NESTA Making Innovation Flourish

Big Data

The power and possibilities of Big Data

In November 2010 NESTA hosted an event to discuss 'The power and possibilities of Big Data'. The event brought together Hans Peter Brøndmo from Nokia, Haakon Overli from Dawn Capital, Max Jolly from dunnhumby and Megan Smith from Google to discuss the issues of Big Data with a diverse audience of entrepreneurs, academics, business people and investors.

This report draws together some of the key concepts that were discussed on the day. For the full detail, you can watch a video of the entire event on the NESTA website. You can also access a list of further reading, videos, case studies and examples on the Big Data resources page.

Hot Topics is a series of NESTA events driven by ideas and technologies. They aim to introduce the technological tools that will change how we do things in the coming years and are designed to bring together the best of business, academia, start-ups and investors.

Introduction and background

'Big Data' is a term applied to the rapid growth of data that has resulted from more automated collection methods and greater capacity for storage and processing. Digital data is growing at an estimated 60 per cent per annum. Around 1,200 exabytes (or 1.2 billion terabytes) of digital data are forecast to be generated this year, compared to 287 exabytes in 2007. This exponential rise is driven by the proliferation of sensors for gathering data automatically, including those in mobile phones, and more activity taking place online, which can be more easily recorded.

Although the use of large data volumes for business is not new, some things have changed, creating new opportunities for innovation. There are three key changes that have brought the issue of data onto many more agendas.

Firstly, data storage, processing power and cloud services continue to make large scale data analysis more and more accessible. You no longer need to build your own data centre to use this technique, expanding the pool of users.

Secondly, there are many more opportunities to capture data, from sensors in phones and RFID tags in products, as well as a greater social acceptance of contributing manually entered data to social services.

Thirdly, it is now possible to analyse unstructured data, so it is not necessary to run your business with detailed customer forms or electronic point of sale terminals to benefit from

this form of analysis. Natural text in emails, photographs and sound can all be analysed and 'mined' for insights, rather than only structured, coded information that needed to be captured electronically or manually coded.

These trends have expanded the pool of data that is available to be analysed, as well as the number of firms that are interested and able to use it. This raises issues of data ownership and privacy that are lagging behind the technological opportunities.

'Big Data' is not a new concept. Companies have been using data on their customers to improve their businesses for hundreds of years. Actuarial science has been used since the 17th century to calculate life insurance premiums. Walmart has collected huge volumes of customer data since the 1970s – by 2004, the New York Times reckoned its databases contained more data than the entire internet (460 Terabytes at that time).

Walmart made huge strides in retail, and understanding large data volumes to make operational decisions about its stores. Building upon the data capture capabilities of electronic point of sale (EPOS) systems, Walmart tracks sales of goods, identifying purchasing trends and optimum store arrangements. They discovered in 2004 that a hurricane forecast increased sales of Pop-Tarts along with flashlights, batteries and other emergency supplies.

> "Data is useless if we don't apply any intelligence to it" **Haakon Overli**

These insights were used to design better stores and to allocate stock more efficiently, ultimately pushing responsibility for stock management back up the chain to the supplier, by allowing them access to real-time sales data from the stores on their Retail Link system. Sam Walton's focus on continuous improvement – he is said to have tasked every regional VP with travelling to their stores every week, and finding an improvement that would save at least the cost of the flight – is likely to have contributed to this wave of innovation.

In the nineties, Capital One built a business by mining customer data to establish characteristics that predicted credit risk, and then testing predictions with credit card offers. They created many different credit card offers based on different customer characteristics, and sent them out by mail order. The response rates were tracked, and used to refine the offers made the next time, and the way people were targeted.

The power and possibilities of Big Data – a Hot Topics event at NESTA

This report draws together some of the key concepts that were discussed on the day. To watch a video of the event, go to: http://www.nesta.org.uk/events/previous_events/assets/events/silicon_valley_comes_to_nesta_the_power_and_possibilities_of_big_data

- Data needs attribution and context
- Using data to make happy customers
- Using data for social good

- The electronic soul – an ownership model for personal data
- Digital Media literacy

Data needs attribution and context

Data is useless without some intelligence being applied. Data is part of a continuum that extends to insight and then to knowledge only when intelligence and context are applied. As Hans Peter Brøndmo illustrated with a short opening line in Norwegian, you can have a perfectly structured piece of data, but without the right interpretation, it is meaningless (and was to most of the audience).

Max Jolly described the way that dunnhumby or Tesco consume data as similar to the way you do when waiting in a checkout queue. You look at the basket of the person ahead of you and make some sort of judgment about what that person is like – have they got a cat, have they got a baby, what kind of person are they buying for?

