the things it has enabled us to do in the news business.

Even so, I think the problems of connected data and the unthinking way in which governments and commercial organisations are stumbling into this, is potentially quite problematic.

BP: So do you want to see more regulation?

Well let's start with transparency and think about regulation down the line. I think there has to be education for a start, because I think most people are quite shockingly ignorant about it all.

BP: Our research shows that people say they are anxious about data sharing, but they also admit mostly that they don't actually read any of the terms and conditions they tick on websites. There are great examples of terms and conditions where, on page 33, there is a phrase, 'Claim $1,000 if you have read this far'. And I think it takes 3,000 people to download it before the first person rings up to get the $1,000.

That's interesting. The producers of the play Privacy at the Donmar Warehouse worked out that the Apple terms and conditions were the same length as The Tempest. The actor Simon Russell Beale read The Tempest alongside Eric Walter reading the Apple terms and conditions.

I think there needs to be an education all round for politicians, business leaders and consumers - we've all got to get smarter.

BP: So are we right to be anxious about privacy, transparency and data sharing?

I think we're right to be anxious. The Snowden story drove home to me the potential menace of how these technologies can be used and their power is considerable. We're all relying on oversight mechanisms to say - well, that's the thing that stops this from being bad, is that firmly in control? I don't think it is firmly in control.

BP: People are particularly anxious about specific types of data. For example, there is more public anxiety around institutions like the NHS using their medical data to save time and public money, than there is about commercial businesses like Amazon knowing what they buy. Why do you think this is?

Well, it's the same picture really. I can see the benefits of sharing that type of data, but I think there is a lot of reassuring yet to do. People need to be reassured that they are in control of the security of these systems, and that so-called metadata - information which is supposedly anonymous - can't actually be traced.

When we live in a world where the NSA, the Department of Justice in America and the military can't prevent massive leaks, what are we to make of something like the NHS spine where two million people can access our data? Until we can build safe systems, people are bound to feel anxious about use of their medical records.

BP: Many people will tell you that virtually anything can be hacked or accessed. Do you think it is possible to build safe systems?

Probably not. So I'm quite interested in the people who are coming up with ways in which we are much more in control of our personal data. I would be more comfortable, for example, if I had my medical records in my own wallet which I could personally take to the hospital or my doctor, rather than a system which controlled it.

There are people who are working on these decentralised systems of data, where it can be centralised, but in a way that I feel in control of it. I don't know enough about the technology to know where that experiment leads, but what's more worrying is governments and commercial organisations that believe they can blunder in without seeking any form of consent or attempts of any kind at transparency. That is a world where you are just storing up trouble.

But I think after Snowden, there will be quite serious industries being built around peer to peer encryption and gauging meaningful consent. Also, the notion of what is meant by meaningful consent is interesting – how much of it is consent as opposed to how quickly can I get this box to disappear off my screen? This is something that businesses and governments will have to think about.

BP: Which country has the best approach to privacy and security, and which has the worst?

People need to be reassured that they are in control of the security of these systems, and that so-called metadata - information which is supposedly anonymous - can't actually be traced.

I think the countries which have not been blessed with the kind of stability that we have in the UK are generally more attuned. So when I meet my Spanish colleagues, who were not living under a democracy until 1975, they are more highly attuned. America had Nixon, Hoover, McCarthy, in living memory, and have a more rights-based society, and so they are more highly attuned. And I think sometimes in Britain we sort of pat ourselves on the back for Magna Carta, and we don't think this really affects us. But when you sit down and explain what's going on, I think people then are worried and do understand the issue.

BP: Thinking a bit more about the opportunities, how has new technology changed journalism's relationship with its audience, and how much of a focus is there on personalised news?

We're experimenting all the time. In 1994, the Massachusetts Institute of Technology designed a virtual daily newspaper, customised for an individual's news preferences, called the Daily Me. In fact, we tried printing off a few Daily Mes. And actually last week somebody came in with something called Paper Later, which is a personalised newspaper, so it's sort of taken 20 years to get to what we were imagining in the early '90s. We could completely tailor-make anything for anybody, depending on their preferences. At the moment there's a limited take up, because I think most people's idea of a newspaper is actually something that challenges and broadens their horizons.

BP: Reading a paper online or on an iPad, as opposed to reading in print, means that you can miss stories, depending on your filtering system. You are also immediately drawn to some stories through availability or attention bias than others.

Yes, but we don't know that this will always be the case. I think we're still in the very early days of what location data means, and whether people will want to use location services to limit their news or their entertainment and sporting choices. So we will continue to experiment, but we don't know the degree to which personalisation will be really important.

BP: The Guardian has been at the forefront of citizen and digital journalism, embracing online and data blogs, leaving other national papers to catch up. How do you tread the line between engaging your readers as citizen journalists, whilst meeting the standards of professional journalism?

We call it open journalism, because I think citizen journalism has a rather sort of clunky feel to it. But you can say in the world that we live in now, everybody is potentially a publisher of information.

I use the analogy of Michael Billington, our theatre critic. Being the professional expert, we send him into the stalls every night. A typical night at the theatre will have 900 people there with him. Is it conceivable to say that out of all 900 people, none of them have anything interesting to say? The answer has to be no, of course. So then the question is, how do we filter the interesting ones from the

We don't know the degree to which personalisation will be really important.

uninteresting ones? Even if you only get 10 interesting opinions of the play out of 900, you are bound to have a more complete account. So, if it works for theatre, would it work for sport, science, war reporting and investigations?

BP: And does it work?

Yes, it always works. We are trying to build a model where we acknowledge that there are some things that we as journalists do best, and if we are really smart about harnessing what other people are doing, we could build something that was even better. And that is how we knit a newspaper into the ebb and flow of information today, including the distribution and sharing of it.

If you look at our science or environment coverage, you can see we have a whole network of sister bloggers, commentators and articles. What we are doing is handpicking the best of what's out there, some of which is good, if not better than what we can do. When you look at Twitter, you see that it will always be faster and more comprehensive. But newspapers can amplify what is interesting. Suddenly, you have a different idea of what a newspaper is.

BP: What's the next big thing after Snowden for you?

The aftermath of Snowden, as I've said, is that we understand that we need some of this stuff to keep us safe. The question is, how do we keep us safe from technologies that could be totalitarian? I don't think that anyone believes that the system of oversight that we currently have is up to the task, or that Parliament has ever debated it properly. We still don't have anything that you can reasonably call consent when it comes to sharing information. I think if you had all those conditions in place, then yes, you could move on from this subject and write about something else. But I think there is still some mileage to go.

BP: And of course, if you had a brilliant browser that allowed you to click here and nobody could ever see what you were looking at, you can guarantee it would get plenty of take up.

