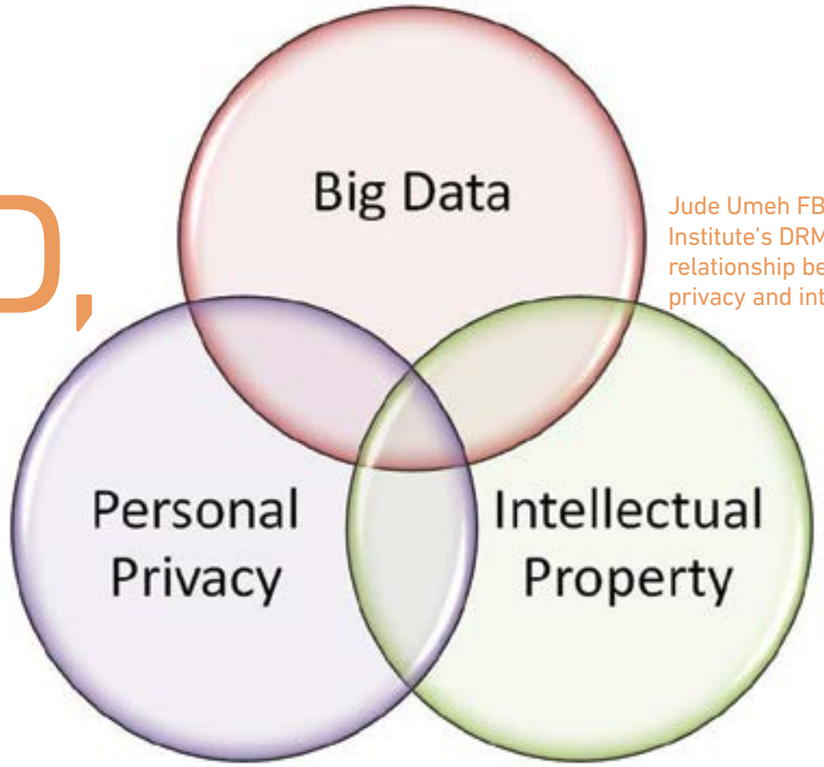
BD, P, IP



Jude Umeh FBCS CITP, the Institute's DRM blogger, looks at the relationship between big data, privacy and intellectual property.

Along with cloud, social and mobility, big data (aka information) is one of four key technology forces which, according to Gartner's Nexus of Forces, have combined to create a paradigm shift in the way we do business.

In a previous article (see refs)on this topic, I discussed how the nexus of forces impacts the rather more fundamental concept of intellectual property. In this article, we shall dive a little deeper into the key issues that impact and influence big data.

A little web research will bring up vast amounts of information and links to articles on the topic of big data. On closer inspection, however, only two or three main issues appear capable of making or breaking the promise of big data, and these are related to: solution approach, personal privacy and intellectual property (IP).

The first issue deals with technology, deployment and the organisational context, whereas the latter two big-ticket items raise concerns about the nature and applicable use of information or big data.

For the purpose of this article we'll pay more attention to the latter issues, mainly because sparks tend to fly whenever the commercial exploitation of information and content enters into the realm of personal privacy and IP rights.

Big data

According to a recent Forrester Research paper, typical firms tend to have an average of 125TB of data, but will actually utilise only 12 per cent of it. This shocking statistic brings home the key attribute/challenge of big data; which is the sheer volume, velocity and variety of data that reside and travel across multiple channels and platforms within and between organisations.

Think of all the personal information that is stored and transmitted through ISPs, mobile network operators, supermarkets, local councils, medical and financial service organisations (e.g. hospitals, banks, insurers, and credit card agencies).

Also, not forgetting information shared and stored on social networks, by religious organisations, educational institutions and or employers. Each organisation has the headache of organising, securing and exploiting their business, operational and customer data.

Incidentally, the information is increasingly comprised of unstructured data, such as: video, audio, image and written content, which require a lot more effort and intelligence to process.

As a result, many organisations have turned to ever more advanced analytics and business intelligence (including big data and social media) solutions to extract value from this sea of information, in order to create and deliver better and more personalised services to the right customer, at the right time.

Personal privacy

Given such powerful tools, and the large amount of replicated information spread across various sources, it is much easier to obtain a clear picture of any individual's situation, strengths and limitations. Furthermore, the explosion in speed, types and channels of interaction, enabled by components of Gartner's nexus of forces, may have brought about a certain degree, (perhaps even an expectation and acceptance), of reduction in personal privacy.

However, people do still care about what and how their personal information is used, especially if it could become disadvantageous or harmful to them. There is a certain class of data which can easily become 'toxic' should a company suffer any loss of control, and it include: personal information, strategic IP information, corporate sensitive data (e.g. KPIs and results)

The situation is further complicated by differing world views on personal privacy as a constitutional or fundamental human right. The UK's Data Protection Act is not applicable to personal information stored outside of the UK, yet we deal daily with organisations, processes and technologies that are global in scale and reach. On the other hand, some users are happy to share personal data in exchange for financial gain.