But the data, the list of goods, is not enough. You need to attribute meaning to that data, and put it into context. Attribution allows you to understand the customer, but requires a degree of interpretation to identify meaning behind the data. A customer who buys a large bag of value frozen chips can be sending a number of different signals: they chose own label not a brand, frozen not fresh. A large bag could mean large portion size, buying ahead or a large household. Context allows you to set the data point against other purchases made at the same time, the trend of previous purchases and other customer information such as location to generate more insight. Is this an unusual purchase, or part of a pattern?

Attribution allows you to understand the customer and context allows you to connect it together and build up a strong picture. These activities are used for many other types of data – filling in the gaps between data points to build trends and forecasts that can be used for business decisions.

The *Ushahidi* website (above) that uses SMS to map information into geographic data, and *Health Speaks* (below), a collaboration with Wikipedia to translate the most popular health articles into Arabic, Hindi and Swahili.

Using data to make happy customers

Haakon Overli painted a vivid picture of the old-fashioned shop keeper, managing customer data in his head. He knows his regular customers and what they like so is able to give them special offers to tempt them to buy more. He makes quick judgements about new customers to sell them premium items on the basis of their likely needs. It works well for him most of the time, but this approach isn't scalable (and there's nothing worse for a venture capital firm than a business that doesn't scale).

However, the principle remains strong – you can use data to make your customers happy by giving them what they want. Retailers are trying to scale up this process. As Max said, the challenge is how to put the customer at the heart of the all the decisions the business makes. To do this you need to 'democratise' the data – cut it into chunks that makes it possible for a category manager to make meaningful decisions on the slices of data that relate to his products and customers.

Retail has already had many years to absorb the possibilities of Big Data, but is only starting to get to grips with it. Max Jolly took a look at some other areas he expected to be influenced, including other forms of retailing, such as petrol stations. Even where there is a single product being sold, the location, time, payment amount and payment method can still give you information. What conclusions could you draw from a £10.01 cash purchase of petrol compared to £45.38 paid by credit card?

Television is another area likely to see the impact of Big Data. TV ratings are still provided by a panel with set-top boxes. With digital and on-demand TV services, it becomes much easier to calculate real viewing figures, and to segment audiences. That will open up much greater opportunities to use the data.

Haakon Overli also highlighted online targeting as an area of growth, describing a Dawn Capital investment, Cognitive Match. By segmenting online audiences using non-private data, they can show that a promotion that does badly overall still converts best for those browsing at the weekend using the Safari browser, for example.

There seems to be a minor market failure here as well: Haakon said that companies waste a lot of data that's free and pay a lot for data they can't use. The challenge for those programming the machines is to find the data that is predictive. Currently the process is too manual, and looks at what is good for most people rather than what's best for you.

Using data for social good

Recent government open data initiatives, as well as recent crises in Kenya and Haiti, have demonstrated the potential of data to be used for public services and social good.

Ushahidi was used by Megan Smith as a great example of the potential for good. Ushahidi is an open-source platform that started with a need to share data on the Kenyan riots in 2008. It uses SMS to map information onto geographic data, crowdsourcing knowledge about current events. The platform has since been expanded and deployed at sites across the world to collect and visualise data from Atlanta to Gaza.

> "Our 'electronic soul' is a valuable asset. Who owns it?" **Hans Peter Brøndmo**

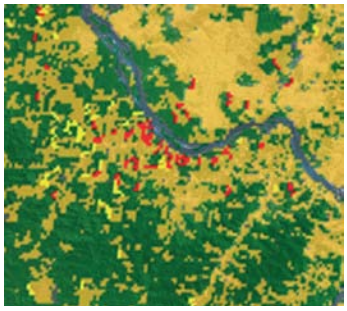
Sometimes just documenting what is happening in an accurate way, or with better visualisation, can be helpful on its own, and to help target resources.

Megan described the Google Earth engine project, part of the philanthropic work that is done through Google.org. The project makes historic satellite imagery, much of which is stored on tape, available online for groups to compute against, and allows them to map deforestation over time in detail. Google supplies data, storage, and computing muscle, speeding up analysis that took weeks with an offline system.

As well as mapping, translation is another area of opportunity for Big Data. Machine translation algorithms are advancing, and can be combined with crowdsourced approaches to provide very efficient translation services. Google has a great resource for translation – EU documents are produced in many languages and professionally translated. The same is true for many books in Google Book Search. They are currently collaborating on Health Speaks with Wikipedia to translate the most popular health articles into other languages, starting with Arabic, Hindi and Swahili. Increasing the amount of quality health information available in a local language has huge potential to improve healthcare.