This word, metadata, is the word of 2014, I think. And whenever I hear politicians like Malcolm Rifkind or William Hague on the radio, talking about Snowden - politicians whose job it is to reassure - go straight to 'nobody's reading your emails, we need a warrant to look at this, we're not interested', and so on. And they fail to mention the information that they don't need warrants for, which is metadata.

There is a moment in the Privacy play where everybody is asked to go through their mobile phones. The actor takes everyone who has an iPhone through the settings where you can see everywhere that your phone has tracked you for the last three months. And you hear a gasp in the audience. I've done it with my friends – you can take them through it and tell them 'so you were in that house in Tottenham on that night at 8 o'clock'. When people realise how much information on their lives is being stored, they suddenly feel that it's creepy and then the penny drops. You don't need a warrant to access that information and nobody has any privacy.

I don't think that anyone believes that the system of oversight that we currently have is up to the task.

And just like phone hacking trials illustrated an explosion of bad practices in the newspaper industries, this will happen in other walks of life too. Or there will be accidental leaks. Either way, that's why I think it is an issue that is not going to go away.

Brilliant, thanks so much for your time.

Privacy or security: a false choice?

European citizens' perceptions of privacy, personal data, surveillance and security


Carolina Haita, Bucharest


Daniel Cameron, London

The relationship between privacy and security is often described as a trade-off. The argument goes something like this: securing against the many threats individuals and societies face requires people to give up at least some of their personal privacy. Many of the security and surveillance technologies used by governments and others rely on this reasoning to justify the ways they reduce individual privacy.[12] There are certainly different views about the extent to which privacy should be traded-off, but this basic model is rarely questioned in the course of public and political debate.

Ipsos MORI is working with a consortium of partner organisations on an EU-funded project to investigate this traditional privacy-security trade-off model.[13] The aim is to re-examine privacy and security in order to explore whether there could be alternative ways to characterise the relationship. Is it possible to enhance security without an inevitable reduction in privacy? Can new technologies be designed to strengthen both privacy and security?

As part of the project, Ipsos MORI conducted discussion groups in eight European countries during summer 2013.[14] The discussions sought people's views about specific security technologies in contexts where they are or might be used, alongside more general perceptions of privacy and security in today's world. The findings were used to help design a large-scale representative quantitative survey currently being carried out in EU Member States.

This article summarises some of the key findings from the qualitative research. These give a snapshot of European citizens' perceptions of privacy and security.

The importance and limits of privacy

Participants were clear that privacy is essential to them personally. This instinctive sense that privacy should be considered a basic human right was near-universal. Yet participants' understanding and expectations of what privacy looks like in practice were much more nuanced. This was often informed by what they thought was realistic rather than by any ideological commitment to a specific understanding of privacy.

For example, participants did not think it was possible to have privacy in many circumstances where they considered it desirable. They were often willing to accept considerable intrusions into their privacy if they thought this would strengthen local, national and international security, and if they could see limited drawbacks for them personally.

> I think it's scary that they know when and where we go, but I can understand why.

Female, under 40, Belgium

Protecting personal data was the usual starting point for thinking about privacy. How personal data is collected, stored and used was the most current and contentious privacy issue. Participants discussed many different examples of the sorts of personal data they knew was collected and stored about them, from financial data to medical records, from location data to pictures of their most recent holiday on social media. They felt that they should have the ability to control their personal data, as it was theirs to share because of the benefits it brought them.

> I only give away the data that is really necessary, especially online.

Female, under 40, Germany

Yet participants also believed they had very limited control over the personal data held by organisations, including both government and private companies. Often they felt they had no choice but to provide personal information in order to access services. They therefore saw a strong role for the state in developing appropriate regulatory frameworks for protecting personal data, and for policing and enforcing agreed standards.

> What we do doesn't matter… our personal data is stored in many places… when you are online all your data can be available.

Male, under 40, Hungary

Government more trustworthy, businesses more competent?

How much participants trusted government and private companies varied considerably. In general, there was more trust in government's intentions but less trust in their competence to protect personal data. The opposite was often the case for businesses. There were concerns about the reasons behind private companies' frequent requests for personal data. However, if participants could see clear benefits for them, for example in terms of improved service, many were comfortable providing data if the right reassurances were in place.

Levels of trust in government when it comes to privacy were different in each country. This tended to be rooted in broader attitudes to the state: where government was viewed as largely benign there were fewer concerns about its role in storing and using personal data. In countries with more recent experience of oppressive regimes, participants were much more guarded about trusting the state to respect privacy and use information appropriately.

Overall, some types of privacy were seen as untouchable – sensitive data like medical records, individuals' political and religious views, their personal communications, and particularly what happens in people's homes. Any intrusion in these areas without very good reasons and strict legal controls was seen as completely unacceptable.

> I don't want to be monitored at home, they can do that all they want in other places, but not at home.
>
> Female, under 40, Denmark

> My family and my home are private, and also what I tell my best friend via mail. I want to keep that.
>
> Female, over 40, Germany

There was an assumption that privacy, at least in the strictest sense, is very often difficult to maintain if you want to be part of a modern society.

But there was also an assumption that privacy, at least in the strictest sense, is very often difficult to maintain if you want to be part of a modern society. In particular, being able to move around and meet with other people in public without being watched was not thought to be as important as some of the other types of privacy discussed. Many thought freedom of movement and association were now unrealistic ideals and, more than that, relatively unimportant provided appropriate safeguards were in place. They assumed this sort of information would only be of interest if people were doing something wrong.

> I think privacy in public spaces is not possible anymore. There are cameras everywhere.
>
> Male, over 40, Germany

> I don't know what the problem is – if you follow the rules there's nothing to worry about. If not you deserve to be caught.
>
> Female, under 40, UK

This reflects the clear concerns participants had about protecting security. Threats from crime and terrorism were discussed frequently, as were the responses people thought were required. The need to protect security was an important priority – and for some, more important than protecting their privacy.

> I favour the use of video surveillance in the cities because of safety concerns.
>
> Female, over 40, Portugal

> We rely on government – the police and the army and so on – for our security, ultimately. There are things you can do but we need outside help.
>
> Male, over 40, UK

Attitudes to both privacy and security were broadly shared across countries, but there were differences in the detail. For example, in some countries like Hungary and the UK there were few concerns about internet surveillance, while in others like Germany and Denmark participants had considerable reservations about these technologies becoming more embedded without proper controls.

Context matters

Participants also discussed specific security technologies using a number of different scenarios. Examples included airport scanners, smart meters and fingerprint recognition in schools. Discussions around these scenarios highlighted how context-dependent attitudes to privacy and security are.

One scenario that generated good discussions focused on internet monitoring. The scenario described a Muslim student who researches and discusses terrorism on the internet as part of his studies. His parents become concerned about internet monitoring and ask him to stop his research. Participants were able to see the tensions inherent in protecting both privacy and security in this context.