According to a recent SSRN paper, data protection and privacy entrepreneurship may have their place, but 'people should not have to pay to protect their privacy or receive coupons as compensation', especially as this might further disadvantage the poor.

> Every last piece of the aforementioned 125TB of big data held within your average organisation will have some associated IP rights.

Intellectual property

In addition to the above points, organisations also have to deal with the drama of IP rights and masses of unstructured data. Simply put, every last piece of the aforementioned 125TB of big data held in your average organisation will have some associated IP rights that must be taken into consideration when collecting, storing, processing or storing all that information. According to legal experts, companies need to think through fundamental legal aspects of IP rights e.g. 'who owns the input data companies are using in their analysis, and who owns the output?'

An extreme scenario: Imagine how that corporate promotional video, shot on location with paid models and real people (sans model release), plus uncleared samples in the background music, which just went viral on a number of social networks, could end up costing a lot more than was ever intended. Oh, by the way, the ad was made with unlicensed video editing software, and is freely available to stream or download on the corporate website and on YouTube. Well, such an organisation will most likely get sued, and perhaps should just hang a sign showing where the lawyers can queue up. Every challenge brings an opportunity, but not always to the same person.

Now imagine all that content, and tons more like it (including employees 'personal' content), just sloshing around in every organisation, and you might begin to perceive the scale of the problem. In fact, this creates a lucrative opportunity for big data mining and analysis algorithms, specifically designed perhaps for the computer audit and forensic investigations market.

Pointing the way forward

Here are three key things that organisations should bear in mind when seeking to deal with issues and problems posed by big data, privacy and IP:

1. Information is the lifeblood of business – therefore treat with due respect and implement the right policies for big data governance. The right information, at the right time, and for the right user, is the holy grail for business and it demands capabilities in data science, and increasingly in data art (visualise data in a meaningful / actionable ways).
2. Soon it may not even matter who owns personal data - personal information is becoming another currency with which the customer can obtain value. There is a growing push to focus big data governance and controls on data usage rather than data collection.
3. It's not the tool, but how you use it – technology is not really that much a differentiator, rather it is the architecture and infrastructure approach that make all the difference - e.g. Forrester recommends the 'hub and spoke' model for decentralised big data capability.

It would seem that the heady combination of big data, privacy and IP could be lethal for any organisation; basically, if privacy issues don't get you, then IP issues will likely finish off the job.

On the contrary, there are real opportunities for organisations to get their houses in order, by putting in place the right policies and principles for big data governance, in order to reap the immense benefits that big data insights can bring.

REFERENCES

Gartner - 'Information and the Nexus of Forces: Delivering and Analyzing Data' - Analyst: Yvonne Genovese

BCS TWENTY:13 ENHANCE YOUR IT STRATEGY - 'Intellectual property in the era of big and open data.

Forrester - 'Deliver On Big Data Potential With A Hub-And-Spoke Architecture' – Analyst: Brian Hopkins

SSRN - 'Buying and Selling Privacy: Big Data's Different Burdens and Benefits' by Joseph Jerome (Future of Privacy Forum) http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2294996

Out-law.com - 'Big data: privacy concerns stealing the headlines but IP issues of equal importance to businesses' – http://www.out-law.com/en/articles/2013/march/big-data-privacy-concerns-stealing-the-headlines-but-ip-issues-of-equal-importance-to-businesses-says-expert/

BCS Edspace Blog 'Big data: manage the chaos, reap the benefits' Marc Vael - www.bcs.org/blogs/edspace/bigdata

Capping IT Off – 'Forget Data Science, Data Art is Next!' - Simon Gratton www.capgemini.com/blog/capping-it-off/2013/07/forget-data-science-data-art-is-next

doi:10.1093/itnow/bwt035 ©2013 The British Computer Society



WHAT IS BIG DATA?

Keith Gordon MBCS CITP, former Secretary of the BCS Data Management Specialist Group, looks at definitions of big data and the database models that have grown up around it.

Whether you live in an 'IT bubble' or not, it is very difficult to miss hearing of something called big data nowadays. Many of the emails hitting my inbox go further and talk about 'big data technologies'. These fall into two camps: the technologies to store the data and the technologies required to analyse and make sense of the data.

