University of Hull
Hull University Business School

Research Memorandum 94
July 2014

Creating business value from Big Data and business analytics: organizational, managerial and human resource implications

Prof Richard Vidgen
Centre for Systems Studies
Hull University Business School

Creating business value from Big Data and business analytics: organizational, managerial and human resource implications

Richard Vidgen

Hull University Business School (HUBS)

Email: r.vidgen@hull.ac.uk

Executive summary

This paper reports on a research project, funded by the EPSRC's NEMODE (New Economic Models in the Digital Economy, Network+) programme, explores how organizations create value from their increasingly Big Data and the challenges they face in doing so. Three case studies are reported of large organizations with a formal business analytics group and data volumes that can be considered to be 'big'. The case organizations are MobCo, a mobile telecoms operator, MediaCo, a television broadcaster, and CityTrans, a provider of transport services to a major city. Analysis of the cases is structured around a framework in which data and value creation are mediated by the organization's business analytics capability. This capability is then studied through a sociotechnical lens of organization/management, process, people, and technology.

From the cases twenty key findings are identified. In the area of data and value creation these are: 1. Ensure data quality, 2. Build trust and permissions platforms, 3. Provide adequate anonymization, 4. Share value with data originators, 5. Create value through data partnerships, 6. Create public as well as private value, 7. Monitor and plan for changes in legislation and regulation.

In organization and management: 8. Build a corporate analytics strategy, 9. Plan for organizational and cultural change, 10. Build deep domain knowledge, 11. Structure the analytics team carefully, 12. Partner with academic institutions, 13. Create an ethics approval process, 14. Make analytics projects agile, 15. Explore and exploit in analytics projects.

In technology: 16. Use visualization as story-telling, 17. Be agnostic about technology while the landscape is uncertain (i.e., maintain a focus on value).

In people and tools: 18. Data scientist personal attributes (curious, problem-focused), 19. Data scientist as 'bricoleur', 20. Data scientist acquisition and retention through challenging work.

With regards to what organizations should do if they want to create value from their data the paper further proposes: a model of the analytics eco-system that places the business analytics function in a broad organizational context; and a process model for analytics implementation together with a six-stage maturity model.

1. Introduction

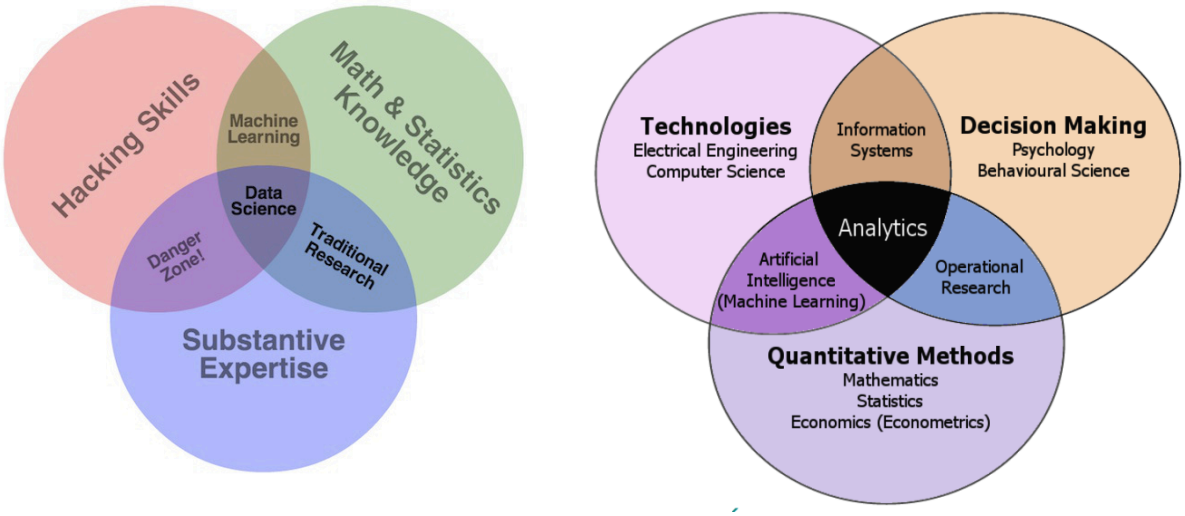
Big data and data science was everywhere in 2013 and the hype and over-inflated expectations have continued to build. According to Davenport and Patil (2012), being a data scientist is the "Sexiest Job of the 21st Century". The Economist says that data scientists are the "New Rock Stars", that they will be in short supply (Swantee, 2013) – McKinsey (2011) is already forecasting the US will face a shortage of up to 190,000 data scientists by 2018 – and that they won't come cheap. There's a distinct sense that this is all new and exciting and that "machine learning algorithms were just invented last week and data was never "big" until Google came along" (O'Neill and Schutt 2014, p. 2). Of course, data science has been around for decades and in the case of statistical techniques for centuries. For example, the Reverend Thomas Bayes (1701 – 1761) devised his eponymous theorem to include the updating of beliefs using prior probabilities, a method that is used in many machine-learning applications to perform classification (e.g., the probability that this email is spam, that this woman has breast cancer).

Robinson (2014) charts the stages of data science development, identifying the eras of 'management information systems' (mid 1960s to mid 1970s), 'decision support' (mid 1970s to late 1980s), 'business intelligence' (late 1980s to mid 2000s) and 'business analytics' (mid 2000s onwards). Davenport and Harris (2007) provide a succinct and widely adopted definition: "By *analytics* we mean the extensive use of data, statistical and quantitative analysis, explanatory and predictive models, and fact-based management to drive decisions and actions." (p. 7).

However, analytics is not simply business as usual for operational researchers, modellers, and statisticians: "… data science is not merely statistics, because when statisticians finish theorizing the perfect model, few could read a tab-delimited file into R if their job depended on it." (Mike Driscoll, quoted by O'Neill and Schutt, 2014, p. 7). Or, put another way, Josh Wills' definition: "data scientist (noun): Person who is better at statistics than any software engineer and better at software engineering than any statistician". Drew Conway's (2010) data science Venn diagram captures the blend of skills needed by a data scientist nicely (Figure 1a) – a blend of hacking, maths and stats, and substantive expertise. The last reflects the data scientist's role in asking questions, creating hypotheses that can be tested using statistical methods, i.e., "questions first, then data" (ibid.). This suggests data scientists need to acquire domain knowledge and work with subject area specialists to work out useful and interesting questions and to formulate explanations rather than rely on brute force analysis by machines. This is in distinct opposition to Anderson's (2008) claim that "We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot."

Robinson (2014), Figure 1b, comes at analytics from an operational research background and focuses on decision-making, technology and quantitative methods (maths/stats/econometrics). Although Robinson's model seeks to make sense of 'analytics' as a field, rather than the attributes of the data scientist, there are clear overlaps and parallels between the two Venn diagrams. Thus, based on Conway's diagram, supported by practitioner and media reports, we expect firms to be looking for data scientists that have (1) a sound maths and stats capability, (2) enough IT capability to be able to code up models and solutions, access and transform data,

and interact with data technologies as and when needed, and (3) the ability to work with business and domain specialists to ask relevant and useful questions that lead to the formulation of testable hypotheses. Robinson's model is more applicable to organizations, highlighting the central role of (1) improved decision-making supported by (2) a maths/stats capability and (3) technology.



(a) (b)

Figure 1: (a) the data scientist (Conway, 2011); (b) business analytics (Robinson, 2014)

This research aims to dig deeper into the role of the data scientist, how they might create value through the business analytics function, and the organizational implications of becoming data intensive. To investigate how organizations can create value from their data we adopt a framework used by Nerur et al., (2005), who analyzed the organizational change implications of the migration from traditional software development to agile software development. The model looks at the implications of change through four dimensions: (1) organization and management, (2) people, (3) process, and (4) technology. This model has a long and distinguished provenance in socio-technical systems and Leavitt's (1965) diamond model of organization. We use this framework to study the change implications of becoming a data-driven organization. The business analytics capability of the organization can be thought of as a mediator between the data the organization generates and accesses (internal and external) and the value the organization can leverage from that data through better decisions (Figure 2). Thus the research addresses the following questions:

1. How can organizations create value from [Big] data?
2. What dimensions do organizations need to address in building a business analytics capability?
3. What are the organizational change implications of building and exploiting a business analytics capability?

Research method

To explore the role of the data scientist and the business analytics function we undertake qualitative case study. This research design is best suited when a contextual understanding of an existing reality is desired (Yin 2009). Further, it allows gaining deeper and richer insights into emergent phenomena (Willis et al. 2007). Thus the aim of the research is to gain a deeper understanding of the role of the data scientist and to gain insight into the change implications for firms that seek to create value from their data.

Data were collected over four months from November 2013 to February 2014 via interviews with the manager responsible for business analytics at three organizations. The data collection was guided by an interview guide (Appendix A) that contained open-ended questions to encourage interviewees to share their opinions and experiences with us (Yin 2009), but also to allow the researchers to further explore emergent themes. The interview guide was used as a structure and aide memoire for the interview rather than as a rigid template. Each interview lasted between 50 to 100 minutes and was electronically recorded. All interviews were subsequently professionally transcribed.