For most, monitoring this type of behaviour was seen as legitimate – or indeed crucially important – provided it is done by government in a controlled way. But most thought that researching terrorism was legitimate too, even if they thought this should be done in a transparent way that is open to questions from government. Finally, participants also understood parental concern, with some saying they would want to dissuade their own children in similar circumstances.

> I agree with the parents… I would do the same thing.
>
> Male, over 40, Portugal

Some of the scenarios were more clear-cut. One described government asking citizens to record their religion on an identity card, in order to help ensure people from different religious backgrounds were involved in local decision-making.

Participants raised a number of objections to this idea, pointing out that religion is a private matter. They were concerned about linking information about religion to other data on an identity card; risks around the security of government databases; and possible future misuse of the data by radical organisations. More than that, participants could see no obvious benefits, arguing that there were alternative ways to consult local religious groups. Even with reassurances, this scenario was considered completely unacceptable in almost all countries.

The need to protect security was an important priority – and for some, more important than protecting their privacy.

Participants wanted a robust, detailed case to be made for new technologies, including whether they are cost effective and clarity of why any loss of privacy is necessary.

Reinforcement, resignation and resistance

There were three common underlying themes when it came to participants' views on the relationship between privacy and security, reflecting the pragmatism evident in our Global Trends Survey data described earlier in this edition. While participants responded in different ways depending on the specifics, most tended towards one of these groups:

- Reinforcement – many were content with giving up further privacy if it made them feel more secure. They said they had nothing to hide from authorities and would welcome further surveillance if this helped prevent crime and terrorism.

- Resignation – this group were instinctively uncomfortable about how much their privacy is currently compromised, and this was exacerbated because they expected the situation would continue getting worse. But they could see few options for them to claw back privacy, and seemed willing to live with the situation because it made little difference to their daily lives, and indeed brought some benefits.

- Resistance – there were some participants who wanted privacy to be strengthened. This was usually because they had a principled objection to intrusion into their privacy. They felt things had gone too far in terms of surveillance and security technologies, and did not accept that these really improved security in any case.

What next?

Participants did not agree on the best way forward, but there were some shared concerns. They highlighted several issues they thought should be taken into account when introducing new surveillance and security technologies.

They wanted better protection of personal data by controlling access to data within organisations, and by strongly limiting data transfer to third parties. Many felt that individuals should be treated better too, so that people are respected, have choice and control, and are not viewed with a general suspicion as a result of new technologies. Finally, participants wanted a robust, detailed case to be made for new technologies, including whether they are cost effective and clarity on why any loss of privacy is necessary. Taking these concerns on board will be important for developing new technologies that protect both privacy and security in a way that is consistent with public priorities.

Is people-powered data possible?

The future of citizen control

Geoff Mulgan,
Chief Executive
of Nesta

So far the public seem to have accepted a dramatic increase in use of their personal data because it doesn't impinge much on our freedom, and helps to give us a largely free internet. That's because the public generally take a pragmatic view of privacy: it's not and never should be something absolute. Instead we're willing to make trade-offs – if using a loyalty card gives a supermarket information about what we eat, that doesn't matter much so long as we get something in return and there's no obvious downside. On the other hand, we're far more sensitive about data relating to our physical health, or even more our mental health, because we can see how easily that information could be used against our interests.

But the ways in which we make these trade-offs could be changing. Edward Snowden's NSA revelations have fuelled a growing perception that the big social media firms are cavalier with personal data (a perception not helped by Facebook and Google's recent moves to make tracking cookies less visible) and the Information Commissioner in the UK has described the data protection breaches of many internet firms, banks and others as 'horrifying'.

Ipsos MORI's research suggests that the more people are familiar with social media the more cautious they are about what's done to their data, and of course over time an ever larger proportion of the world's population is becoming more digitally aware.

According to some this doesn't matter. Scott McNealy of Sun Microsystems famously dismissed the problem: 'you have zero privacy anyway. Get over it.' Mark Zuckerberg claims that young people no longer worry about making their lives transparent. We're willing to be digital chattels so long as it doesn't do us any visible harm.

That's the picture now. But the past isn't always a good guide to the future. More digitally savvy young people put a high premium on autonomy and control, and don't like being the dupes of big organisations. We increasingly live with a digital aura alongside our physical identity – a mix of trails, data, pictures. We will increasingly want to shape and control that aura, and will pay a price if we don't.

That's why the movement for citizen control over data has gathered momentum. It has been 30 years since Germany enshrined 'informational self-determination' in the constitution and other countries are considering similar rules. Brazil recently passed the world's first Digital Rights legislation, partly inspired by Sir Tim Berners Lee who thinks we need a new Magna Carta fit for a digital age.

Organisations like Mydex[15] and Qiy[16] now give users direct control over a store of their personal data, part of an emerging sector of Personal Data Stores, Privacy Dashboards and even 'Life Management Platforms'. In the UK, the government-backed Midata[17] programme is encouraging firms to migrate data back to public control, while the US has introduced green, yellow and blue buttons to simplify the option of taking back your data (in energy, education and the Veterans Administration respectively). Meanwhile a parallel movement encourages people to monetise their own data – so that, for example, Tesco or Experian would have to pay for the privilege of making money

out of analysing your purchases and behaviours.

Few of us have a very realistic understanding of what actually happens to our personal details. But when people are shown what really happens to their data now they are shocked. That's why we may be near a tipping point. A few more scandals could blow away any remaining complacency about the near future world of ubiquitous facial recognition software, Google Glass and the like, a world where more people are likely to spy on their neighbours, lovers and colleagues.

Geopolitics will also play its part. Europe is much more concerned about identity and privacy than other parts of the world, and that concern has been fuelled by revelations of NSA's far reaching surveillance. Europe will set much more stringent rules for protecting personal privacy and given that the EU now has 500m of the world's richest citizens it will be hard for social media firms to ignore these standards.

Meanwhile some firms are distancing themselves from the pack – like Microsoft, developing technologies to help citizens control unauthorised reuse of their data. The big consultancies and accounting firms see privacy accreditation as a huge new source of business. And The Boston Consulting Group recently warned that two-thirds of the total value of greater use of personal data, estimated to reach a potential €1trn in Europe by 2020, will be lost if organisations fail to establish trust.

The next few years will bring a further explosion of data, and of data awareness in daily life. In the end we may be able to choose more easily just how open our lives are to be – and some of us will be willing to pay more to stay more private. It's possible too that a new generation of technologies will be developed that design in privacy and personal control over data, just as a previous generation of technologies, dominated by the military, security services and big business, did the opposite.

We're some way off the new Magna Carta that will at some point need to establish the ground rules of privacy, power and identity in a digital world. But these issues are fast moving from the margins to the mainstream of daily life – and some very powerful organisations risk being on the wrong side of history.[18]

Few of us have a very realistic understanding of what actually happens to our personal details. But when people are shown what really happens to their data now they are shocked.

