nominettrust ›

# Big Data and social organisations

**A STATE OF THE ART REVIEW** written for Nominet Trust by Duncan Ross

This document is designed to be read with **Adobe Acrobat**

# Foreword

At Nominet Trust we seek to understand where digital technologies and the practices that can be developed with them can best be used to create new ways of generating social value. Making use of Big Data is a clear example of one such area: linking the transformational insights of data science with the creative and tenacious characteristics of social entrepreneurs offers incredible potential for finding new ways to address persistent social challenges.

If social organisations can realise the potential of Big Data then we can create new practices and interventions that offer radically different approaches to addressing some of the most persistent social challenges. Understanding what Big Data is; how it can be utilised, and the potential it offers social organisations is a key step in creating new insights from where such interventions and practices can be developed. It is our hope that this review does just this, and we look forward to working with you to make the most of this opportunity to redesign ways of addressing social challenges through the use of digital technology.

**Dan Sutch**
Head of Development Research
Nominet Trust - November 2013

# About the series

Nominet Trust State of the Art Reviews are undertaken by leading academics to collate and analyse the latest research at the intersection of digital technology and society. Drawing on national and international work, these reviews aim to share the latest research to inform the work of the Trust, those applying to the Trust and our wider partner organisations.

We value your comments and suggestions for how to act on the recommendations in these Reviews, and how we can build the series, so that it is more useful to us all as we explore how digital technology can be used to design radically new solutions to address specific social problems.

We look forward to your comments and suggestions at:

**developmentresearch@nominettrust.org.uk**

# About the author

**Duncan Ross** is Director Data Sciences, Teradata Ltd, a provider of analytical data systems, where he leads a small team of senior data scientists in a role combining demand generation and analytical innovation. He regularly speaks on the topics of Big Data, analytical innovation, and data philanthropy.

He is a founder and Director of DataKind UK, a charity that puts data scientists in touch with social organisations in order to promote the use of data for the public good. He is also a founder and director of the Society of Data Miners, a Fellow of the Institute of Direct Marketing, and a member of the Cabinet Office Open Data Users' Group. In the past he has started an award winning farmers' market, been Chair of Trustees of a national children's charity, and served as a City Councillor on Birmingham City Council.

## Acknowledgements

I would like to thank everyone who has helped me put this document together, special thanks to Nominet Trust for their support and encouragement, to my team of fantastic analysts at Teradata (Judy, Frank, Chris and Zunnoor), to the DataKind UK team (Kaitlin, Fran, Stewart, Jake and Jessie), to DataKind in the US, to those who kept me sane (Ivanhoe Runners, Tilla and xkcd), and to all the contributors, especially Ben Gilchrist, Policy and Participation Manager, Community and Voluntary Action Tameside (CVAT), Ian Carey, former CEO of Barnsley Hospice, Hannah Underwood, CEO, Keyfund, and HyeSook Chung, Executive Director of DC Action for Children.

Finally thanks to my family, Laura, Eve and Alys for putting up with long weekends and evenings.

# Contents

Written by

**Duncan Ross**
Director of Data
Sciences for Teradata

for Nominet Trust



Designed by **Ben Carruthers**

› www.nominettrust.org.uk

# Introduction



This is a document about Big Data and social organisations, although hopefully it will be useful to anyone who is interested in using Big Data and data science to improve society.

The term 'social organisation' will mean different things to different readers. Rather than getting lost in definitions though we will assume that it could mean a charity, a voluntary organisation, a community interest company, a not-for-profit, a non-governmental organisation, or even a governmental organisation or a commercial organisation – provided that the goal is to make the world a better place. I will use a variety of terms interchangeably for this – please don't be offended if yours isn't mentioned explicitly.

It is aimed at the enthusiastic Big Data amateur, who has some knowledge, a lot of interest, but isn't (yet) an expert. Inevitably it will touch on some technical areas, both of hardware, software and approaches.

In order to help define how Big Data can be useful to the third sector, this publication will:

- explain the key features of Big Data,
- look at where Big Data might be found,
- describe some of the key data issues facing users of Big Data,
- run through the technologies and concepts,
- and finally identify the key success factors needed to make Big Data work for your organisation.

I'm not going to provide code, or deep statistics – the language throughout will be relatively straightforward, although you are encouraged to go and research anything that catches your eye!

# Executive summary



Big Data may be the subject of hype, but it can and will provide social organisations with opportunities to improve and reshape their services. Commercial companies are already making the transformation to being data-led organisations – using insights and understandings from their data to transform their business.

Companies like Walmart have used data to revolutionise the way they do business. As Sam Walton, the founder of Walmart, put it: "People think we got big by putting big stores in small towns. Really we got big by replacing inventory with information". In the case of Walmart this means knowing every transaction from every store within minutes of it occurring, and being able to analyse this data and, most importantly, take action on it.

For online retailers such as Amazon, eBay and others, data not only helps them understand their customers, it actively shapes the websites that we see when we visit their pages. All this is done through analysis, experimentation, and data.