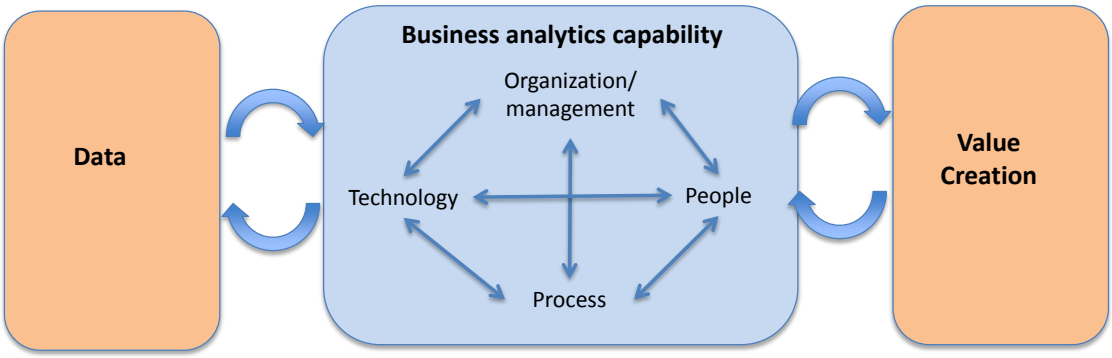


Figure 2: the research framework (adapted from Leavitt 1965, Nerur et al. 2005)

To analyze the data, we utilized Strauss and Corbin's (1998) open coding and axial coding techniques. Hence, we seek to identify codes and categories on data science not purely from the data, but rather based on the dimensions in Figure 2. During open coding, we first deconceptualized data by breaking it into smaller units that were repeatedly compared, categorized and reexamined based on the dimensions of Figure 2. During axial coding, we then reconcepualized the data in new ways that enabled connections between categories to emerge, that is, the different categories were assembled into higher-order themes to give meaning to the role of the data scientist and value creation (Strauss and Corbin 1998). During the entire data analysis process, we followed guidelines by Miles and Huberman (1994) regarding evaluation criteria of qualitative research (e.g., authenticity, plausibility, and transferability).

1.1. Cases

Case organizations were selected on the basis that the organization is large (i.e., has the potential to generate big data) and that it has an established business analytics function. Our three cases are, pseudonymously, MobCo (a mobile

telecoms company), MediaCo (a broadcaster), and CityTrans (a transport provider for a large UK city).

2. Findings

2.1. MobCo

Data

MobCo is a large international mobile telecoms provider whose primary focus is selling airtime to consumers. MobCo has a substantial share of the UK market and revenues in the billions. The mobile phone network creates a vast amount of data associated with mobile phone usage; data that can be used in many ways by MobCo to support network operations, billing, and customer service. The network also allows the location of users to be tracked: by triangulating from mobile phone base stations the user of a device can be identified to an accuracy of 50 meters.

Value creation

MobCo consider that total revenues could be increased by 20% through exploitation of its data, with 65% coming from internal opportunities and 35% from external opportunities. However, this is an aspiration rather than a reality today. Currently, analytics is concentrated in small islands, such as customer management and marketing (e.g., gaining insight into the customer experience) and network operations (e.g., identifying black spots and coverage holes):

> *"But once you got outside those sort of two islands, the rest of the company had actually very, very low analytical ambitions." (Head of Analytics)*

More broadly, MobCo see significant opportunities for analytics as a basis for competitive advantage:

> *"And then, the one that's the hardest to quantify, but I think it's potentially the biggest, is the businesses that use this data more effectively are going to be able to compete by offering products at lower prices. And so the consumers will pay lower prices. If you have two companies and one is very data-centric and knows exactly how it can cost, I mean, these customer satisfactions, they're going to have a lower price offering. And they're going to drop their prices because they'll get more market share." (Head of Analytics)*

The external opportunities fall into three major categories: (1) data licensing, (2) data products, and (3) public service offerings. Data sharing involves making data extracts available to third parties, such as the market research company Nielsen, who augment the data sets with additional data sources and then provide the data to their customers with MobCo getting a share of the revenue in return. MobCo can also use these augmented data source themselves to support internal planning, such as marketing. The second area of external activity is data products, typically based on location awareness:

> *"And the number of use cases that you can apply that to is really, it's almost unlimited. You'd be hard-pressed to think of an industry that wouldn't benefit in some way from*

understanding where their customers were, or what their movements were." (Head of Analytics)

Mobile phones are particularly appealing for location-based services because:

"It's stronger than the chip and pin because people bring their phones with them everywhere. If they leave the phone at home, they'll go and get it. And if they lose it, they'll report it stolen instantly and so it's actually a very, very effective mechanism for identifying fraud." (Head of Analytics)

Fraud protection solutions can be provided to credit card companies:

"So you check into a hotel, give them your card, the merchant will dial up to Visa and say, "Here is this person's card," and Visa will do, right now, will do a fraud check. If they know that you're standing right in front of that merchant, that will eliminate almost all fraud." (Head of Analytics)

Further opportunities include insurance companies, who could, for example, check that a driver was actually at the scene of an accident, and whether the passengers claiming whiplash damages were actually in the car. For life assurance purposes the insurance company could verify whether an applicant does indeed go to the gym every evening and how often they visit hospital, to create a profile that would allow the insurance premium to be better tailored to the individual's lifestyle.

The commercial applications of location-aware technology are many and varied. The mobile phone network also enables public service value creation, for example to provide warnings to the general public:

"You can send them a text message. We want to tell everybody that's within an area that's at risk from flood damage or if there's been a chemical spill, or a terrorist alert … Or, even doing tsunami warnings with everybody that's within a mile of shore." (Head of Analytics)

Or to assist the emergency services:

"We can identify witnesses to specific crimes like child abduction. We can run it very, very quickly, say, show us everybody that was in the area of where the child was last seen … And then, where have they moved to? And are there any other people that were in that area now and are they still there?" (Head of Analytics)

However, it is only fair to say that these novel and innovative value propositions for the most part have yet to be implemented.

Business analytics function

Organization/management

The business analytics function in MobCo was originally a joint venture established with the aim of exploiting the huge data asset generated by MobCo's customers and networks. After some months MobCo decided this capability was too important to be left outside as a joint venture and MobCo bought out their partner and created a centralized analytics function servicing MobCo and its international operations. The business analytics team at MobCo comprises around 30 people, directly and indirectly, of whom a small number are core.

Process

The core members of the analytics team are business analysts – MobCo does not currently employ data scientists directly, choosing instead to outsource this role together with the related IT support:

> *"Most of our data resource, our data scientist resource –we outsource on a case-by-case basis as required" (Head of Analytics)*

The analytics process includes checks to ensure that any use of data is legal and meets regulatory requirements. With regard to the ethical dimension of data use, while MobCo don't have a defined process they do have a heuristic:

> *"Yeah, we do have a saying internally that we want to be spooky but not creepy, spooky but not creepy. Now, the difference between spooky and creepy is, it's very, very thin. And spooky's sort of, you know, "Ooh, how do they do that?" and creepy is sort of, "Ugh, how do they do that?" And that's a fine line." (Head of Analytics)*

The fine line between 'spooky' and 'creepy' points to the data scientist, the team, and the organization needing to be aware of ethical considerations:

> *"There is a huge ethical dimension to this. If you are in any way inclined towards Orwellian type of schools of thought then this could be the worst thing ever. On the other hand, you know, a lot of data scientists look at the benefits side that this type of data can create, and it seems like the biggest opportunity ever. So yeah, there is, it's a lot of responsibility and that's imbued to the work teams, we're all responsible in this." (Head of Analytics)*

People

From a people perspective, data scientist skills by themselves are not enough (and can be bought in as and when needed). Data skills have to be complemented with business knowledge, problem-solving ability, story-telling, and communication skills. For example, in terms of exploiting the value of data:

> *"The hard part is constructing a business opportunity. And that's what, really, the challenge is. And people that have enough of a background in statistics … but also, are deep enough into the business side to be able to bring the two together, to define the problem we're trying to solve, or the opportunity. And that, right there, is what's missing, that is the biggest problem from my perspective." (Head of Analytics)*

Making sense of data and being able to communicate that insight effectively is a key part of the data scientist's role:

> *"I think that, you know, being able to look at masses of data and tell a story, which is really how I describe to people in the business world the data scientist's role" (Head of Analytics)*

And the data scientist needs a problem-solving approach that takes account of the application domain and context:

> *"And I talk to lots of data scientists and they're brilliant people, but they're, it's just that context that's missing. And it's like, it's a classic, you know, academic business type of trade-off. But the ones that are most effective, and I think that ultimately will*

command the biggest price, are not just the pure scientists but those who can identify a problem and then create a process to solve it." (Head of Analytics)

Technology

At their current stage of development, MobCo don't see technology as an issue, neither enabling nor impeding their data science activity:

"The technology has never been anything that's standing in our way. The technology, you can go to any one of five or six different vendors to get what we require. That's not the hard part." (Head of Analytics)

Challenges

MobCo are seeking clarity on their business analytics strategy and recognize that this will involve a significant level of cultural change:

"but it is a cultural recognition that being data-centric as a company is a way to more effectively compete" (Head of Analytics)

And in becoming data-centric the head of analytics must be a champion within the organization:

"Most of my effort is not spent externally, it's spent internally, trying to act as a champion for the benefits of becoming more data-centric as a company, and trying to garner support for creating [an] integrated analytic strategy" (Head of Analytics)

Coupled with the need for a clear strategic direction for business analytics and the associated culture change is the ability to innovate. The hesitance to invest internally in business analytics and data science may be related to past failures to apply innovation:

"We're a very slow-moving corporate behemoth, and we don't do innovation well. If you're familiar with any of our past attempts, you know, they've been business school case failures" (Head of Analytics)

In using consumer data MobCo face challenges of informed consent:

"The hardest part about all these types of models, and I would say this is both internally and externally, is being able to get the consent, you know, the architect for a consent strategy to be able to access this data." (Head of Analytics)

MobCo want to develop the idea of informed consent through a customer dashboard that will allow the value of the data collected by MobCo to be shared with its customers:

"We're actually going a step further and, we are launching a new permissions platform and if you are a MobCo customer you will see your personal details … you can log on, enter your credentials, and you will see all the data that we hold about you. That data will be field-editable, so you can go in there and you can delete it, or you can make it more specific. And then, depending on who you are, let's say that you have a resemblance to a demographic character, it's just who you are, you will see a series of use cases, that we would like to use your data in. And within those use cases there will be a series of incentives and you can go through and grant and revoke a percentage of these use cases, according to your comfort level." (Head of Analytics)

Summary

The findings from the MobCo case are sumarised in Table 1, structured around the research framework presented in Figure 1.