So, what is big data? In an attempt to find out I attended a seminar put on by The Institution of Engineering and Technology (IET) late last year. After listening to five speakers I was even more confused than I had been at the beginning of the day. Amongst the interpretations of the term 'big data' I heard on that day were:

- Making the vast quantities of data that is held by the government publically available, the 'Open Data' initiative. I am really not sure what 'big' means in this scenario!
- For a future project, storing, in a 'hostile' environment with no readily-available power supply, and then analysing in slow time large quantities of very structured data of limited complexity. Here 'big' means 'a lot of'.
- For a telecoms company, analysing data available about a person's previous web searches and tying that together with that person's current location so that, for instance, they can be pinged with an advert for a nearby Chinese restaurant if their searches have indicated they like Chinese food before they have walked past the restaurant. Here 'big' principally means 'very fast'.
- Trying to gain business intelligence for the mass of unstructured or semi-structured data an organisation has in its documents, emails, etc. Here 'big' equates to 'complex'.

So, although there is no commonly accepted definition of big data, we can say that it is data that can be defined by some combination of the following five characteristics:

- Volume – where the amount of data to be stored and analysed is sufficiently large so as to require special considerations.
- Variety – where the data consists of multiple types of data potentially from multiple sources; here we need to consider structured data held in tables or objects for which the metadata is well defined, semi-structured data held as documents or similar where the metadata is contained internally (for example XML documents), or unstructured data which can be photographs, video, or any other form of binary data.
- Velocity – where the data is produced at high rates and operating on 'stale' data is not valuable.
- Value – where the data has perceived or quantifiable benefit to the enterprise or organisation using it.
- Veracity – where the correctness of the data can be assessed.

Interestingly, I saw an article from *The New York Times* about a group that works for the council in New York. They were faced with the problem of finding the culprits who were polluting the sewers with old cooking fats.

One department had details of where the sewers ran and where they were getting blocked, another department had maps of the city with details of all the restaurants and a third department had details of which restaurants had contracts with disposal companies for the removal of old cooking fats.

Putting that together produced details of the restaurants that did not have disposal contracts and were close to the blockages and which were, therefore, possible culprits. That was described as an application of big data, but there was no mention of any specific big data technologies. Was it just an application of common sense and good detective work?

The technologies

More recently, following the revelation from Edward Snowden, the American whistle-blower, the *Washington Post* had an article explaining how the National Security Agency is able to store and analyse the massive quantities of data it is collecting about the telephone, text and online conversations that are going on around the world. This was put down to the arrival, within the last few years, of big data technologies.

But it is not just government agencies that are interested in big data. Large data-intensive companies, such as Amazon and Google, are taking the lead in some of the developments of the technologies to handle big data.

Our beloved SQL databases, based on the relational model of data, do not scale easily to handle the growing quantities of structured data and have only limited facilities for handing semi-structured and unstructured data. There is, therefore, a need for alternative storage models for data.

Collectively, databases built around these alternative storage models have become known as NoSQL databases, where this can mean 'NotOnlySQL' or 'No,NeverSQL' depending on the alternative storage model being considered (or, indeed, your perception of SQL as a database language).

There are over 150 different NoSQL databases available on the market. They all achieve performance gains by doing away with some (or all) of the restrictions traditionally associated with conventional databases in exchange for scalability and distributed processing. The principal categories of NoSQL databases are key-value stores, document stores, extensible record (or wide-column) stores and graph databases, although there are many other types of NoSQL databases.

Our beloved SQL databases, based on the relational model of data, do not scale easily to handle the growing quantities of structured data.

A key-value store is where the data can be stored in a schema-less way, with the 'key-value' relationship consisting of a key, normally a string, and a value, which is the actual data of interest. The value itself can be stored using a datatype of a programming language or as an object.

A document store is a key-value store where the values are specifically the native documents, such as Microsoft Office (MS Word and MS Excel, etc), PDF, XML or similar documents. Whilst every row in a table in an SQL database will have the same sequence of columns, each document could have data items that are completely different.

Like SQL databases, extensible record stores, or wide column stores, have 'tables' (called 'super column families') which contain columns (called 'super columns').

However, each of the columns contains a mix of 'attributes', similar to key-value stores. The most common NoSQL databases, such as Hadoop, are extensible record stores.

Graph databases consist of interconnected elements with an undetermined number of interconnections and are used to store data representing concepts such as social relationships, public transport links, road maps or network topologies.

Storing the data is, of course, just part of the story. For the data to be of use it must be analysed and for this a whole new range of sophisticated techniques are required, including machine learning, natural language processing, predictive modelling, neural networks and social network mapping. Sitting alongside these techniques are a complementary range of data visualisation tools.

Big data has always been with us, whether you consider it as a volume issue, a variety issue, a velocity issue, a value issue or a veracity issue, or a combination of any of these. What is different is that we now have the technologies to store and analyse large quantities of structured, semi-structured and unstructured data.

For the data to be of use it must be analysed and for this a whole new range of sophisticated techniques are required, including machine learning, natural language processing, predictive modelling, neural networks and social network mapping.

For some this is technically challenging. Others see the emergence of big data technologies as a threat and the arrival of the true big brother society.