The trouble with numbers

Ben Page interviews Nate Silver

BP: Survey research is mostly about asking people questions and adding the answers up, and then trying to interpret what they say. Increasingly, it's moving into passive measurement as well. You have become one of the aggregators-in-chief of surveys. What do you find most interesting these days? Are you going to try and forecast the US job market?

NS: Probably not! The evidence suggests that economic forecasts are not very good, and have not been getting any better.

BP: The meteorologists claim to be better than the economists.

NS: They are; the meteorologists are kind of the heroes of my book The Signal and the Noise. There are a few things that help: one is that we have a good physical understanding of dynamics of the weather system, so they're actually building a simulation of the atmosphere – it's not really statistical. It's been a success story where hurricanes now are forecast about 250% more accurately than 25 years ago. Hurricane Sandy was predicted almost exactly five days in advance. The sports gamblers are also very good. Part of it is that if you're putting money on the line, it has a disciplining effect. A lot of people in American politics in particular are just pontificating – kind of like in the essay, you know 'On Bullshit' – where it's not that you're lying, so much as you're entirely indifferent towards whether you're accurate or not.

BP: I read you as somebody who is big on Bayesian statistics. Tell us about Mr Bayes.

NS: So Thomas Bayes was an English reverend and kind of a pioneer in the theory of statistics, and – this was 250 years ago – his basic insight is something called Bayes' Theorem, which is simple algebraic equation. It says, first of all, you have a probabilistic view on the world – and it's by the way not a metaphysical statement, it's not saying the world is uncertain, it's just saying our knowledge of the world is uncertain. The other thing is that it assumes you start out with a prior belief – a presumption of a bias – and you weigh new evidence against that.

BP: I guess there is still an element of subjectivity or judgement involved in the things you're doing?

NS: I tend to see subjectivity as intrinsic to the way human beings view the world: we all have one narrow lens through which we perceive the shared 'objective' reality and making predictions is a way to verify whether our subjective view matches this reality, which is hard to do.

The evidence suggests that economic forecasts are not very good, and have not been getting any better.

It's about understanding when you are really in a data-rich environment, because it's not just the number of observations you have.

BP: I'm interested in just how good both ordinary people and decision-makers in business and in government are at dealing with probabilities.

NS: Yeah, people don't have an intuitive grasp of that at all. They think if you say something is 80% likely to occur, it's tantamount to saying it's guaranteed to occur, when obviously it's not. If you woke up every day and you had an 80% probability of not being stabbed, you wouldn't survive very long. Those 20% outcomes occur quite a bit. If you follow sports as I do, or especially playing poker, where you're on the losing side of that 20% enough, your intuition for probability is better. There were a couple of years where poker was my main source of income.

BP: You've stopped this now? You've found the blog is better, the writing is better?

NS: Well, the poker's worse! In 2003 you had a kind of fat accountant slob named Chris Moneymaker, this everyman character – nice guy – who won the World Series of Poker in Las Vegas. He's an OK player, but basically got very lucky. And ESPN would edit out all the hands that he would lose or play badly. That deceived people as to how easy it was. So for a while, there was a lot of dumb money in the poker economy – but now it's gotten very, very competitive.

BP: People talk a lot about Big Data. To me, some of that is like that guy from Google who said, 'Social media is like teen sex: lots of people talk about it, lots of people look forward to it but when its finally achieved, most people are disappointed.'

NS: If you read the Harvard Business Review, every other ad is about big data. They're pitching this as the magic serum: that's always dangerous. The problem with some of this data is that it's big, but it's not structured. It's about understanding when you are really in a data-rich environment, because it's not just the number of observations you have. The credit rating agencies in advance of the crunch had millions of observations on individual mortgages, but all from a period when housing prices were increasing. Part of the issue with social media metrics is that you're collecting data that's only a couple of years old. But Google's product, Flu Trends, searches in real time.

BP: This is people typing 'flu' and they can predict actual influenza cases.

NS: They say they can, but they considerably overestimated the flu in the United States last year. My hunch is, you build a model that can accurately describe what search terms people were using in, say, 2008 to 2010, but people change their search habits. Now that you have Google auto-complete, that will suggest different queries to people than you might have had, and when people are typing on mobile phones you tend to have shorter queries, so that will change things. Three years is a lifetime in social media.

BP: We know people aren't rational: Nudge by Thaler & Sustein is very popular here, as it is in US policy circles, but what's your advice to the average human being? Most people's financial planning, for example, seems to involve them dying at about 70, when actually they should know it's very unlikely to happen.

NS: People apply a lot of attention and a lot of bandwidth to where they're going to go for dinner – and I'm a foodie, so I don't begrudge that entirely – but those heuristics don't work as well for things like career or education planning, where you have to be more detached. One thing we're seeing now is that once you're unemployed for more than about six months, it becomes very difficult to find a job again. Employers begin to assume, 'A lot of other employers have passed on this person, there must be something wrong with them'.

BP: So the advice is if you become unemployed, take any job, and don't wait for the right job? OK. Any other examples for people?

NS: There was a column I did in the Sun – slightly embarrassing! But there is a dating site called OKCupid that analyses a lot of data, and a couple of years ago, we looked at which is the best night of the week to go out to meet someone. They looked at what percentage of people had updated their status to say 'I wanna get laid', basically, and the percentage peaks on Wednesday night. We think the issue is that early in the week, people go out to drink; then on Friday and Saturday, everyone goes out with their friends. It's kind of amateur hour. The headline of the Sun article was 'Maths can be fun!'

People think if you say something is 80% likely to occur, it's tantamount to saying it's guaranteed to occur, when obviously it's not.

This is an edited version of an article originally published in the Ipsos MORI Almanac 2013.

The rise of the forecaster-pundit

Big data, data aggregation and the 2012 US Presidential Election


Clifford Young, Washington


Elisa Bernd, Washington

The 2012 US presidential election was a watershed for a number of reasons. Our first African-American president—Barack Obama—successfully defended his presidency against Mitt Romney, the Republican challenger. For most, Obama's victory was no surprise. He did, though, outperform both the polls and the electoral forecasts by a fair margin, beating Romney by 3.9 percentage points (51.1 to 47.2), when most put his probable victory margin at between 1 and 2 points.

But the 2012 electoral cycle also marked the rise of big data and advanced analytics in politics. One such example is the Obama campaign's use of advanced analytics and large datasets to push their base to the polls on election-day. Such data allowed the Obama campaign, through statistical propensity models, to identify likely Obama voters who without a nudge would not have shown up on election-day. Many analysts attribute Obama's strong electoral showing to his campaign's analytics advantage over Romney and the Republicans.[19]

Another example of 2012's big data DNA is the sheer volume of polls. The advent of new methodologies like online polling and robo-polls has significantly reduced cost barriers when compared with the face-to-face and telephone polls of previous decades.