Dimension	Finding
Data	• Customer profiles/demographic data • Data generated by mobile phone users (customers) interacting with the mobile phone network • Location of user can be tracked
Value creation	• Internal - improved network operations enabling better data and services to be offered to existing mobile phone users and to attract new users • External - creation of data products based on mobile phone usage and location awareness (e.g., anti-credit card fraud, location-based marketing) • Substantial revenue opportunities for analytics - potential to uplift revenue by 20% (35% external, 65% internal) • Potential for public service offerings (e.g., flood warning by text message)
Organization/ management	• Established as a joint venture then brought in-house as a global analytics function • Emphasis on external projects rather than supporting internal decision-making and development • Islands of analytics expertise in MobCo, centred on customer management and network management
Process	• Internally the analytics function is focused on the core areas of network and customer management with new projects taken on ad hoc as needs are identified • External projects are developed as opportunities arise • Data science and IT skills are acquired “on demand” to resource projects
People	• Technical data science skills are sourced externally on an as needed basis • Business analysts need business knowledge, a problem-solving aptitude, and strong communication and story-telling skills
Technology	• Agnostic – the most appropriate technology is used for each project
Challenges	• Lack of buy-in from MobCo business functions – low analytical ambition outside of customer management and network management • A strategy for business analytics needs to be developed and communicated • Cultural change will be required if the analytics strategy is to succeed • Organization takes a cautious approach to innovation • Requires robust data use policy (e.g., seeking user permission to use data and sharing the value created)

Table 1: MobCo summary

2.2. MediaCo

Data

MediaCo is a broadcaster (television company) whose revenue is primarily gained from selling advertising. MediaCo delivers content through the digital terrestrial network and via the Internet as an ‘on demand’ service. Through the Internet service MediaCo can capture details of users’ viewing habits, which allows MediaCo to place appropriate adverts and to promote content to its viewers through recommender applications. MediaCo is also engaged in promoting societal change and thus helping users discover content that helps achieve more than simple entertainment is also an important part of its mission. Users wishing to access content that is more than 30 days old must register with the MediaCo Web site, allowing details of online behavior to be attributed to identifiable individuals:

> *"Once they've registered with us across all of our different platforms that you can watch [the content], we then capture who you are from a registration perspective, and then all of your behaviour after that. So everything you watch, whether you pause the content and then re-view it, how you navigate through our websites, all of the digital information that you're tracking, so we capture all that information." (Head of Analytics)*

MediaCo use Adobe's Omniture for Web analytics, allowing them to "track every single mouse click, every single navigation" (Head of Analytics). External data sources include matching viewers to Experian credit-ratings (with varying degrees of success) and weather data, "because weather has a high correlation to ratings and viewings" (Head of Analytics). Although not currently used, it would also be possible to match viewers with their social media data.

MediaCo store their data in the Cloud, suitably anonymized:

> *"I think we've put enough checks in place such that there's nothing personally identifiable in the cloud. What happens is we, once you register with us we create a globally unique identifier, so it's like a 12 digit number or something. So everything gets anonymized, then from that point forward that identifier and all of your behavioural data is all put in the cloud … you know, there's nothing that's identifiable down to the individual, at best what you're going to get is just viewing behaviour." (Head of Analytics)*

Value creation

A major use case for data analytics revenue generation is advertising sales, in which online viewing behavior is used to predict the social class and demographics of the user:

> *"if you're a media buyer today you can buy certain audiences on linear TV, and very different audiences on video demand. And with predictive product that basically predicts someone's age, gender, socioeconomic class, housewife and housewife with children status, based on all of that viewing behaviour." (Head of Analytics)*

This intimate insight into the user:

> *"then allows us to sell a video on demand advertising product across all those different things by age, gender, socioeconomic class, so you can buy for example an ABC1 1634 housewife with children as a segment. And it's all done using predictive analytics." (Head of Analytics)*

The quality of the predictions is audited and once established allows MediaCo to achieve a price premium for its advertising:

> *"So we've built a predictive model, that model was audited by PWC as being 80% plus accurate in terms of when we say you're an ABC1 1634 housewife with children, how many times are we getting that right? And 15% of all of our video on demand inventory will be traded on that product, at a price premium of about 20% to 30% above our existing price premium for the product that you currently can trade on." (Head of Analytics)*

MediaCo also use advertising to market and promote its own content, where other shows are promoted at the end of the show being viewed:

"So because we've built the model at all those different levels, it now allows marketing for example to promote certain content to certain users very different in video on demand. And some of our early trials have shown about twice as much, two times effectiveness in terms of when we're promoting a show and getting a viewer who then sees the promotion to then go on and watch that content" (Head of Analytics)

As well as placing advertising and promoting its own content MediaCo has a remit to take the viewer on a "voyage of discovery", to challenge conventional thinking, to raise public awareness of issues, i.e., to view things that would not be picked up by a recommendation engine and "introduce some noise into the model":

"it isn't about just because I watch comedy you're not introducing me to more and more and more comedy, you're helping me to discovery something in factual perhaps that I would never even consider. And I may not enjoy it but it perhaps will evoke some kind of reaction." (Head of Analytics)

In terms of increasing future value, MediaCo sees this in further internal developments such as assisting commissioning of content and scheduling activities rather than in external activity.

Business analytics function

Organization/management

The Data Planning and Analytics group was established in 2011 following the appointment of a new CEO who recognized that a change in user behavior was taking place as users migrated from fixed times on terrestrial digital channels to on demand viewing via the Internet on smartphones and tablets:

"he was fundamentally convinced that data was going to transform all aspects of broadcast, and all elements of broadcast. And part of that was because he saw a big change in viewing behaviour from a broadcast type relationship to mass intermediation happening … [the] consumer starting to watch more on their mobiles and tablets, they have a natural return of data. So he was absolutely convinced that from a strategic perspective data enabling the organization was going to be key.

This strategic orientation to data led to the establishment of the analytics function that goes beyond the traditional analytics areas of marketing and customer relationship management (CRM):

"So as a result of that the new capability was set up that reports directly into the CEO. The reason for that was it was important that data wasn't just going to be used just for marketing or CRM, which is a very classical way in which you could use data. It was going to look at all aspects of our business, and look at it horizontally first, and then start to look at okay, what can analytics and what can these things do?" (Head of Analytics)

To implement this vision MediaCo has a co-located team:

"The team at the moment is 15, and that's split out across three that are more business consultants and business facing, so those are the viewer insight managers is what I call them. And the rest are all data scientists, a combination of data scientists and analysts, that's what the rest of the team are." (Head of Analytics)

The structure of the team is data scientists and business analysts, the former focusing on data science methods and the latter on working with the business to define and shape problems and opportunities.

Process

The team works with the business to help define problems and solutions. It also finds new opportunities through exploration of the data:

> *"I think what's happened is the data scientist has been looking, understood the business a bit more and sort of said actually, you know, why aren't we looking at this sort of view of our viewers … I think that's what accelerated adoption of the business and people asking more and more questions, which then thought actually there is something here because if you can do this in scale it might change the way you might think about scheduling or the way in which you might think about what content you might want to commission." (Head of Analytics)*

A key part of the process of exploring the data with the business is visualization methods, which can highlight patterns in the data and open up new conversations:

> *"So for example our scientist, one of our data scientists has built out there just a visualisation using some gravity modelling, just a visualisation of content clusters. And that's, that was developed just to showcase visualisation as a concept, but as a result of that it's now having, it's been touted around the business and it's inspiring quite interesting conversations." (Head of Analytics)*

And visualization can help with adoption of analytics by the business:

> *"And I think because it was quite visual and quite easy for people to interact with it, I think that's what accelerated adoption of the business" (Head of Analytics)*

While projects use standard techniques, such as regression, CHAID, support vector machines, and Markov chains, because the data is unstructured and changing it's not simply a case of taking out the data and running it through a model. Rather, MediaCo need to use machine-learning techniques to understand and learn about new content through supervised learning:

> *"And content in itself isn't, it's not like a consumer packaged goods company where you can describe the packaging in a certain way, or the product in a certain way in terms of the UPC codes [universal product code], there isn't a standardised way to describe content. So because of that you need machine learning type techniques, because it needs to understand and learn about this new content, and then understand and re-tweak the model." (Head of Analytics)*

People

MediaCo want data scientists who are curious about the world:

> *"… a lot of my analysts certainly will describe how they were just fundamentally curious around how the world is structured, or curious as to why, you know, patterns emerge the way they emerge. So it wasn't about the vocation necessarily itself, but it was an element of curiosity. And that curiosity is what you want in an analyst." (Head of Analytics)*

As well as being curious and problem-focused the data scientist needs to be able to work independently with initiative:

> *"... the other thing that we find as well is the ability to think independently, so we have interviewed a lot of the people who we then, who have done brilliantly academically and performed terribly in the work place" (Head of Analytics)*

With regards to skills and tools, MediaCo reports a disconnect from the world of business intelligence, data warehouses, and enterprise data analysis tools such as SAS and SPSS with a world of open source tools used on an ad hoc and problem-centred (rather than tool-centred) basis:

> *"And what we found was those analysts were very good at that tool, but when you asked them around what was emerging on big data, when you asked them around open source and some of these things, they were very dismissive of some of the other technologies ... they sort of lacked the ability to kind of look beyond kind of what SAS could do, what SPSS could do." (Head of Analytics)*

It further seems that experienced and senior analytics people may have become too far removed from the hands-on methods and business of analytics; it is even possible that many of these people don't have the technical data science skills that are needed when moving from intelligence to predictive analytics:

> *"at the more senior end of the analyst cycle, is they almost became glorified resource managers ... they were so far removed from the analytics that they were more project managers administrating teams of analysts, and they couldn't, when I've asked them how might you do a clustering or why would you use clustering versus regression, they just genuinely didn't know." (Head of Analytics)*

MediaCo then forged alliances with Universities and began developing their own resources:

> *"So we then formulated partnerships with London universities, we fund a fully funded PhD programme that runs for five years, and we fund some Masters programmes as well." (Head of Analytics)*

MediaCo look for a combination of statistics, maths, and programming skills in their data scientists. An applied statistics degree may mean a data scientist is strong on techniques but not able to program, in which case learning to program statistics in the R programming language would help them understand the basics of programming and object oriented design. On the other hand, it may be more difficult to start from a programming background:

> *"Computer science probably less so, because I think that's, it's harder I think to teach statistics and maths to a computer scientist." (Head of Analytics)*

MediaCo don't see business domain knowledge as essential – experience of the media industry can be picked up "on the job". Indeed data scientists can be too narrow in terms of industry and techniques:

> *"You then look at sort of investment banking and actuarial science, brilliant people there but they almost have again quite a niche narrow view of the sort of techniques they're deploying, same thing with hedge funds." (Head of Analytics)*

Depending on the industry, domain-specific knowledge may or may not be a requirement to working as a data scientist. Regardless, business awareness is essential in terms of understanding how data is linked to outcomes

> *"So I think the business school brings to life a lot of things around, you know, decomposing an organization to its component parts to say actually, you know, could value be created from, you know, from the value chain, and at what part? Or is it about driving your intel efficiencies, or is it about how might we go about developing a new consumer facing proposition? So I think all of those sort of macro questions, I think there's still value in that." (Head of Analytics)*

Business schools can also help with managing the transition to becoming a data-driven organization:

> *"Where they may play a role for big data is something around, you know, data enabled, fact based decisioning, and what does that look like? So in a lot of organizations they talk a lot about, you know, and it may be part of a change programme, so for example customer centric change or data driven organizations as part of a change agenda" (Head of Analytics)*

The career path for data scientists is either a technical one in which the scientist moves up a "sort of analytics data science hierarchy" or becomes more business facing and act in a role that is "a more traditional business consultant".

Technology

MediaCo found that traditional relational technologies would not cope with the volumes of data:

> *"So we looked at building our own internally, we looked at cloud era technology, we looked at a whole host of technologies, including traditional technology like Oracle and that sort of thing. With the traditional ones, so the relational databases, we found that for the volumes we were dealing with, for the unstructured nature of web data, it couldn't cope with it. You could eventually load it, it would take something like nine months to load. Whereas with elastic cloud computing we loaded the same amount of data in about two days that would have taken us nine months to load." (Head of Analytics)*

The IT operations are based around the use of Amazon Elastic Compute Cloud, with limited in house support. This gives MediaCo flexibility in the scaling of its operations:

> *"at the moment what we're saying to the business is we need two things; we need flexibility of infrastructure, so there's some analysts will run, some modelling we'll run that just needs a significant amount of computational power and the ability to kind of almost infinitely scale wide, so having 10,000 servers for example for a particular query or a particular model. Now once that's done we want the flexibility then to shut them all down and go back to one." (Head of Analytics)*

Within the Amazon cloud environment the MediaCo analysts make use of open source tools such as MapReduce for computations running on a Hadoop platform, and Mahout for machine learning. The analysts use R as their programming interface to the data tools, rather than proprietary software:

> *"We started with SPSS initially, but we found it wasn't flexible enough and didn't enable machine learning in the cloud, and that's what we needed." (Head of Analytics)*

Challenges

One challenge for the analytics function is the ability of the rest of the organization to keep pace with opportunities identified by the data scientists, i.e., for the organization to be responsive and agile:

> *"So what is the business outcome you want, understanding the organizational agility, so how quickly can the organization respond to something that you've identified on the inside? And if you have a very slow organization … it begs the question why are you doing the analytics … for example how quickly can we launch a new programme from a piece of insight, how quickly could we change the schedule?" (Head of Analytics)*

There can be further resistance to the techniques of data science and particularly the use of control groups:

> *"…control is an interesting one because in marketing there's always the challenge around I don't want to have a control route because, you know, I want to talk to all the customers, I don't want to have a control." (Head of Analytics)*

The issue of personally identifiable information (PII) arises in many aspects of analytics – what data can be used, for what purposes, how will value generated from the data be shared between the consumer and the organization, and what are the brand implications of using PII. A robust data usage policy is in place to reassure consumers that these activities will not be detrimental to them. Being careful about the uses consumer data is put to is an ethical and business issue. In making an assessment of whether it is acceptable and wise to use viewer data MediaCo look at it from the viewer's perspective:

> *"The way we're doing it though is so we have what's called a viewer promise … it's a promise to our viewers around what we're doing with data, how we're capturing data, and what the benefit is. So it's based on two principles around transparency and control, and what we say is if we can't explain to the viewer what we're doing with the data and why in a way that makes sense to the viewer, we won't do it." (Head of Analytics)*

This means establishing a compelling proposition for the viewer:

> *"I think what you need to do is define a very compelling consumer proposition that sits on top as to why they would want to link, you know, my bank details with my retail details with my viewing details." (Head of Analytics)*

In terms of external opportunities, MediaCo are wary of being involved in selling consumer data or sharing it with third parties. If value is to be created from consumer data then the question of how that value is shared arises:

> *"… the other concern we have with that is that it goes back to the viewer promise in that one of our principles is we'll never sell their data, because we feel that from an ethics perspective it isn't right, and therefore shouldn't the consumer also enjoy some sort of benefit. However I know there's sectors that have emerged that try to be data aggregators and these things. I think fundamentally the problem with that model is there isn't a natural trusted third party that consumers will put all their data in." (Head of Analytics)*

The issue of PII is further complicated by potential developments in online regulation: the right not to be tracked and the right to be forgotten, both of which

have implications for how data is collected, stored, and used. Looking further ahead MediaCo can imagine a world in which PII need not be held (thus avoiding issues of tracking ad being forgotten), where predictive analytics are good enough to make decisions based on each viewing session in real-time:

> *"So we would love to get to a point where we saw no PII at all, and we're running all of that in real time … and then be able to be in a position for the first time, we say but we track nothing, we don't do any history, we don't track any history within session"*

And not holding PII might help protect and promote the brand as a point of differentiation with competitors:

> *"I think the other thing is, it's highly advantageous as a brand to be able to go to market and say look, we no longer storing any of your information, whereas you know, a lot of the others, you're going to keep your history, and there's all these other concerns, I think that could be quite, quite an interesting competitive advantage for us." (Head of Analytics)*

From a people and HR perspective, holding on to data scientists is likely to be challenge for MediaCo in the near future:

> *"In the next year and a half I expect churn just to rocket, and that's because there's going to be a lot of other companies now that will probably double their salaries … we can't compete with investment banking, we can't compete with, you know, some of the larger players." (Head of Analytics)*

Dimension	Finding
Data	• User viewing habits tracked online with detailed Web analytics • Potential for links to external data (e.g., credit-ratings, social media, weather)
Value creation	• Advertising revenues • Marketing and promotion of content • Social benefit through education re promotion of content novel to viewer
Organization/ management	• Established as a strategic initiative • Sourced in-house, organized as data scientists, business analysts, and IT services
Process	• Exploitation through a focus on the core area of advertising revenue generation • Opportunities identified through ad hoc exploration supported by visualization techniques • Unstructured and changing data requires a flexible approach utilizing machine-learning
People	• Data scientists need to be curious, problem-oriented, and capable of independent working • Data scientists need strong skills in statistics, maths, and programming • Open source skills (e.g., mapreduce, R) are valued over enterprise software (e.g., SAS, SPSS) • Prior domain knowledge is not essential and can be learnt on the job • Business awareness and knowledge are needed by business analysts to identify value creation opportunities and to manage change
Technology	• Amazon cloud and related open source technologies (e.g., Hadoop, Mahout, R)
Challenges	• Lack of a culture of data-centric thinking throughout the organization, e.g., embracing data science methods such as control groups • Personally identifiable information (PII): permissions, use, and value sharing with consumers • Impact of potential introduction of legislation such as "do not track" and "right to be forgotten" • Brand protection and enhancement • Talent retention

Table 2: MediaCo summary

MediaCo are working with their HR team to identify ways in which they can retain their staff:

> *"And one of the biggest things we can offer is variety of problems. So most analytic teams, if you go and talk to someone at O2 or Vodafone in an insight team, what you'll find is the analyst is a CRM analyst." (Head of Analytics)*

However, churn is not necessarily bad:

> *"... it's healthy for some churn to happen as well, because I think you then can bring in some external perspective, so let's bring somebody in from, I don't know, airline. We have no one from the airline industry in my team, what would that bring?" (Head of Analytics)*

Summary

The findings from the MediaCo case are sumarised in Table 2, structured around the research framework presented in Figure 1.

2.3. CityTrans

Data

CityTrans is a governmental, not-for-profit provider of an integrated public transport system for a major city, dealing with every aspect of how people move across the city using different modes of public and private transport. The organization works with many data sets, such as network operations, travel data, traffic data, loadweigh data, infra-red data, and customer data. Some of these data sets have been linked together for operational analysis and planning but there are still a number of unexplored opportunities in joining up these numerous and diverse datasets. CityTrans collects the bulk of its public transport travel data through a smart travel card (STC), which can be used anonymously or as a registered (and therefore identifiable) customer. Any one passenger may have a number of anonymous, pay as you go or low-value season ticket STCs, making identification of an individual difficult. To further protect individuals' identities CityTrans' data scientists and research partners use psuedonymised ("hashed") IDs for their analyses of customer behaviour.