The BCS Data Management Specialist Group web pages are at: www.bcs.org/category/17607

Image: Thinkstock/Stockbyte
doi:10.1093/itnow/bwt037 ©2013 The British Computer Society



Adam Davison MBCS CITP asks whether big data means big governance.

For the average undergraduate student in the 1980s, attempting to research a topic was a time consuming and often frustrating experience. Some original research and data collection might be possible, but to a great extent, research consisted of visit to a library to trawl through text books and periodicals.

Today the situation is very different. Huge volumes of data from which useful information can be derived are readily available - both in structured and unstructured formats - and that volume is growing exponentially. The researcher has many options. They can still generate their own data, but they can also obtain original data from other sources or draw on the analysis of others. Most powerfully of all, they can combine these approaches allowing great potential to examine correlations and the differences. In addition to all this, researchers have powerful tools and technologies to analyse this data and present the results.

In the world of work the situation is similar, with huge potential for organisations to make truly informed management decisions. The day of the 'seat of the pants' management is generally believed to be on the way out, with future success for most organisations driven by two factors: what data you have or can obtain and how you use it.

However, in all this excitement, there is an aspect that is easy to overlook: governance. What structured and processes should organisations put in place to ensure that they can realise all these possibilities?

Equally importantly, how can the minefield of potential traps waiting to ensnare the unwary be avoided? Can organisations continue to address this area in the way they always have, or, in this new world of big data, is a whole new approach to governance needed?

What is clear is that big data presents numerous challenges to the organisation, which can only be addressed by robust governance.

Most of these aren't entirely new, but the increasing emphasis on data and data modelling as the main driver of organisational decisions and competitive advantage means that getting the governance right is likely to become far more important than has been the case in the past.

Questions, questions

To start with there is the question of the overall organisational vision for big data and who has the responsibility of setting this? What projects will be carried out with what priority? Also one has to consider practicalities – how will the management of organisational data be optimised?

Next we come to the critical question of quality. Garbage in, garbage out is an old adage and IT departments have been running data cleansing initiatives since time immemorial. But in the world of big data, is this enough? What about the role of the wider organisation, the people who really get the benefit from having good quality data? There is also the issue that a lot of the anticipated value of big data comes not just from using the data you own, but from combining your data with external data sets. But how do you guarantee the quality of these externally derived data sets and who takes responsibility for the consequences of decisions made based on poor quality, externally derived data?

Although garbage in more or less guarantees garbage out, the opposite is not necessarily true. There are two elements involved in turning a data asset into something useful to the organisation; good quality data and good quality models to analyse that data. As was clearly demonstrated in the banking crisis, however, predictive models rarely give perfect results.

How therefore can organisations ensure that the that the results of modelling are properly tested against historic data and then re-tested and analysed against real results so the models and the data sets required to feed the models can be refined and improved? Above all, how can organisations ensure that the results of analysis are treated with an appropriate degree of scepticism when used as a basis for decision-making?

Confirmation bias

Also, when considering how such models are used, the psychological phenomenon of confirmation bias needs to be considered; the human tendency to look for or favour the results that are expected or desired. Inevitably analysis of data will sometimes give results that are counterintuitive or just not what was looked for, leading to the age old temptation to dismiss the results or massage the figures. What policies and processes are needed to ensure that this doesn't happen?

Another important governance issue is around how to protect the valuable data. The information security threat is constantly evolving and as big data becomes the critical driving force for many organisations, the risk of having their data asset compromised or corrupted becomes acute. Great clarity on who is responsible for managing this issue and how it is managed will be critical.

So, when starting to consider all these issues, the most fundamental question is; where should responsibility for these issues lie?

Generally speaking, four options tend present themselves:

- The CIO as the person responsible for managing the data asset;
- The person or people who get the benefit from the data asset;
- With a neutral third party;
- A mixture of the above.

As things stand, in many organisations, the CIO is the default answer. After all, the 'I' in CIO stands for information, so surely this should be a core responsibility? This approach does have some justification. CIOs are often the only people who have an overall understanding of what data, in total, the organisation owns and what it is used for. Also, the CIO tends to have practical responsibility for many of the issues listed above such as IT security (not quite the same as information security, however) and data cleansing (not quite the same as data quality).

However, the CIO typically has responsibility for managing the data. Is it therefore appropriate that he/she should also own the governance framework under which this data is managed? Furthermore, CIOs tend to have a wide range of responsibilities, so their ability to give sufficient focus to data/information governance could be limited. Finally, CIOs may not be ideally positioned when it comes to influencing behaviours across the organisation as a whole.

Diversity of Organisational Activities

Level of information dependency		Low	High
	Low	CIO or User	CIO
	High	User	CDO

Is it appropriate that the CIO owns the governance framework under which data is managed?

Responsibility with the user?