And Big Data is an important issue for social organisations today. It represents a combination of a series of trends: the rapid growth in data creation, the ability to store this data at a reasonable price, and the ability to apply sophisticated techniques to it in order to extract knowledge.

Big Data is often defined in terms of three Vs: Volume, Velocity, and Variety. Volume describes the amount of data that has become available, Velocity describes the rate of change of data, and Variety describes the range of data types that are now capable of analysis. But organisations should not allow the exact definition to prevent them from taking action.

'Data science' is a phrase that describes the techniques for turning raw Big Data into actionable insights. Data scientists are the people who undertake this work. For the third sector, data science is the way that charities can transform and grow their activities by analysing and using data. The raw material for data science is Big Data – data that because of its size, complexity, or the sophistication of analysis required, is outside the comfort zone of the charity concerned.

Social organisations can use Big Data and data science to improve their actions through analysis and prediction. Areas that provide immediate opportunities for this include, but are not limited to:

- fundraising,
- development,
- policy and influence,
- health,
- community involvement,
- children and young people.

Data for these analyses is available from a wide range of sources, and organisations should look both internally and externally to source the most appropriate data.

Organisations that have committed to using Big Data need to be aware of three key areas of concern – handling data, dealing with statistics and analysis, and legal and reputational issues. These issues are not necessarily technical in nature, but addressing them correctly will be relevant in reassuring trustees, funders and other stakeholders.

Big Data opens the doors for organisations to try new approaches. Social organisations should be willing to explore the opportunities that these provide, but using an agile 'fail fast' approach so that lessons can be quickly learned and absorbed into the organisation. Chapter 4 provides an insight into some of these new opportunities.

In order to exploit Big Data effectively, social organisations will need to have the right technical infrastructure in place. This will include storage and analysis technologies.

Big Data opens the doors for organisations to try new approaches. Social organisations should be willing to explore the opportunities that these provide

Organisations should investigate open source technologies such as Hadoop and R, but need to be aware of the potential long term costs. Technology should, above all, fit with the needs of the organisation.

The key elements to successfully implementing Big Data are putting data at the core of your organisation, focusing on data, identifying the right people and structures, and matching technology to the problem at hand.

When putting data at the core of your organisation you should consider:

- Are your organisational leaders able and committed to using Big Data to change actions?
- Are you willing to change your actions as a result of Big Data learnings?
- Is your organisation capable of measuring the results?
- Will the use of analytics worry or scare funders or clients, and can you mitigate that risk?
- Do you understand the business problem sufficiently from an analytical perspective?

In order to deal effectively with data you should consider:

- Can you effectively collect Big Data when (and over the timespans) that you require?
- Is your data well curated, or will you put in place ways of doing this?
- Do you have consent for use of personal data?
- Is the data quality appropriate for the actions you need to take?

To create the right structure and identify the right people you need to know:

- Do you have access to relevant data science experience?
- If not, can you recruit or develop it?
- Does your data science team cover the key roles of data science?
- Are procedures in place to retain data scientists and their knowledge?

To implement the most appropriate technology:

- Do you understand your future analytical needs?
- Does the technology match your user base needs?
- Will the technology be cost effective in the long term?
- Can you identify external resources that can help you with decisions?

Big Data is not a journey that you need to take by yourself. The final thing to remember is that Big Data is not a journey that you have to take by yourself. A number of organisations exist to support you in identifying best practices, and to link you to volunteer data scientists.

# Big Data today



It's hard these days to go far without seeing an article on data, usually with the word 'big' attached to it.

But is there any place for it in the third sector? Can charities use it? What could it mean? How could it be used effectively?

If you're reading this you probably work for, or with, a social organisation.
You may have been involved in finance, either fundraising or perhaps responsible for a project, in which case you might be thinking that you have a pretty good grip on your data. After all, your funders are increasingly looking to you to provide data as evidence of the impact of your work.

It's possible you have been designing services, or just working at the coal face with clients, and think that there isn't really a place for data in what you do.

Perhaps you're a data practitioner working in commerce or science, and you're wondering if what you do could have an alternative, direct benefit to society.

Or perhaps you're just an interested bystander trying to get a grip on this whole Big Data thing.

But whatever your position, one thing is clear, Big Data can add significantly to the effectiveness of data analysis, and hopefully to the impact of charities. As Hannah Underwood, Chief Executive Officer (CEO) of Keyfund put it, "data is the opportunity to move from prove to improve".

## Why the excitement?

There have been numerous attempts to define Big Data, mostly focusing on the type of data or the amount of data involved. We will look to a definition of Big Data in the next section, but in reality you can usually treat data as Big Data if it takes you into

areas that make you uncomfortable, either in terms of technology, skills, or ability to interpret. This is a very subjective decision – what is Big Data for you may not be Big Data for Walmart.

So why are commercial organisations – and, we hope, charities – making such a big deal about Big Data? Largely because it provides them with the opportunity to know things that were previously unknowable, and to move decisions from the world of hunches to the world of measurement.