By our count, 17,058 polls were conducted at all levels of government (local, state and national) in 2012. This number probably undercounts the total number of polls by a factor of two to three as it excludes private political polls for campaigns.

Front and center in this poll volume revolution were websites, like RealClearPolitics[20] and HuffPo's[21] Pollster, that aggregated the polls in almost real-time using 'web scraping' and other automated curation techniques. These 'aggregators' became essential in serving as a virtual meeting place both for political junkies and by making poll data easily assessable to the public.

By our count, 17,058 political polls were conducted at all levels of government (local, state and national) in the US in 2012

The exponential increase in the number of polls also made statistical analysis easier and more robust. No longer was the single poll of much inferential interest. This innovation, in turn, led to another equally new and unique phenomenon in elections: the 'forecaster-pundit'.

Statistical election forecasting found its way into the national lexicon and the media's election coverage in 2012. Nate Sliver—the best known forecaster of the bunch—together with others, like Sam Wang, Drew Linzer, Alan Abramowitz, and Simon Jackman, became important stakeholders in the national narrative and ultimately predicted quite accurately the final election outcome (see Figure 5). Mis-predictions by well-established polling firms and the Romney campaign led many to talk about the rise of the 'forecaster-pundit' and the imminent demise of the traditional political pundit and pollster.[22]

What did the forecasters do?

In broad strokes, the electoral forecast models employed in 2012 depended on data aggregation. Some aggregated individual polls—sometimes hundreds or even thousands. Others aggregated past US presidential elections; while still others combined both poll and election aggregation. However, all models used aggregated data rather than a single poll or other 'special sauce' heuristics.

At their core, all these models maximize the amount of information employed in order to reduce uncertainty. They also go beyond forecasting of vote share by estimating the probability of victory for a given candidate (e.g., Obama has an XX% chance of winning).[23] This is a radical departure from the 'deterministic' nature of single poll analysis where outcomes are thought of as binary (e.g., Obama will win or will lose). This innovation, however, only becomes possible in a world of big data where the frequency of an event can be more easily ascertained (e.g., how many polls show Obama ahead of Romney).

The first family of models—poll aggregation or model-based averaging approaches (see Figure 6) – basically do two things: (1) they combine multiple

Figure FIVE.

Forecasters' final predictions of the 2012 US presidential election result

	Final Prediction Vote Share	Probability of Obama Victory
Actual	3.9	-
Linzer	3.9	99%
Jackman	1.6	91%
Wang	2.5	99%
Silver	2.2	91%
Abramowitz	3.8	80%

Figure SIX.

Summary of aggregated forecasting models

	Poll Aggregation	Election Aggregation	Poll + Election Aggregation
Foundational Article	Jackman (2005)[24]	Abramowitz (2008)[25]	Linzer (2012)[26]
Model Definition	Model-based Averaging	Fundamental Model	Combinatorial Model
Method	Aggregating multiple polls and then correcting for specific polls bias	Using past election polling and other data to predict future elections	Aggregating multiple polls and using extrapolation from past elections
Data	National and States polls 100s to 1000s	Past elections: 16 US elections since 1948 (1948 is the advent of modern poll)	National and States polls plus past elections
Variables	Single Polls and poll character-istics (polling firm, sample size, etc)	Vote Share, GDP, incumbency, consumer confidence, approval ratings	Single Polls and poll character-istics (polling firm, sample size, etc)
Estimation technique	Markov Chain Monte Carlo[27]	Linear regression	Markov Chain Monte Carlo
Forecasters	Jackman, Silver, and Wang	Abramowitz	Linzer
Strengths	Polling information is real-time information that captures the campaign dynamic	Not dependent on campaign noise like the polls	Combines the strengths of the poll and election aggregation approaches
Weakness	Polls often are false positives, or capture noise. If there is a systematic bias in all polls then no way to correct	Very dependent on variables in the model. Also, assumption that past behavior predicts future behavior	More complicated. Still is depended on the assumption of poll and election aggregation—always problem of being out of sample

polls into an overall average and (2) they correct for biases in individual polls (known as house effects). Jackman of HuffPo's Pollster.com and Stanford, Wang of Princeton, and Nate Silver all employ their own version of model-based averaging.

Such models are very dynamic as they capture real-time information from voters via polls. However, there can be systematic bias in polls, like under-coverage of minorities or misidentifying who will show on election-day. Additionally, polls can be noisy and produce 'false positives' in the short term as they jump around due to events such as party conventions or debates.[28]

In contrast, election aggregation approaches (or fundamentals models) do not incorporate horse race polls, instead relying on past electoral behavior to predict future behavior. Often such models employ both political and economic variables to estimate vote share:

Vote Share = Political variables and Economic variables

There are many derivations of these models.[29] But political variables may include party-of-the-president, approval ratings, incumbency, or number of years in power. Economic variables may include unemployment, consumer confidence, or GDP among others. Fundamentals models overcome the specific noisiness of poll aggregation as they depend on past elections and not current horse race polls. That said, there are three weaknesses to fundamentals models.

First, they assume that past is prologue: that past behavior will predict future behavior, which is not always the case. Second, fundamentals models depend on the variables inputted into them – omitting key variables is a risk which can be especially problematic when data is scarce. And third, such models rely on the election as the unit of analysis. Many countries have only a handful of elections, leading to challenges with small sample sizes and a heightened risk of rare events or outlier elections.

Finally, combinatorial models integrate the poll and election aggregation approaches. They include both the averaging of multiple polls as well as inputs from fundamentals models. Theoretically, this increase in information should improve model accuracy. That said, combinatorial models do suffer from the same underlying weaknesses of each individual model. It is always important to remember that all models are only as good as the data they use.

Figure SEVEN.

Performance of re-engineered forecasting models

Method	Adjustment/variables	Diff	Prob obama win (%)	Pundit forecaster
Model-Based Averaging	Simple average; no adjustments	0.9	80	Naive Poll Watcher Prediction
Model-Based Averaging	Taking out house effects	1.65	93	Jackman Model (actual diff 1.6)
Model-Based Averaging	Taking out house effects Adjustments for N-size Adjustment for most trusted firms	2.1	91	Silver Model (actual diff 2.2)
Fundamentals Model	GDP Approval rating # of years of incumbent party in power	3.7	81*	Abramowitz Model (actual diff 3.8)
Combined Model	Combines Model-Based Averag-ing with Fundamentals Model	4.0	98	Linzer Model (actual diff 3.9)

** To calculate the probability of an Obama victory for the fundamentals model (which does not count polls), we simply count the number of elections in which the model picks the right winner for US President (or 13/16).*

Is it possible to replicate these models?