For many, having overall responsibility for data governance resting with the users, the people who gain benefit from the data, is an appealing concept. They are, after all, the people who have most to lose if good governance isn't applied. Again, however, there are downsides to this. Only in the relatively small organisation will it be practical for the user side to be represented by a single individual. More frequently, one runs the risk of ending up with a sort of governance by committee, with a range of stakeholders each with their own viewpoints. In this scenario, the chance of a consistent and appropriate governance model being created and such a model being successfully applied are very limited.

Faced with these issues, some organisations have chosen to take a third way and create the post of chief data officer (CDO): someone who has overall responsibility for organisational data but who sits outside of either (usually) IT or the end-user communities. This approach is in many ways attractive. It means that overall governance responsibility rests with someone who is able to focus themselves entirely on the issues related to data (not the case with either the CIO or the user community) and who can take an entirely neutral viewpoint when setting rules on how such data is managed, and used. However, issues again emerge.

The CDO concept can be undermined by the question of organisational authority to ensure that the decisions that they make are binding, particularly as CEOs, already under pressure from multiple directions for increased senior level representation, will naturally be reluctant to create yet another C-level role.

Finally there is the hybrid approach, for example sharing governance responsibility between the CIO and the users or putting a CDO in place to report to the CIO or a senior user figure such as a COO. It is certainly true that all significant stakeholder groups will need to be involved at some level in ensuring good governance around data. However, this again brings in the issues around governance by committee and unclear overall responsibilities.

Any of the above models could work, but ultimately, which of them will work is most likely to be highly influenced by the nature of the organisation. In general terms, therefore, the pictured model might apply.

However, this model does not take account of some further vital factors. For example, corporate culture is a key issue. In an organisation with a very strong cooperative culture, the hybrid approach might be the one to choose.

Last but not least, giving this important responsibility to an individual with the right experience and personality can be seen as being at least as important as their job title. Give the job to the right person and the chances are it will get done, give the job to the wrong person and the chances are it won't. What remains true in all cases, however, is that this issue will become more and more important and addressing it successfully is going to be of vital importance for all organisations.

Adam Davison MBCS CITP writes the Strategy Perspective Blog for BCS.
www.bcs.org/blogs/itstrategy

Image: iStockPhoto/DigitalVision/Ryan McVay
doi:10.1093/itnow/bwt038 ©2013 The British Computer Society



DATA MOUNTAIN

In our ever connected world we are relying more and more on data centres, but they use a lot of power. With this in mind, Henry Tucker MBCS went to see the self-proclaimed, greenest data centre in the world.

The dark blue water laps gently on the hard granite shoreline. Take just one step into the cold water and the drop is 70m straight down. Go a little further out and it can get as deep as 150m. These Norwegian fjords have been used for many things ever since man first laid eyes on them. Now they are being used to cool a data centre.

Green Mountain is no ordinary data centre though, even before it started using 8ºc fjord water to cool its servers. That's because not only is it quite green, literally and figuratively, but it is also a mountain. Well inside one.

Smedvig, the company that owns the data centre isn't the first to operate inside the mountain though. The tunnels that run up to 260m into the granite were drilled by NATO in the early 1960s after the Cuban missile crisis. Initially it was used to store field hospitals and then later to house and repair torpedoes and mines, now NATO has gone and the mountain is almost empty.

In many ways it makes the perfect location for a data centre. They aren't places that you want a lot of people going to, they need to be secure and they need to be cooled efficiently and effectively. With only one way in it is secure, the fjord isn't at risk from tsunami or earthquakes and with all the mountains and lakes around it, it isn't short of cheaper electricity from the network of hydroelectric plants in the area.

The fact that it has the hydroelectric power nearby would always be in its favour when it came to being environmentally friendly. However, in addition to this the company running it has designed an efficient way to cool it using the other thing that it has on tap; the fjord water.

As the water is so deep, when you get down to 100m it is a constant 8ºc all year round. This water is then drawn into a large concrete tank without using pumps because it is at sea level. As it is sea water it can't be used to directly cool anything inside the data centre as it would be too corrosive.

So what they do is use a closed, fresh water pipeline that draws the heat away from the racks, this is then cooled using the sea water and titanium heat exchangers. After this the sea water then goes out into the fjord again at a temperature of 18ºc.

As water is far more efficient and effective than air, Green Mountain claims that it gets 100KW of cooling from just 1KW of power.

A potential downside of its remote location, of course, could be the time data takes to get to and from the servers that reside there. Stavanger though is the hub of the Norwegian oil industry and so it already has fast connections to other parts of Northern Europe and to London. It claims to have only a 6.5ms latency to the UK.

As for its claim to being the greenest data centre in the world, Green Mountain CEO Knut Molaug explained this claim.

'It is because of the low CO2 emissions, we have virtually no CO2 emissions. As you know the data centre industry is very energy hungry and we, in Norway, have close to 100 per cent renewable energy.