The STC generates millions of 'taps' a day as customers use it on different modes of transport throughout the city. However, this data typically tells CityTrans where a passenger entered the system and where they left it – it is more difficult to determine what route they took through the network. CityTrans also accept contactless payments from credit and debit cards on its bus network and has plans to roll out contactless payment to the rest of the public transport network in 2014.

CityTrans has been active in data analysis and modelling of passenger flow for many years, with analysis grounded in traditional operational research methods. Making use of data has become more prominent since the introduction of the STC and as storage capacity has increased and new software tools have become available:

> *"STC data has long been used for customer service purposes, which was the primary reason to collect it in the first place. We started to analyse our STC data for transport*

planning and for operational analysis starting in 2005. In 2011 we took the strategic decision to devote significant senior management focus to exploring how we could use our data even further to answer strategic questions about our customers and operations to inform planning and decision-making. At that point, I became the head of a full-time team focussing on how we use our data to deliver value for the business and our customers." (Head of Analytics, Customer Experience Directorate)

In the past CityTrans had to rely on surveys of individual users, which involved making small amounts of data work hard in complex planning models:

"the challenge in those years, was that we didn't have granular, comprehensive data. We had to, therefore, make many assumptions based on these small-sample surveys We then did very complex and sophisticated modelling based on what limited information we had, and our very large strategic models would have to make simplifying assumptions to predict customer behaviour." (Head of Analytics)

CityTrans are now making use of large datasets produced by their operational transport systems:

"Now we have a lot of the data that reflects how people are travelling on the network. There's a real opportunity to turn that into detailed and insightful analysis of what's going on. We are now using our data intensively, but there's obviously a transition path as we worked across the organization to demonstrate our new data tools and the potential to use them to answer questions with new data sets." (Head of Analytics)

Value creation

The data collected from users of the transport system and the system operations are used to measure the reliability of the service and to gain insight into the customer experience. Rather than focus on average journey times as a proxy for customer satisfaction, CityTrans are now able to use the STC data to get closer to the customer experience and to use the data to minimise travel delays:

"We used to only be able to measure the reliability of our train services for our customers by solely looking at the data from our train systems. And that gives you good understanding about how your trains are performing but it doesn't always correlate to how customers are actually experiencing the services. So we're supplementing the train service data with customer data based on STC journey times, and looking at the distribution of journey times rather than using an average journey time metric which can mask some of the distributions" (Head of Analytics)

When there is disruption to the network, a measure of central tendency, such as the average or the median is a poor representation of the impact on passengers:

"We know that the travel times for the people caught at the worst point of our disruption will be very poor. But if we're only looking at the average over the course of the entire day, we'll be capturing both the disrupted and non-disrupted journeys. So if we look just at daily averages, we wouldn't highlight the impact of disruption for those unlucky people who were caught in the worst of it" (Head of Analytics)

Data is automatically collected from the STC system, which provides a vast amount of journeys to be analysed each day. This reduces the need to conduct expensive surveys to ask passengers to report their journey details. While this is straightforward for the City's train network, where there is a tap in and tap out associated with each journey, it is more difficult for the bus network, where customers only tap on when entering a bus and do not tap on exit. CityTrans has an

algorithm—based on looking at the customer's next tap—that uses the journey taps alongside its bus location data to infer where passengers exited the bus. This provides passenger bus loadings at a route level and thus informs the design of routes and frequency of service. Based on this detailed data the transport system design can be further nuanced:

> *"Using our algorithm, we look at how our bus routes are grouped at a stop level. We can then amend our groupings of routes based on where we see passengers interchanging between buses. This is a benefit for our customers, as it minimises the walk time required between bus changes.*

CityTrans are also investigating the use of data for behavioural change, to shift travel patterns so as to spread the load across the network:

> *"We've also been looking whether there is an opportunity to provide customers with better information about the busiest times and places on our network, in order to encourage some customers to shift their travel times slightly. If we can shift even some customers from the most crowded times and places, all customers could have a better journey experience. Using our data, as well as surveys to understand our customers' motivations, we are assessing whether this might work. We could provide some information to tell our customers that if they leave right now, they would be travelling at what is generally the busiest time of the peak." (Head of Analytics)*

CityTrans envisage changing traveller behavior using 'recommender' systems that will assist the traveller in planning their journey:

> *"We could provide some information to tell our customers that if they leave right now, they would be travelling at what is generally the busiest time of the peak. We could explain that if they were able to shift their journeys slightly—say fifteen minutes one way or the other, they would have a better journey." (Head of Analytics)*

Business analytics function

Organization/management

A single team of 20 provides the core ticketing analytical services and is divided into three areas: a small strategy team who work with the business to establish what data is needed to answer key strategic business questions, manage stakeholders' requests for analysis and analytic tools, and drives the future data strategy; an operational research and analytics team whose members concentrate on the maths/stats and data crunching and providing interpretation of the data; and an IT and operations team (currently the largest team given the development programme underway) who provide the technology solutions, such as the new data warehouse, develop reports, and provide operational support for the data, systems and tools.

The strategy team is responsible for promoting the use of data across the organization, to help the business identify strategic questions that data can answer and to translate a data need into analysis or reporting tools for one-off insight or for self-service reporting. The operational research analysts (data scientists) work both on designing the methodology for answering questions brought to them by the strategy team, while also having time for investigative analysis. The data scientists need a good understanding of the business "because the ticketing systems is so complex and the different data sets are so complex and the network is complex". The data scientists use statistical methods including t-tests, and cluster analysis, as

well as data mining techniques to extract data. They also provide interpretation of the analysis so that it can be understood by both technical and non-technical audiences.

In order to enable even better use of STC data, CityTrans has established research partnerships with several universities, to explore their data and apply new analytic techniques, such as visualization. The relationship with MIT has proved particularly effective:

> *"They have a very effective model, that demonstrates the deep understanding of the transport network by the students and faculty. The faculty spend a significant amount of time working with [CityTrans] to define suitable research questions, and through constant review of students' work, provide assurance and quality control that the research output is robust. Several of the tools that MIT students have provided as part of their dissertations have been adopted and industrialised by [CityTrans]." (Head of Analytics)*

CityTrans also have data sharing agreements with a number of other universities, allowing access to anonymised STC data to researchers with the provision that they do not publish card-level data. This provides researchers the opportunities, for more arms-length investigations.

Process

The decision to develop a new data warehouse has led CityTrans to look at its processes for extracting, transforming and loading data, and in report development to align them with best practice for business intelligence systems development. A modified Agile Software approach is used for development.

CityTrans has used the opportunity of the new data warehouse to ensure that its data quality and data governance processes reflect best-practice guidance. Data stewards have been identified from across the business to ensure that reference data is kept up to date. Data profiling tools will also be established to highlight when numbers are outside the bounds of what is expected. This is useful for both identifying unexpected trends in travel behaviour, or highlighting if there were a systems issue, such as a power or communications failure, where data might be held at the devices rather than reported back to the ticketing system. Where possible, data profiling in the new data warehouse will done using rules and applied automatically to accomplish data cleansing. Data cleaning, transformation, and extraction takes up the vast majority of effort using the legacy tools. With a reengineered data warehouse and improved extraction routines CityTrans have an aspiration that this will be reduced substantially.

People

CityTrans are looking for statistical expertise and familiarity within SPSS, SAS, R, and the ability to join datasets up using SQL. Analysts must be able to make sense of data, be able to write it up and "understand what your data's telling you, what it isn't telling you, and the underlying biases". Analysts have to be able to convey their findings to technical and non-technical audiences. While the current data scientists were developed internally, CityTrans also looks for data scientists externally. CityTrans finds that "because we have such exciting data to work with, we are able

to attract a lot of interest." Data scientists are retained by providing them with an opportunity to work on interesting questions. From a career progression perspective, some of the data scientists will look to move into data science management, while others will prefer to focus in on the complex questions that data science can solve, rather than take a managerial or strategic role.

CityTrans have a demand for analysts who interpret the data and provide the analysis of the output that is insightful and potentially provides a basis for action:

> *"Your data scientists need to be able to craft a compelling story. If data is presented without interpretation it's just a table of numbers. You don't understand the context, it's not compelling and it's not useful"*
>
> *"You need to have the story about what does this mean for your organization and what action decision-makers should take. Your data scientist needs to take the complicated maths and explain the conclusions in such a way that someone who is not a data analyst can understand it. In some ways, that may be the hardest skill for the data scientist" (Head of Analytics)*

Technology

Up until this year, CityTrans relied on extracting data from the operational systems using a legacy SAP Business Objects reporting stack, which resulted in a "very slow, very cumbersome" process. The operational systems were designed for revenue collection and for customer service—CityTrans' intensity of data science was not envisioned at the time the ticketing system was built.

A new data warehouse is currently being deployed that will dramatically change the way that data is structured, and will offer a vast improvement in terms of speed and flexibility and supportability of the reports that the Analytics team provides to the business. It will also enable a much greater level of autonomy for basic queries so that non-data scientists can self-serve for a variety of questions.

The migration of the existing reports, starting with identifying which ones are used and are valid, and then developing them to best-practice methodology has taken up much of the development team's time. There is also a very large work programme to support future developments in the ticketing system, including the ongoing rollout of contactless bank cards on the network.

Challenges

The Customer Analytics team at CityTrans has started by incorporating the STC data and Contactless bankcard data alongside bus location data in the data warehouse. The team has on its strategic roadmap the challenging of joining up a large number of databases, which requires a deep business and technical understanding of how the data is structured. It also requires that new operational system tools, when designed, build in the capability to record the granular data so that it can be transferred to the data warehouse in a timely manner. This principle has been embedded into the project management development processes for new infrastructure investment. There is a challenge in terms of retrofitting this requirement into some of the legacy systems.