In Figure 7 we detail results from re-engineered versions of the different forecasting models employed in 2012. In order to this, we employ two different datasets: (1) 539 national polls from Jan 1, 2012 to November 4th and (2) 16 US Presidential elections from 1948 to 2008.

Three benchmarks are used to assess these re-engineered models:

1. The last estimate of each of the 'forecaster-pundits';
2. The actual election results; and
3. A naïve poll aggregation estimate which does not correct for house biases.

So how close do I get to the forecaster-pundits?

The results in Figure 7 are based on simplified approximations of each forecaster's model. Even so, we get within a tenth of a point of their final estimate, showing that it is possible to replicate these models and to validate how well they predicted the election outcome.

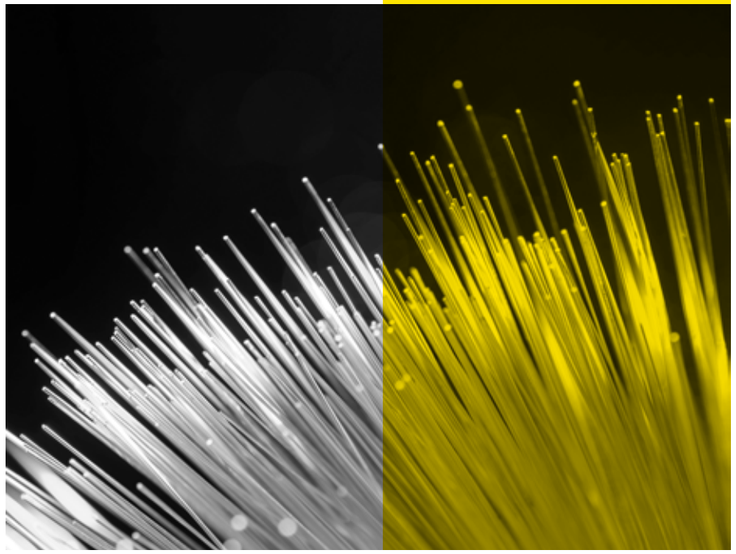
How good are the models?

All models out-perform the naïve averaging model which puts Obama only at about a 1 point advantage over Romney. Remember Obama ultimately wins by 3.9, meaning the polls as a whole were off by about 3-points!

Both the Jackman and Silver models are improvements over the naïve model, picking the right winner and assessing an Obama victory as highly probable. But we wouldn't call them 'highly accurate' being off by more than 1.5 points. The Abramowitz model does do a great job of estimating the correct vote share (3.8 vs. 3.9). So why not employ fundamentals models in place of poll aggregation?

Well, for two reasons: First, such models have an error rate of around 20% which is primarily a function of the small N-size of elections at hand (which increases the risk of an outlier election having more influence than it should). Second, fundamentals models work well until they don't; they depend on past behavior to predict future behavior. For most elections, this is a reasonable assumption, but not for discontinuity elections.

The Linzer combinatorial model performs the best. Such models have an advantage over their 'fundamentals' counterparts because they combine structural information from past elections together with real-time polling. Again, while they don't eliminate all problems, they do minimize the specific flaws of each individual model.



The power of big data in polling

So what did the 2012 election teach us about big data in politics? And what does it mean for polls and the polling industry?

First and foremost, the 'forecaster-pundit' is here to stay in one form or another. With the quantity of polling data out there, media and political analysts will increasingly take advantage of the inherent superiority of poll aggregation over single-poll analysis. With this, any one poll or polling firm becomes less relevant to the overall political narrative.

The poll aggregation movement and the 2012 election have provided many insights into big data as a concept. Specifically, big data does not always have to be, well, BIG. The aggregated data sets we used here are no larger than a few gigabytes of memory – much smaller than those usually associated with big data.

Cheaper field costs allow for much larger sample sizes and for polls which can always be in the field.

So if not size then what does 2012 tell us about big data? In our view, the primary lesson is about the aggregation of heretofore nonexistent or inaccessible information and how aggregation, in turn, can innovate our analytical reasoning. Specifically, many repeated measures of the same thing increase our analytical power. It reduces the uncertainty of our estimate and pushes us from thinking deterministically to probabilistically. No longer do we need to assess the world in 'binary single poll' terms: Obama will win yes/no. Instead, we can assign a number from 0 to 100. This gives us incredible analytical flexibility.

But the 2012 Presidential election also shows the shortcomings of big data. In particular, that your models are only as good as the data they use. Most of the models based solely on poll aggregation underestimated the Obama victory because the polls themselves underestimated Obama.

So what of the traditional pollster? Will they go by the way of the typewriter? Our answer is both yes and no. The single poll loses its relevance in the context of poll aggregation. But, at the same time, polls still help us understand voter motivations which often are washed out in a big data context.

Cheaper field costs allow for much larger sample sizes and for polls which can always be in the field. We call this continuous polling. The best example is the present Ipsos-Reuters online poll which is in the field every-day and includes about 11,000 interviews per month. Such continuous polling allows for a much more detailed analysis of subgroups (e.g., blue collar democrats in Ohio) as well as the ability to capture public opinion on key events within the same day.

The 2012 US presidential election could be classified as a high information election: it had many polls and elections to aggregate. Not all countries and electoral contexts have such large volumes of information. Would our models here work in low information environments? In countries with only a few polling firms? Or with only a few elections? Our initial response to this question is no—in low information elections our models must be different. But, even so, aggregation in some form or another will be our solution.

The data rich society

Exploring a new way of working with numbers

Henri Wallard, Paris

The basis of many published articles – in science, economics, research, media and so on – includes data to support the authors' arguments. Yet data and statistics cause a wide range of reactions: from skepticism about accuracy and objectivity to the current hype and fascination around big data. In some cases data is revered as an idol, while in others it is disparaged as a possible source of bias and manipulation.

Is big data the answer?

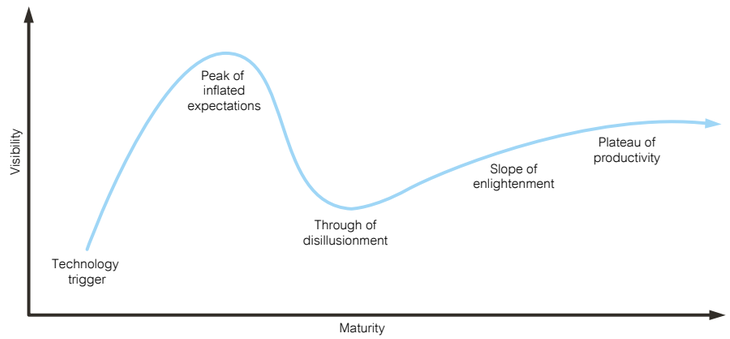
In an essay entitled 'The End of Theory' published in 2008 in Wired the author argues that 'with enough data, the numbers speak for themselves' , and 'correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all'.[30] Clearly, this is a radically different opinion of data and statistics from the refrain popularized by Mark Twain, 'Lies, damned lies, and statistics'.