The quality that has changed social and public sector use of data is adjacency. Being able to work with people or data across geographic and temporal boundaries expands the range of collaborators massively. Governments are starting to see the potential of putting disparate sources of data next to each other. Social entrepreneurship has also been boosted by this adjacency. Megan described trying to set up social entrepreneurship projects by post at college – you sent a letter to a project in Africa you wanted to work with, and you had probably graduated by the time you got the reply. Now social entrepreneurship is exploding in universities, because these connections can be created much more easily.

Connecting social needs to business uses, there are also opportunities for companies to play back messages to consumers to help them make better choices. At the moment, several online supermarkets will highlight cheaper alternatives to products you have selected, to help you save money. Similar information could be used to suggest healthier options, or low carbon options. However, some customers won't welcome these suggestions – it could be seen as helpful, or as interfering. Customers will need to opt-in to these services, but there is a noble goal in encouraging them to do so.



Results from the *Google Earth engine project* using existing satellite imagery and historical data to measure deforestation over time in Brazil.

The electronic soul – an ownership model for personal data

As Haakon said, when talking about Big Data, privacy is never far behind. In his opening statement, Hans Peter advocated an ownership model for the 'electronic soul'.

As he described it, your electronic soul is the collection of your personal electronic data that you might want to grant or revoke access to. You might grant access in return for better ads, special offers or in order to access a service; a government might 'tax' that information to plan public services. However, if you find that data is being misused, you might want to withdraw it. Hans Peter argued that designing systems to address privacy issues needs an asset model for both depositing and withdrawing information. Such a model would need appropriate controls, as well as transparency about where your data is 'deposited' and for what purpose. The electronic soul is an asset with value attached to it, and should be seen as such.

Haakon commented that it's the trade-off that's the interesting thing. What are you willing to give up if you get something of value in return? If you're walking down the street and get an ad for cheaper pizza, you might feel that they know where you are, but the trade-off wasn't worth it for you. People are already choosing to give up a lot of data where they can see some benefit to it.

There is a growing urgency to this issue – consumers and governments are becoming more wary and more demanding, as the problems of Big Data become more apparent. Public awareness of the issues of data control and transparency has been heightened by government open data initiatives, the need to navigate Facebook's complex privacy controls, as well as high-profile lapses in data security. A vacuum cannot be allowed to develop here, and those who currently hold large volumes of data – including the companies represented on the panel – have the most to lose if consumer pressure leads to restrictive legislation.

The panel's view was that users are most likely to keep companies honest, and will be a bigger influence than government in protecting their data. The job is too big for government to do. Current practices are being built from the ground up, starting in a small way. Government's tendency is to start in a large way – how do we do it for everyone – which makes it harder to attack this problem. You're starting to see the emergence of personal data management solutions through the private sector: Microsoft tried Passport, Google has Open

Authentication, Facebook is probably the most innovative at the moment with Facebook Connect.

There's a real disconnect there. The electronic soul is an incredibly valuable asset, but there is no model for how to represent it. There is no complete model that effectively controls ownership, allows you to see who has access and helps you get it back from people who have it. This is a complex subject, and needs a framework that doesn't currently exist. There's a great opportunity to create one.

"It's like Model T days – it's so early" Megan Smith

Digital Media literacy

If an ownership model for personal data, the electronic soul, is to take hold, consumers will need to inform themselves about the meaning of the decisions they make and how to operate the controls. Facebook has discovered that it is not necessarily enough to make the controls available – you also need to publicise and educate users on how to operate them.

Megan Smith highlighted work done by The Aspen Institute to identify digital media literacy as essential for democracy and civic engagement. The Knight Commission recognised that successful participation in the digital age entails two kinds of skills sets – digital literacy and media literacy:

> *"Digital literacy means learning how to work the information and communication technologies in a networked environment, as well as understanding the social, cultural and ethical issues that go along with the use of these technologies. Media literacy is the ability to access, analyse, evaluate, create, reflect upon, and act with the information products that media disseminate."*

Megan's view was that this digital literacy needs to be based within the education system. People need to learn about the data they have and how to control it, as well as learning to understand data, visualisations and the pitfalls of statistics. More education is needed but also more excitement. Megan cited Hans Rosling as someone who has done a huge amount to excite people about data, and how to use and understand it.

For more about this event, visit the NESTA website to watch the video, access a list of links and videos on the resources page and find out more about upcoming events.