Dimension	Finding
Data	• There is a wide range of datasets at the transit organization, being used for a variety of operational and planning purposes. Analysis of customer travel data on the public transport network is predominantly done through analysis of the ticketing system (smart travel cards (STCs) and contactless payment cards (CPCs) • Surveys to gather further information about intent to travel and about route choice within the public transport network are also undertaken from time-to-time. Historically, before STC data was available, surveys were the only way to understand journey behavior. • Additionally, vehicle loadweigh data and infrared counters are used on parts of the network where ticketing data is not complete • Data is stored at the disaggregate level, with some (anonymised) personal information, although registration is not a requirement for travel on a STC • The agency also has access to other customer data sets outside the public transport realm, and uses them intensively for analysis. At the moment these are only partially integrated into the customer analytics platform
Value creation	• Improvements to reliability and quality of service • Insights into the specific customer experience (not an averaged out experience) • Replacement of expensive qualitative surveys by automated travel analysis • Potential to initiate behavioural change in passengers and spread the network load
Organization/management	• CityTrans has a strong background in operational research • Structured around strategic analysts, data scientists, and IT development and operations • Partnerships with universities used for research and development and for exploitation of the data resource and experimentation with new technologies
Process	• Creation of compelling and actionable insight and analysis • Fuller exploration of data, historically, was constrained by difficulty of extracting information as well as the need to have analysts and data scientists provide routine reporting and support for operational systems • Current development of new technology platform and an increase in team capacity have begun to expand the analysis capability
People	• Data scientists required to have expertise in statistics, the ability to differentiate signal from noise, and to be sensitive to biases in the data analysis • Data scientist must be able to ask the right questions • Technical skills in SPSS, R, SQL, and data discovery tools are required • Data scientists are attracted by the opportunity to work with interesting data and retained through being given interesting problems to work on • Data scientists must be able to provide interpretation of their results rather than just prepare raw output. • Development teams must have expertise in developing BI systems and reporting.
Technology	• Historically, dominated by legacy systems and databases that were slow and cumbersome • A new data warehouse has been constructed, which has substantially reduced manual effort and data extraction, transformation, and load times
Challenges	• The technical complexity in joining up a large number of disparate datasets and ensuring that new operational systems are designed to capture, record, and provide data feeds that can be used for analysis. • The need to understand underlying business requirements and a deep business understanding of the subject area. This is required to ensure that the right questions are being posed, the right data is being used, and that the interpretation is accurate and useful for the business. This deep understanding is also required before embarking upon changes to systems and processes in order to create buy-in from colleagues for adopting new data tools. • There is an ever-increasing appetite for new analysis—a challenge and an opportunity

Table 3: CityTrans summary

The CityTrans team has had to develop considerable technical understanding of underlying business requirements, in order to deliver analysis and reporting to business stakeholders that meet requirements for CityTrans. Further, a deep

understanding of the business subject area is needed. This is required to ensure that the right questions are being posed, the right data is being used, and that the interpretation is accurate and useful for the business. This deep understanding is also required before embarking upon changes to systems and processes in order to create buy-in from colleagues for adopting new data tools. This is particularly the case when the adoption of the new tools means that business processes will change to adapt to the new technology:

> *"We've worked closely with our internal stakeholders to understand their business needs and to make sure that there is support for new data tools. This has been a very successful partnership, but is worth highlighting as a potential challenge for organizations looking to change the way that data tools are used in organizations." (Head of Analytics)*

Lastly, CityTrans has seen a dramatic uptake in the appetite for analytics:

> *"With this ever-increasing appetite for new analysis, we need to respond quickly to requests and do so with the highest quality, so that key business questions are answered. Meeting the analysis needs of the business to support data-driven decision-making is a challenge that we're pleased to accept." (Head of Analytics)*

Summary

The findings from the CityTrans case are summarized in Table 3 and structured around the research framework presented in Figure 1.

3. Discussion

From the summaries of the cases twenty recommendations are identified, which, following the research model presented in Figure 2, are grouped into: (1) data and value, (2) analytics organization/management and process, (3) technology, and (4) people (data scientists) and tools. The key recommendations are summarized in Table 4.

3.1. Data and value

Data and value are inextricably interwoven and mediated through the business analytics capability of the organization (Figure 2). Data of sufficient quality, i.e., that is fit for purpose, will be essential and intrinsic dimensions of data quality such as accuracy; timeliness; consistency; and completeness is essential (Wang and Strong, 1996; Strong et al., 1997). For organizations with large investments in legacy systems (e.g., ticketing for CityTrans and billing for MobCo) structured data will need to be extracted and transformed and, while these legacy systems may prove cumbersome and inflexible, they contain valuable data and the quality of that data will count (Haug and Arlbjørn, 2011). As data is increasingly collected from the Internet of Things (IoT) (e.g., smartphones providing location data automatically and tracking software catching every mouse movement) issues of data quality – in all its dimensions - will still need to be addressed, and will vary from use case to use case, e.g., what constitutes timeliness for management reporting might not be the same as timeliness for real-time decision-making.

Data and value	
1. Ensure data quality	Data must be 'fit for purpose', including legacy data
2. Build permissions platforms	Organizations will develop customer self-serve permissions portals. Assurance of trust is paramount – organizations must be transparent about how data is used and generate trust that it is secure
3. Apply anonymization	Establish confidence in the data anonymization process before data is shared
4. Share value	Value created from data may need to be shared with the data originator
5. Build data partnerships	Value is likely to arise from data partnerships rather than selling data as a commodity to third parties
6. Create public and private value	The data managed by the organizations can be used for public and societal benefit as well as commercial benefit (e.g., flood warnings)
7. Legislation and regulation	Changes in legislation may result in fundamental shifts in what can be done with customer data (for example a "right to be forgotten")
Organization/management and process	
8. Corporate analytics strategy	An analytics strategy is needed with a clear articulation of how and where value will be created
9. Organizational change	Becoming a data-driven organization will involve organizational and cultural change and innovation
10. Deep domain knowledge	The business analytics function will need to build deep understanding of the organization and its business domain if it is to create lasting value
11. Team structure	The business analytics team requires a mix of data scientists, business analysts, and IT specialists
12. Academic partnering	Data science expertise and resource can be acquired through partnering with Universities
13. Ethics process	Ethics committees should be established to provide oversight of how data is used and to protect the reputations and brands of organizations
14. Agility	The agile practices of software development can be adopted and modified to provide a process model for analytics projects
15. Explore and exploit	Analytics teams should exploit in response to identified problems (80%) and have slack resource to explore new opportunities (20%)
Technology	
16. Visualization as story-telling	Visualization of data is not simply a technical feature – it is part of the story-telling
17. Technologies	While technology is in a state of flux an agnostic approach is advisable
People and tools	
18. Data scientist personal attributes	The data scientist must be curious, problem-focused, able to work independently, and capable of co-creating and communicating stories to the business that form the basis for actionable insight into data
19. Data scientist as 'bricoleur'	The tools and techniques don't matter as much as the ability of the data scientist to cobble together solutions using the tools at hand ('bricolage')
20. Acquisition and retention	Data scientists are attracted by interesting data to work with and retained if they are given interesting problems to work on and have career paths

Table 4: Key findings

In all of the cases respondents highlighted concerns about how customer data is used and the need to be transparent and to seek customer permission. While it might be an aspiration not to hold personally identifiable information (PII) (e.g. MediaCo), where PII is held then a self-serve permission platform for customers to manage their data usage permissions may be required. Such a platform would allow customers and other data owners to opt in and opt out. It is also likely that customers will expect to be incentivized or compensated for the use of their data in ways that go beyond simply receiving a better service, i.e., the value created from data may need to be shared with the customer. The cases also suggested that selling customer data would be a low value activity and one that could harm the image of the organization. Rather, organizations are more likely to enter into value creating partnerships (e.g., using location based services to reduce credit card theft). When data is shared with research partners it has to be anonymized and this may need to go beyond simply stripping out personal identifiers such as names and addresses if the risk of reidentification is to be managed effectively (Ohm, 2010).

While CityTrans' mission is to provide an integrated transport system for the passengers and travellers in the city, the public service agenda applied also to the two commercial organizations. MobCo could envisage using its location-based services to warn of floods and tsunamis, while MediaCo wanted to actively influence viewing habits from an educational and social awareness agenda. Organizations also need to be aware of potential changes to legislation, such as the 'right to be forgotten', which may require them to not only change the data they retain and how they use it but could have significant impacts on their data-driven strategies and business models.

3.2. Organization/management and process

Organizations need to articulate a clear data strategy and be clear about how value will be created from data, whether that be financial (e.g., increased revenue, decreased costs), intangible (e.g., increased customer satisfaction), or societal (e.g., Tsunami warnings). Implementing the data strategy will require organizational change and innovation, for which senior management support will be essential. Rather than relying on a central analytics function to provide a service to the business, the business itself will use data and evidence-based management in all areas of decision-making as managers and others become more data literate and self-reliant. The cultural and organizational change associated with the move to becoming data-driven will likely encounter resistance and will take years rather than months (Adler and Shenhar, 1990). The business analytics function will need to develop deep subject matter expertise and knowledge of the many and complex datasets, systems, and business operations.

The structure of the analytics teams was similar in all three organizations, comprising data scientists, business analysts, and IT services (Figure 3). The data scientists require strong data skills with basic IT skills in programming (e.g., R), enterprise analysis software (e.g., SAS), data manipulation (e.g., SQL), and visualization. The business analysts need deeper domain knowledge and a focus on creating business value; they work with the business to understand requirements and with data scientists to shape solutions. To turn data science prototypes into production applications and data products requires the stronger development skills of IT professionals. The team structure in Figure 3 can be augmented with

partnerships with research institutions that can provide expertise and resource to tackle projects that might not otherwise be viable, allowing the organization to experiment and explore their data in new ways.