Large datasets combined with algorithms and IT architectures that enable machine learning offer enormous opportunities, and we should definitely think positively of this new world. It is true that it is sometimes possible to predict well without fully understanding the underlying mechanisms. This has become our everyday experience in many parts of our lives (e.g. spam filtering software, and online service optimization).

However we should not disregard the knowledge accumulated from past experience with regard to the limitations of data sources, model design, and statistical expertise. We should grasp new thinking and welcome the contribution of new sources and new methods but not drop our guard. In particular, there should be healthy challenge when data are used in the fields of social and political research.

It is reassuring to see that the vague term 'big data' has already been challenged. An article in the Financial Times even asked the question 'Big Data: are we making a big mistake?'[31]

Figure EIGHT



It is easy to make mistakes with data. Therefore, we must not lose a healthy dose of skepticism.

The Hype Cycle is one useful way to think about the maturity and adoption of technologies and applications like big data.[32] The cycle illustrates the usual stages new technologies go through before reaching their most productive. Some of the key big data technologies are at or beyond the 'peak of inflated expectations', and heading downward to the 'trough of disillusionment'.[33] It may be some time before the real benefits of big data approaches are felt more broadly.

In any case, big 'bad' data is just bad, as the election of Franklin Roosevelt in 1936 reminds us. Famously, a newspaper using a large but biased sample source got its prediction wrong, while George Gallup predicted the outcome more accurately with 800 times fewer responses but on a better selected sample.[34]

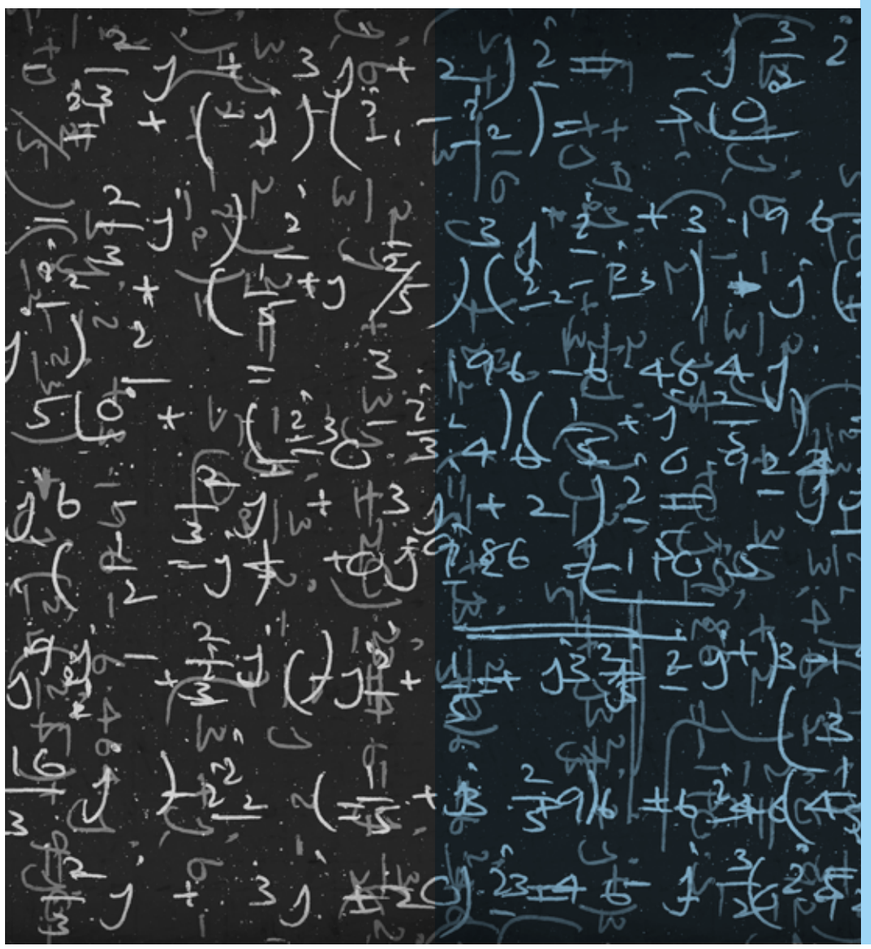
The importance of knowing more

So bigger is not always better and definitely data does not speak for itself regardless of how much is available. With the proliferation of data sources, the growing use of data visualization, and the development of data journalism, we must make sure we allow citizens to be informed with the right context and background when it comes to using data as evidence. Data should not be considered as having a life of its own; it must be associated with other information such as the source, how it was acquired, what precision should be expected, and any other caveats .

An interesting example was the public debt to GDP debate. In 2010, Harvard professors Carmen Reinhart and Kenneth Rogoff published papers on public debt and growth.[35] They suggested that the average real GDP growth declines when the ratio of public debt to GDP is high: from circa 4% growth if the ratio is below 30% to a small or slightly negative growth when the ratio is above a threshold of 90%. This conclusion has been widely publicized and was used in the political debate both in the US and Europe to justify controlling public debt and to promote austerity. [36]

In 2013, a student from the University of Massachusetts Amherst found errors in the calculation. He discovered that countries at the beginning of the alphabet had been omitted because of an Excel error, noticed that some observations had been rejected, and challenged the computation methods.[37] Together with his professors they published their own findings in April 2013 rejecting the idea of a sharper decline in growth above the threshold of 90% for debt to GDP ratio. This has led to numerous comments in the press and blogs.

But there was a much more substantial issue here than just differences in computation. The basic problem is that these data were too dispersed to provide a convincing link between the debt to GDP ratio and the rate of GDP growth. With a sufficiently large number of observations, you tend to find that variables are correlated at least a little. However this does not provide a convincing link if the dispersion of the data is too large. Statisticians use a coefficient of determination, noted R^2 (R squared) to measure the linear strength of a relationship between variables. This R squared value always sits between 0 (no correlation) and 1 (perfect correlation). In the GDP debt dataset, the R squared value was 0.04. Statisticians would say that 96% of the variation in growth is 'unexplained' by the debt to GDP ratio. This is far too low in terms of explanatory power and cannot lead to convincing conclusions.



Bigger is not always better and definitely data do not speak for themselves regardless of how much is available.

As for the classic confusion between correlation and causation, assuming causality can lead even the best intentioned towards blinkered views. For instance, people do not commit crimes just because they are poor or belong to certain groups; other factors will be important in explaining criminal behaviour. But in many articles simple correlations are treated as a comprehensive explanation, and these important caveats are absent. This can lead to potentially dramatic misinterpretation.[38] So just as data does not speak for itself, neither does correlation supersede causation.