'We use 100 per cent hydro power here. So that's number one. Number two, is that we built the system to be extremely efficient utilising the fjord outside the site for cooling. This means that we have one of the world's most efficient data centres in combination with using green energy and using former buildings in our efforts, and in everything we have built, we have put a green element in all the designs.'

Going green

The company's rationale for building the data centre came from wanting to build the facility, but also to try and do it in as environmentally a way as possible.

'It was a combination of both (wanting to build a data centre and wanting it to be green). We were discussing the possibility of building a data centre in other places around us, because the owners of Green Mountain are already owners of another data centre in Stavanger, and during the process of evaluating the possibility of building another data centre using water for cooling, this site came up for sale. So it was a combination of we were looking to build a green data centre and an opportunity that came along.'

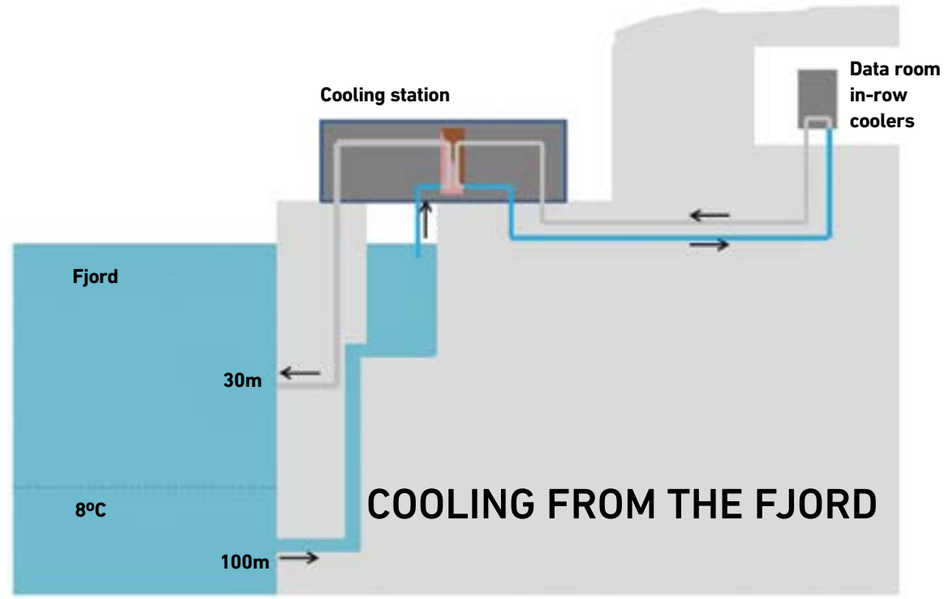
By creating what Green Mountain likes to claim is the greenest data centre they are hoping that other companies take their lead and make potentially greener data centres.

'It's the beauty of competition that when somebody stands out, someone will want to level you or pass. We hope that this spurs further development within green data centres.'

As to what is driving companies to choose to use the data centre, although it has excellent green credentials, Knut doesn't think that it is the main reason companies chose it.

'I think that the main driver for any business is money. The fact that we have green energy available, at low cost, is the main driver for almost all of them. Everyone would like to be green, but they don't want to pay for it. We can offer a cheaper alternative that, in addition, is green.'

Green Mountain is a good example of making the best use of the things you have around you in order to be as efficient as possible. With the data centre industry growing hopefully more will take on some of the features of Green Mountain to reduce their CO^2 footprint.



doi:10.1093/itnow/bwt040 ©2013 The British Computer Society



Image: iStockPhoto/DigitalVision/Ryan McVay

doi:10.1093/itnow/bwt041 ©2013 The British Computer Society

Louise Bennett FBCS, Chair of BCS Security, looks at the opportunities and dangers of one of the implications of big data: identity discovery through data aggregation.

Think for a moment about all the data that you have given to organisations when you signed up for a subscription or purchased a ticket. Add to that your loyalty card data and what you have posted to social networks, your browsing history and email, your medical and education records. Then add in your bank records, things friends and others have posted about you, memberships and even CVs posted to job sites. Pictorially it will look something like the picture above.

Are you happy about people joining all this data together into an aggregated view of your life and mining it? If they do so, what are the implications for privacy and will it benefit you or 'them' more?

There are many commercial models on the internet. Some services are free or below cost because there is value in the data that customers give up when they use those sites or services. The quid pro quo is usually targeted advertising. As Viviane Reding of the European Commission said on 22 Jan 2012, 'Personal data is the currency of today's digital market'. It is widely said that if you are not paying the full cost of a service you are a product, not a customer. Most young people either do not think about this or they accept it, and it can be a win-win situation. You can apparently get something for nothing, or almost nothing, if you pay for it with your identity attributes.

Do you need to get offline?