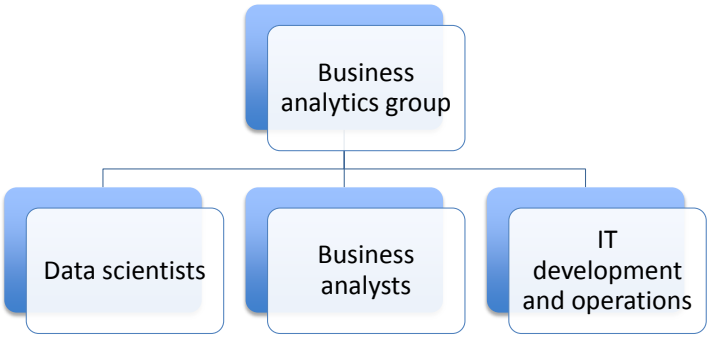


Figure 3: Typical business analytics structure

Undoubtedly, how the data collected by the organization is secured and how it is used will have an impact on the trust of customers and perceptions of the brand image of the organization. While some things may be legal and acceptable within the regulatory framework they may not be in accord with the values of the organization and the image that it wants to project (as MobCo commented, “spooky” may be acceptable but “creepy” is not). An ethics committee would consider requests to use data for commercial purposes and assess whether it is legal, whether it is in accord with the values of the organization, and the potential risks (e.g., to the brand value).

Data science projects can learn much from agile software development, which has a proven and tested process model for delivering software that creates business value through iterative delivery rather than the stepwise definition and execution of a specification. Many of the values and practices from agile software development can be adopted in data science projects. For example, agile methods, such as eXtreme Programming (XP) and Scrum, emphasize engagement with the customer (or subject matter expert), frequent and delivery of working software (or data science solutions), co-location of resources (e.g., subject matter expert, business analyst, and data scientist), learning through rotation of roles (e.g., data scientists can learn new techniques from other data scientists), and establishing a culture of professional excellence (Vidgen and Wang, 2009). As well as being agile, a truly effective data science team will explore as well as exploit; while the bulk of an analyst’s time, say 80%, is spent working with data to solve defined business problems, the remaining time, say 20%, should be retained as slack resource for experimentation and exploration – such as searching for new patterns in the data, trying new tools, and learning new techniques.

3.3. Technology

Visualization and data interaction are more than technical ways of presenting data to the business– they are an integral part of the communication process in which non-technical business people can see and understand how data can help

improve decisions and business operations. As for the underlying big data technologies, these are currently in a state of flux and will take some time to shake out and making bets on which technologies will win out is a risky proposition. While not all organizations will be able to be agnostic about the technology, as MobCo is, these decision should be made on the basis of data requirements and the value that can be created from that data rather than fashion.

3.4. People and tools

The overriding message from the cases is that organizations want data scientists who are curious, problem-solving, capable of independent working, and able to work with the business to co-create plausible and convincing stories through data that lead ultimately to actionable insights.

While the data scientist undoubtedly needs strong statistical and mathematical skills they also need IT skills, notably an ability to program (e.g., R) and an ability to manipulate data (e.g., SQL). While open source solutions may be in the ascendance when compared with enterprise offerings such as SPSS and SAS, it is not really about the toolset. Rather than rely on one tool, whether it be an enterprise product such as SAS or an open source product such as R, the data scientist needs to be able to use the most appropriate tools to hand, to combine different technologies, toolsets, and analytic techniques to fashion a local and relevant solution. Thus, the data scientist is more 'bricoleur' than engineer.

There will be intense competition for data scientists that are technically competent and able to create innovative and practical solutions to business problems through data analytics. Firms that have interesting data will have an edge in attracting good data scientists; firms that let their data scientists work on interesting problems and build a strong culture of data science professionalism will retain their data scientists. Some level of movement of data scientists will be positive as experience in one sector is applied to another. Data scientists will need career paths; the cases suggest that these will encompass paths to becoming a senior data scientist, becoming a manager of data scientists, becoming a business analyst/strategist, or moving to work in the business as an analyst or manager.

4. Implications

The transition to a data-driven organization in which business decisions are routinely made on the basis of data and hypothesis testing will require deep changes to processes, management, structure, and ultimately to the culture of the organization. In Figure 4 we present the business analytics framework of Figure 2 as an eco-system.

The elements of the eco-system are distinct but inter-twined with reciprocal relationships. While all the elements are interrelated, only those relationships that are particularly pertinent to business analytics are shown. For example, the ICT strategy will constrain what can be done with data and the analytics strategy. At the same time, the data held and the analytics strategy will influence the ICT strategy. The analytics eco-system illustrates the mutual co-dependencies that are of particular interest when defining an analytics strategy. Drawing on the research

model of Figure 2 and the findings in Table 4, three triangles of influence are overlaid on the diagram: data, strategic, and tactical. The data triangle is concerned with the data assets, the ICT strategy and their relationship with the analytics strategy. The strategic triangle is concerned with how the analytics strategy creates value for the organization and how it enacts and shapes the business strategy. The tactical triangle is concerned with the analytics function and the ICT and HR strategies, reflecting the technology and people dimensions of the research model. Thus, the three triangles reflect data management, value creation, and the analytics function.

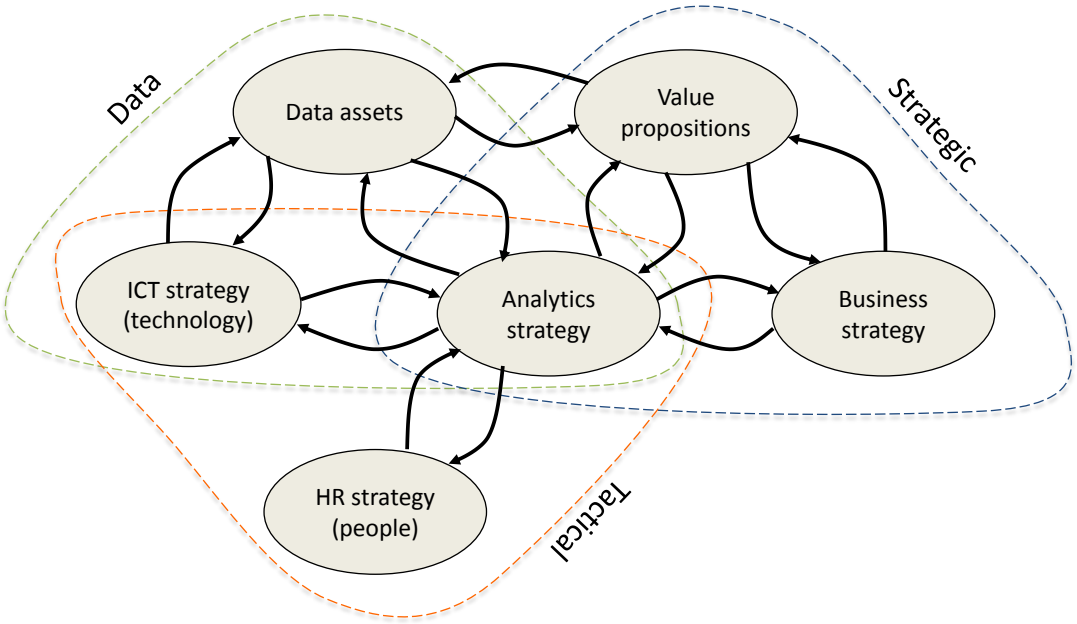


Figure 4: Business analytics eco-system

The eco-system represents the part of the system and their inter-relationships – it highlights the interdependencies and connectedness of business analytics across the organization. As such, it does not give a guide to what action to take. In Figure 5 a process model is presented, which shows the steps that might be taken to achieve an effective analytics capability.

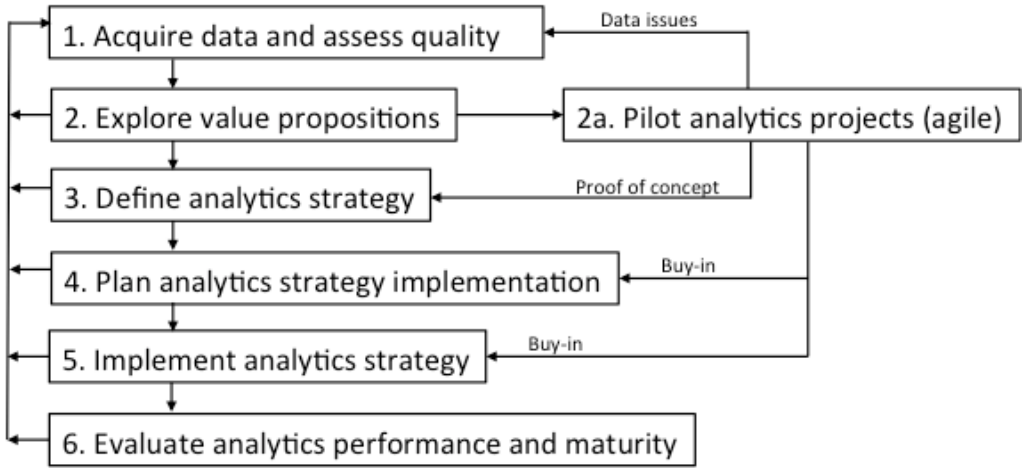


Figure 5: Business analytics implementation model

The model represents an iterative and ongoing cycle of data acquisition, value exploration, strategy, and execution. Step 2a is intended to avoid the pitfalls of the grand plan – small pilot projects grounded in agile methods are used to gain credibility, to learn about analytics and methods, to identify data issues, and to provide proof of concept for the strategy. Further, the pilot projects raise the profile of the analytics function and help to gain buy-in from the various stakeholders who need to be engaged if the strategy is to work.