Lastly, the availability of large datasets from social networks does not mean that all data should be seen as representative. For instance, there can be multiple anonymous users or even involuntary selection bias, particularly where only partial access to the data is available. As such who has access to the data is also a source of power and influence over the political debate.

It is easy to make mistakes with data. Therefore, we must not lose a healthy dose of skepticism, and always take into account where data comes from and how they are processed. Big Data and machine learning offer new avenues of knowledge and reinforce the permanent thirst for valid data and sound interpretations. One of our challenges will be to fight prejudice, biased, wrong or unethical use of data. As such, data scientists and statisticians will play a key role in the political and social debates to come.

REFERENCES

1. http://www.demos.co.uk/files/The_Data_Dialogue.pdf
2. http://www3.weforum.org/docs/WEF_IT_RethinkingPersonalData_Report_2012.pdf
3. http://www.deloitte.com/assets/Dcom-UnitedKingdom/Local%20Assets/Documents/Market%20insights/Deloitte%20Analytics/uk-da-data-nation-2013.pdf
4. http://www.ipsos-mori.com/newsevents/events/120/Personalisation-vs-Privacy-The-Balance-of-Public-Opinion.aspx
5. See http://reports-archive.adm.cs.cmu.edu/anon/isri2005/CMU-ISRI-05-138.pdf for more details
6. http://www.people-press.org/2013/07/26/government-surveillance-a-question-wording-experiment/
7. http://conversation.which.co.uk/technology/length-of-website-terms-and-conditions/
8. http://www.theguardian.com/world/the-nsa-files
9. http://www.pulitzer.org/citation/2014-Public-Service
10. http://www.shirky.com/
11. http://www.cjr.org/cover_story/evgeny_vs_the_internet.php?page=all
12. http://www.pbs.org/newshour/bb/new-police-surveillance-techniques-raise-privacy-concerns/ http://www.nap.edu/catalog.php?record_id=18749
13. http://prismsproject.eu/
14. Belgium, Denmark, Estonia, Germany, Hungary, Portugal, Romania and the UK
15. Mydex: http://mydex.org/
16. Qiy: https://www.qiy.nl/
17. Midata: http://www.midatalab.org.uk/
18. I've had three main involvements in this field in the past: at Demos in the 1990s we ran some big research programmes on privacy, the Internet and trust which in retrospect were far ahead of their time http://www.demos.co.uk/files/thefutureofprivacyvolume2.pdf. In government I oversaw the Cabinet Office review of privacy and data sharing in 2002 (which, ironically, can no longer be accessed on-line). In both cases we concluded that, with the right safeguards, the public stood to gain more from the free flow of personal data than they were losing. But the balance of the issue has changed markedly – one reason why more recently I became involved in helping the creation of Mydex, a project to give the public much more control. A good recent overview of the field which sets out current European thinking is 'Digital Enlightenment Yearbook 2013: the value of personal data' published by the Digital Enlightenment Forum.
19. Sasha Issenberg (2013) The Victory Lab: the Science of Winning Campaigns
20. http://realclearpolitics.com/
21. http://www.huffingtonpost.com/news/pollster/
22. Alex Knapp (2012) http://www.forbes.com/sites/alexknapp/2012/11/07/election-forecaster-sam-wang-on-the-future-of-polling-and-punditry/
23. Silver, N. (2012). The signal and the noise: Why so many predictions fail-but some don't.
24. Jackman, S. (2005). Pooling the polls over an election campaign. Australian Journal of Political Science, 40(4), 499-517
25. Abramowitz, A (2008) Forecasting the 2008 presidential election with the time-for-change model PS: political science & politics 41: 691-5; Montgomery, J. M., Hollenbach, F. M., & Ward, M. D. (2012). Improving predictions using ensemble Bayesian model averaging. Political Analysis, 20(3), 271-291
26. Linzer, D. A. (2013). Dynamic Bayesian Forecasting of Presidential Elections in the States. Journal of the American Statistical Association, 108(501), 124-134.
27. Jackman (2000) Estimation and Inference Are Missing Data Problems: Unifying Social Science Statistics via Bayesian Simulation Political Analysis, Vol. 8, No. 4 (Autumn)
28. Holbrook, T. (1996). Do campaigns matter? (Vol. 1). Sage.
29. Hummel, P., & Rothschild, D. (2013). Fundamental models for forecasting elections.
30. http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory
31. http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz33fzwJ3n8
32. http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp
33. http://www.informationweek.com/big-data/software-platforms/big-data-meets-trough-of-disillusionment-gartner/d/d-id/898939
34. http://pollingmatters.gallup.com/2010/10/75-years-ago-first-gallup-poll.html
35. http://www.nber.org/papers/w15639
36. http://thinkprogress.org/economy/2013/04/16/1875541/11-republicans-who-cited-a-flawed-study-to-push-for-drastic-spending-cuts/
37. http://www.peri.umass.edu/236/hash/31e2ff374b6377b2ddec04deaa6388b1/publication/566/
38. http://www.motherjones.com/kevin-drum/2012/11/single-mothers-now-hook-70s-crime-wave

FIGURE References

1. Source: Ipsos Global Trends Survey
   Base: 16,039 adults across 20 countries, online, 3-17 September 2013, data is weighted
2. Source: Ipsos Global Trends Survey
   Base: 16,039 adults across 20 countries, online, 3-17 September 2013, data is weighted
3. Source: Ipsos Global Trends Survey
   Base: 16,039 adults across 20 countries, online, 3-17 September 2013, data is weighted
4. Source: Ipsos Global Trends Survey
   Base: 16,039 adults across 20 countries, online, 3-17 September 2013, data is weighted
8. http://www.gartner.com/technology/research/methodologies/hype-cycle.jsp

About Ipsos Social Research Institute

The Social Research Institute works closely with international organisations, local public services and the not-for-profit sector. Research staff focus on public service and policy issues. We provide clients with information that helps them understand how they can build efficient and effective policies, programs, communications, strategies and marketing initiatives. This, combined with our methodological and communications expertise, ensures that our research makes a difference for decision makers and communities worldwide.

Chairman

Henri Wallard
+33 1 41 98 90 15
henri.wallard@ipsos.com

Global Chief Executive Officer

Darrell Bricker
+1 416 324 2001
darrell.bricker@ipsos.com

Global Director and Europe

Bobby Duffy
+44 20 7347 3267
bobby.duffy@ipsos.com

North America

Clifford Young
+1 202 420 2016
clifford.young@ipsos.com

Mike Colledge
+1 613 688 8971
mike.colledge@ipsos.com

South America

Alfredo Torres
+511 610 0100
alfredo.torres@ipsos.com

Asia-Pacific

Mark Davis
+61 3 9946 0888
mark.davis@ipsos.com

Middle East and Africa

Mari Harris
+27 11 709 7800
mari.harris@ipsos.com

Ipsos Public Affairs
The Social Research and Corporate Reputation Specialists