However, you may not want your identity attributes to be used and privacy may really matter to you. If that is the case, do you need to get offline and lose out on some deals you might be offered? What does big data mean for your privacy? Can you retain online privacy or is identity discovery through the aggregation of your personal data attributes inevitable?

Personal information disseminates over time into many different areas and once published on the internet it is improbable that it can ever all be deleted. There are also powerful commercial tools available to mine information about an individual or organisation. The next time you use a social media site or search engine consider what adverts or suggestions are made to you. They will often be tied to your habits.

For this reason, many people will want to use different identities for different activities on the internet to frustrate potential data aggregation. Many of us will feel there has been an invasion of our privacy if, out of the blue, a connection we deliberately withheld is made about us. For example, you may wonder: 'How on earth did the organisation my husband has just bought something from know my mobile phone number. We did not give it to them and it is in another name. So how could they text my smart phone to tell me his purchase will be delivered to our home tomorrow?'

Increasing regulation

Concerns about data aggregation and data mining on the internet are likely to increase rather than decrease in the coming years. There is also likely to be pressure for regulation because of the potential privacy implications. One example of this is the proposed new EU Regulation on Data Protection. This includes a section on 'the right to be forgotten'. However, if the Regulation ever gets agreed (which is unlikely with about 4,000 amendments tabled and a 2014 deadline before the EU elections), the right to be forgotten is one thing that will probably be removed.

Such a right is certainly technically challenging, if not impossible, in the internet age. The best privacy activists can hope for is a right to relative obscurity.

The online world increasingly uses a network of attributes to determine identity. If these attributes are just matched for a one-off identity check that is one thing, if they are stored and aggregated in big databases it raises more concerns. When we think about privacy, particularly in relation to commercialisation of the internet, government surveillance and data collection, it is revealing to consider the outrage at Edward Snowden's revelations about elected governments engaging in lawful espionage compared to the absence of concern that businesses (accountable only to their shareholders) have all this data in the first place.

Many individuals object to identity discovery through data aggregation, whether by governments or business. This is especially true where it is used to find out about a person's preferences and life, using data that the individual regards as sensitive, personal data. It is of even more concern when it is used for cyber-stalking and cyber-bullying, or transfers into the real world as stalking or other criminal activities.

It is revealing to consider the outrage at Edward Snowden's revelations about elected governments engaging in lawful espionage compared to the absence of concern that businesses have all this data in the first place.

This in turn can lead to people feeling it is legitimate to withhold information about themselves or provide incorrect information in responding to requests they feel are unjustified (e.g. mandatory fields on their age, ethnicity or religion being requested before they receive their goods or services). This is especially important where identity discovery is looking for attributes that are not actually identity attributes, but give information about a person's preferences or life choices (such as sexuality or membership of organisations).

Attributes of identity

The 'attributes' aspect of identity are key to the responsible use of big data. Everything is context dependent. We rarely engage completely online. Often the trust context is developed offline (through our friends or trusted brands) and carried through to the online experience. It is vital to determine what attributes are required in a particular interaction and how trustworthy attributes can be conveyed in a manner that maximises the benefits of the availability of those attributes, while minimising the disbenefits of revealing more attributes than are strictly needed. This requires detailed analysis and not broad generalisations.

While technology solutions may exist, the social and economic aspects of implementation are very complex. They are also very personal and will change for an individual over time, even in identical contexts. What was a playful prank at school could have implications when applying for jobs years later if it can be linked to your identity.

The opportunities

The pace of innovation in online commerce and delivery of government services is accelerating. By making everything digital, exploiting the power of big data and the ubiquity of mobile communications, there are huge opportunities to improve productivity, enhance the value to individuals and manage risks effectively.

While the potential upsides are great, the downsides are also stark. The downsides lie mainly in the potential loss of privacy (both real and perceived) and the erosion of trust, if those online cannot provide evidence of their trustworthiness in the context of the transactions they wish to make.

Context and demonstrable trustworthiness are key to the use of personal data attributes in the online world. They are blended with our experiences in the offline world. The success of 'bricks and clicks' commercial models is testament to this. Those who mine big data need to think very hard about how they monetise personal data attributes. They need to be transparent about what they are doing and provide evidence that they are trustworthy if they are to handle our attributes in an acceptable manner, and be successful in an online world.



DATA, DATA EVERYWHERE

As individuals living in a rich technology and communication ecosystem we capture, encode and publish more data than ever before. This trend toward greater amounts of data is set to increase as technology is woven ever more into the fabric of our everyday lives, says Ben Banks MBCS, European Information Security Manager, RR Donnelley.

Image: iStockPhoto/16013814 5

doi:10.1093/itnow/bwt047 ©2013 The British Computer Society

As information security and privacy professionals we are in the vanguard of navigating this new landscape. Our challenge is enabling commerce whilst ensuring our stewardship for these new assets remains strong.