In Step 6 the performance of the analytics programme is evaluated. The evaluation at level 6 should include benchmarking the organization against its competitors and against best practice together with an assessment of the analytics maturity achieved. The iterative nature of the process model suggests that higher maturity stages will be reached over time as the organization adjusts to the change in structure and culture that are needed to become data-driven, i.e., the journey to analytics excellence is likely to be long and winding. Following stage models (e.g., Venkatraman, 1994) relevant levels of maturity for analytics might range through:

1. *Fragmented*: the organization makes ad hoc use of analytics within individual departments, such as marketing;

2. *Localized*: the organization begins to exploit medium sized data sets and starts to integrate data from multiple functions, e.g., supply chain, marketing, and sales;

3. *Functional*: establishment of a central analytics service as a part of the formal organizational structure; an organization-wide emphasis on improving revenue and margins and on improving operations;

4. *Data-driven*: decisions throughout the organization are based on data and evidence; management know to ask about the provenance of data and its quality and know how to interpret the results of analytics;

5. *Evidence-based*: data and decisions are aligned with corporate policy and strategy, fully integrated into business processes, and supported by methods such as randomized controlled experiments as the basis for informed action;

6. *Essential*: management and decision-makers focus on exceptions/dilemmas, ethical considerations; the analytics strategy plays a greater role in shaping the business strategy and becomes the essence of how the organization competes and survives in its environment.

At level 3 the organization can begin to follow the stage model of Figure 5 and to design their analytics strategy rather than relying solely on emergence. All three of the case organizations reported here are squarely placed at level 3 and are working to make a fuller transition to level 4. However, the migration from level 3 to level 4 is likely to be pivotal – it will require cultural and organizational change and may become a stumbling block for those organizations that fail to establish an effective analytics strategy, and a stagnation point for those organizations that feel sufficient value has been created through the establishment of a central analytics function.

5. Summary

The findings can be summarized as follows. First, data and value should be considered and managed jointly. To create value data has to be managed (e.g., anonymization, permissions, quality) and that value may have to be shared with those who provide the data. Second, business analytics is not a technical project to be given to the IT department – it is a business transformation initiative that requires a strategy, senior management support, and active and careful change management (Thorpe, 1998). This is not to say that IT is unimportant; it is a fundamental enabler of the business analytics process and essentially embedded in the organization's processes and practices (McAfee 2006).Thirdly, business analytics and data science needs a project process model, for which agile software development provides a strong starting point. Fourthly, the organization needs a change model to navigate the implementation of a business analytics strategy. Fifthly, data scientists need a strong sense of curiosity and problem-solving orientation and the ability to pull tools and techniques together using whatever is at hand, i.e., a process of bricolage.

While not wishing to marginalize analytics technologies and data science methods, this research demonstrates that there are many avenues for future research, including: value sharing models, regulatory impacts, societal benefits/disbenefits, assessment of business analytics maturity, business analytics and organizational change, business analytics project management, data quality, human resource development, and visualization in the context of effective story-telling.

6. Acknowledgments

This research was conducted with support from the EPSRC's NEMODE (New Economic Models in the Digital Economy, Network+) programme. Hasan Bhakshi, Juan Mateos-Garcia and Andrew Whitby of Nesta co-developed the interview protocol. Chris Thomson of HUBS and the OU assisted with the analysis of the data. John Morton of CPM Consulting provided invaluable feedback on drafts of the report and was a co-developer of the maturity framework. Regardless of all the help and support received any errors and inaccuracies of interpretation are entirely the responsibility of the author.

7. References

Adler, P. S., & Shenhar, A. (1990). Adapting your technological base: The organizational challenge. *Sloan Management Review*, (32)1: 25–37.

Anderson, C. (2008). *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. http://www.wired.com/science/discoveries/magazine/16-07/pb_theory (accessed 17 March 2014).

Conway, D. (2010). *The data science Venn diagram*. https://s3.amazonaws.com/aws.drewconway.com/viz/venn_diagram/data_science.html (accessed 17 March 2014).

Davenport, T. H. and D.J. Patil (2012). Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*. http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ (accessed 17 March 2014).

Haug, A., Arlbjørn, J.S., (2011). Barriers to master data quality. *Journal of Enterprise Information Management*, 24(3): 288-303.

Leavitt, H.J. (1965). Applying organizational change in industry: Structural, technological and humanistic approaches. *Handbook of Organizations*, J.G. March, Ed. Rand McNally, Chicago, IL.

McAfee, A., (2006). Mastering the Three Worlds of Information Technology. *Harvard Business Review*, November, 141 - 149.

McKinsey (2011). *Big data: The next frontier for innovation, competition, and productivity*. Available from: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation (accessed 17 March 2014).

Miles, M.B., and Huberman, M. (1994). *Qualitative Data Analysis: An Expanded Sourcebook*, (2 ed.) Sage, Thousand Oaks.

Nerur, S., Mahapatra, R., and Mangalaraj, G. (2005). Challenges of migrating to agile methodologies. *Communications of the ACM*, 48:5, pp. 72-78.

Ohm, P. (2010). Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization. *UCLA Law Review*, 57(6): 1701-1777.

Robinson, S., (2014). *How do you solve a problem like analytics?* Presentation to the OR Society, 29 January 2014.

Strauss, A., and Corbin, J. (1998). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage, Thousand Oaks, CA.

Strong, R., Y. Lee and R. Wang, (1997). Data Quality in Context. *Communications of the ACM*, 40(5): 103-110.

Swantee, O. (2013). Data scientists are the new rock stars as big data demands big talent. *The Economist*. http://eefutureconnections.economist.com/data-scientists-

new-rock-stars-big-data-demands-big-talent/?fsrc=scn/outbrainUK (accessed 17 March 2014).

Thorp, J., (1998). *The Information Paradox*. McGraw-Hill, Montreal.

Venkatraman,N., (1994). IT-Enabled Business Transformation: From Automation to Business Scope Redefinition. *Sloan Management Review*, 35(2): 73 – 87.

Vidgen, R., and Wang, X. (2009). Coevolving systems and the organization of agile software development. *Information Systems Research*, September, 20(3): 355-376.

Vinekar, V., Slinkman, C.W., and Nerur, S. (2006). Can Agile and Traditional Systems Development Approaches Coexist? An Ambidextrous View. *Information Systems Management*, 23:3, pp. 31-42.

Willis, J.W., Jost, M., and Nilakanta, R. (2007). *Foundations of Qualitative Research: Interpretive and Critical Approaches Sage*, Thousand Oaks, CA.

Yin, R. (2009). *Case Study Research*, (4th ed.) Sage, Thousand Oaks, CA.

Appendix A: Interview guide - head of analytics/big data

1 Introduction	
1a	Introductions – focus of our research – working definitions (big data workers)
1b	Would you tell us about what your organization does?
1c	Would you tell us about what you do and your background?

2 Big data context	
2a	What types of data are held by your organization? Of this data, what would you consider to be "big"?
2b	Who uses the data?
2c	Who owns the data (manages add/change/delete)?
2d	1. What big data technologies are used in your organization? a. distributed batch processing e.g. Hadoop b. real time processing e.g. Storm c. NoSQL databases e.g. HBase, MongoDB 2. Who first introduced these tools to the organization (from within IT, from another function)? 3. Who uses these tools? 4. Are these tools managed locally or in the cloud (e.g. Amazon AWS)

3 Big data value creation	
3a	How do big data workers create value in the company? How do they influence decision-making? What types of decisions?
3b	Is value created internally (e.g., to improve customer retention) or externally (e.g., to sell data products)?
3c	How is the business value of your big data evaluated (if at all)?

4 Where are the big data workers, how are they organised, and what do they do?	
4a	How many people with deep data skills work in your company?
4b	How are these people referred to in the company? What is their job role?
4c	In what part of the company do they work? Are they part of a cross-functional team or are they generally found operating within a single corporate function?
4d	Are they generalists or domain specialists?

4e	What types of activities are they involved with?
4f	What proportion of data analysis is exploratory (e.g., Google day)? How is this justified in business terms?
4g	What analytics techniques do they use? E.g., machine learning, recommendation systems, sentiment analysis, time series analysis, SNA, AI, simulations
4h	What desktop analytics and visualisation technologies do they use? E.g., Excel, SAS, Stata, SPSS, R, Python, Mahout, Tableau?
4i	Does the output from these tools feed management reporting? If so, who is the end consumer of the reporting from these outputs?

5 How are big data workers managed?	
5a	Could you describe the lifecycle of a 'data project' (e.g. a recent project)?
5b	What are the best ways to motivate data workers in your company? How do they differ, in this respect, from other workers?
5c	How self-directed is their work? Do they generate new questions or answer questions posed by others?
5d	What is the single thing your company has done that has made your data workers more productive?
5e	What is the main barrier to generating more value from them?

6 HR: Where do big data workers come from and where do they go?	
6a	What knowledge/skills/competences are you looking for when you hire them?
6b	Where do you go to look for them (from universities/from industry/elsewhere)?
6c	How long does it take it for them to be able to make a contribution to the business?
6d	What training and development is given to big data workers?
6e	What are the career progression opportunities for big data workers?
6f	If they leave the organization, where do they go and why?
6g	Are there any particular skill sets or competences in short supply in the market? Are these getting worse or better?
6h	What is the impact of these shortages (if they exist)?
6i	What are you doing to address these shortages?
6j	What skills and competencies should Universities be developing in its analytics graduates?

7 What are the barriers to big data use?	
7a	What regulatory or legal constraints are there? Do these inhibit what can be done with big data?
7b	What ethical issues are you concerned about (these might be legal but not something you would want to do)
7c	Going forward, what other barriers to using and creating value from big data do you see? E.g., organizational, managerial, technical

8 Concluding	
8a	What single piece of knowledge about data talent would make a difference for your company?
8b	Who else (inside or outside your organization) should we talk to?