This article explores one aspect of this challenging new world - when does data become information and what does that change mean for our assurance work?

From data to Information

Data and information are not synonymous. Although the terms data and information are often used interchangeably adopting a more rigorous understanding of them has important implications.

It is fairly intuitive that an instance of data, a data point, when considered in isolation is not information. For example 23 or 54, 46 are perfect good instances of data, but they are not particularly informative.

In order for data points to become information there must be a known relationship between data and what it encodes that makes it meaningful. In contrast information is always data because semantic meaning is always encodable as data. All information is data, but not all data is information.

It is vital that as information security professionals we enrich our understating of the ways data becomes meaningful.

To do this we need to consider in turn issues related to order, truth, association, size, brevity, resolution and causal efficacy.

Ordering the data

Knowing the meaning of data, i.e. what information it is, is the critical step in establishing its true value. So how does data become meaningful? The first way in which data becomes meaningful is order. I give you a list of four digit numbers from 0000 to 9999.

As a dataset I've just expressed all the credit card EMV PIN numbers in use in the UK and as information it has little meaning, and even less value. I now reorder that list putting all the numbers from 1930-1995 at the beginning and put values like 0000, 1111 etc at the end. I give the list a title of the 'most popular EMV PIN numbers in the UK'. Ordering changes the meaning, and by extension, the value of that list.

Truth

If we consider another order of our EMV PIN number list, this time titled 'the most popular EMV PIN numbers in the UK as reported in a secret credit card brands report' received from a legitimate source.

We could say that the ordered data is now more accurate. With complete accuracy the information becomes true. True data becomes more meaningful and highly valuable information.

The more data there is the more meaningful the information can become. Even if you don't see the actual record, the size of a data set changes the risk profile - a news report indicating a breach of five records has a very different negative value to a news report of a breach of 5 million records.

The greater the amount of data captured about a 'thing' the more informative it is likely to be.

Accuracy and truth are not synonymous. But we can assume that without accuracy, the truth of information would be hard to verify.

Association

A list of EMV PIN numbers, however well and truly ordered, has limited meaning.

For a criminal with some stolen credit cards they have better guesses about the appropriate EMV PIN to use with it, but with only three attempts per card the value is still limited.

As part of our thought experiment let us consider what if, on each row of the list, was an example valid primary account number (PAN) example written alongside.

The association of PIN and a valid PAN has given our data much more value. Association of different data elements is another route for adding a valuable meaning to data and the correlations indicated in the associations of data elements in different data sets is a fundamental deliverable of big data.

It is also important to note that when two individually benign data elements are brought together their meaning and value can be increased hugely.

Size

Size always matters. Imagine the increase in meaning if, instead of one valid PAN number, 50 valid PAN numbers were listed with their corresponding valid PIN.

Brevity

'If I had more time, I would have written a shorter letter' is commonly accepted as true. When bandwidth was costly short meaningful messages were more valuable. If we had a list of default PINS and some of the associated PAN numbers it would be a big data set.

That information could be re-expressed in a condensed format as the algorithm for generating a default PIN from a PAN (in truth this isn't actually how it works, but it does help to illustrate the point). Brevity condenses the content of data without loss of meaning and in so doing it becomes more valuable.

Causal efficacy

What can data you hold let you do? In order to dig into this question we need to change our perspective a little. Consider the question of what data you would need to supply to make an online payment when the card is not present.

Typically one needs a credit card number, a name, a card verification value and an expiry date associated together to make a valid transaction (assuming the sites you used didn't require an additional verification step).

Whilst payments require a number of data points to have a level of causal efficacy, consider how many data points you need to identify yourself to get access to online services.

Typically only two data points are required - an email and a password. Knowing what data enables makes it valuable as meaningful information.

Resolution

The greater the amount of data captured about a 'thing' the more informative it is likely to be.

Consider the difference in two CCTV cameras looking at the same scene from the same perspective, where one has an image resolution that can allow people to be identified from the footage, the other not.

One critical point to make with relation to resolution is linked to brevity. Maintaining meaning when aggregating, summarising, or otherwise reducing the resolution of the data set is often a more subtly difficult problem than people imagine.

What does it all mean?

Exploring data and information, and the critical role of meaning in influencing their value, will enhance how we manage the confidentiality, integrity and availability risks to them as assets.

Perhaps it is only as information that data has any inherent value worth protecting at all.

Likewise, by assuming neither data or information have intuitive, self-evident, definitions will help to reduce the subtle dangers associated with being part of a dialog or process that fails to see the dangers in the uncritical use of a common language.

And a final word should go to big data, as we move into datasets that are so vast that it may well blur the line between data and information.

Perhaps there is a notional critical mass after which a dataset, regardless of the content, is de facto meaningful.

www.bcs.org/security