In the world of commerce this has been known for some time. Walmart – possibly the largest retailer in the world – have revolutionised the way that their supply chain works through the use of data. They are centrally aware of every purchase made in every store within 10 minutes of it taking place. They use this information to inform their own purchasing decisions, being able to react to external events such as extreme weather conditions *before they are aware that they are happening*. They are also able to run experiments, a key element of the Big Data revolution, varying the price in one store and comparing their sales to the same item in a similar store elsewhere.

Amazon use Big Data to compare your purchases (and your browsing habits) to those of other people. They don't need to know who you are, just what you do. They use a sophisticated analytical technique called collaborative filtering to enable them to make recommendations based on this behaviour, and then they measure the results. Advanced analytical techniques and measurement – two more features of the Big Data revolution.

Closer to the world of charities, the health industry in the US is looking to use Big Data analysis to help people identify when they have pre-diabetes[1], and in the UK an organisation of doctors and health practitioners, NHS Hack Days, has been working to improve the NHS using similar techniques.

For Community Voluntary Action Tameside, an organisation that participated in a Big Data event, their Big Data was the output from a detailed survey of the hundreds of

Why are commercial organisations – and, we hope, charities – making such a big deal about Big Data? Largely because it provides them with the opportunity to know things that were previously unknowable, and to move decisions from the world of hunches to the world of measurement

1 www.nyu.edu/about/news-publications/news/2013/04/29/independence-blue-cross-nyu-nyu-langone-medical-center-collaborate-to-detect-early-diabetes.html

voluntary organisations in their area, and their insight was about the financial and organisational security of these organisations.

For Hampshire County Council's Special Educational Needs team their Big Data is the breadth of different data types needed to help predict future special educational needs.

In each of these cases the organisation concerned has identified a need and an ability to use new Big Data techniques as drivers for change.

The reason that Big Data is happening now, rather than ten years ago, is that we have a set of technologies that both makes data creation easier, and makes analysis easier. Storage solutions (such as Hadoop) have changed the game for organisations. Predictive techniques and toolsets such as R have grown in popularity. And most of these are open source software, making using them theoretically free.

The reason that Big Data is happening now, rather than ten years ago, is that we have a set of technologies that both makes data creation easier, and makes analysis easier

The most exciting thing is that all this is happening at the same time. It's not for nothing that this is often described as a 'Big Data tidal wave'.

## So why not?

We shouldn't pretend that Big Data is for everyone. There are some important considerations that organisations need to take into account before they spend time and money in this space.

As so often, these can be broken down into issues of people, process and technology.

- Does your organisation have the technological capability to take advantage of Big Data? We will see later on how the technological landscape has been changing with Big Data.

- Do you have the people, often referred to as data scientists, who can take advantage of Big Data? And do you have the senior commitment to push through the changes that might come out of your analysis? We will address this topic when we talk about key success factors.

- Do you have the processes to take advantage of the answers that Big Data can give you? Many organisations, especially in the third sector, have other considerations that may prevent them from taking action on Big Data analysis.

Hopefully by the end of this publication we will have identified both the way to overcome these issues, and the reasons why it makes sense to do so.

# What is ‘Big Data’?



Big Data is very much the description of a movement around the use of data rather than about delineation between ‘big’ and ‘little’ data. It involves using diverse data sources, some of which can indeed be very large, and using mathematical techniques to discover new information that enables you to change your actions. The term ‘big’ can sensibly be said to refer to the importance of data, rather than to any more formal definition.

Big Data in turn becomes the fuel for data science: the approach used to derive the new knowledge.

Nevertheless there have been many attempts to define Big Data, and these at least provide some patterns (and possibly some anti-patterns) that are useful in framing any discussion on the topic.

## The simplest definition

> “Big Data is one byte more than you are comfortable dealing with.”
> Simon Rogers, The Guardian

In his keynote at the Strata conference Simon, who now works for Twitter, remembered one of The Guardian’s first Big Data problems: whilst exploring the nascent field of data journalism they received a dataset of over 100,000 rows. At the time their toolkit was a version of Microsoft Excel that could only handle 65,535 rows of data. They had to find ways to work around their existing technology.

This definition is surprisingly useful. Where you are uncomfortable with either the volumes, the insights or the actions that can be driven by your data then you are entering the world of Big Data.

A more complex definition, and one that is perhaps overused, is that provided by Gartner. In an attempt to provide a more precise definition (and one that was perhaps more memorable) they identified the useful 3Vs mnemonic.

Where you are uncomfortable with either the volumes, the insights or the actions that can be driven by your data then you are entering the world of Big Data

## The three Vs (Figure 1)

### Volume

The most straightforward of the Vs to understand is volume. The size of data files is increasing, and things that we once saw as analogue and not analysable are now digital and predictable. My first computer had a storage capacity of 1 kilobytes (a ZX81 – I bought an additional 16 kilobyte storage unit). The Word document I'm writing this on is already more than 180 kilobytes in size, and it is sitting on a computer with over 256 gigabytes of storage – 268 million times bigger.

Pictures, which were once analogue chemical imprints on paper are now stored as digital files that can be six, seven, eight or more megabytes in size. Movies can be downloaded in digital format. Songs are transmitted digitally. And all of these are stored on spinning hard disks or in solid state memory.

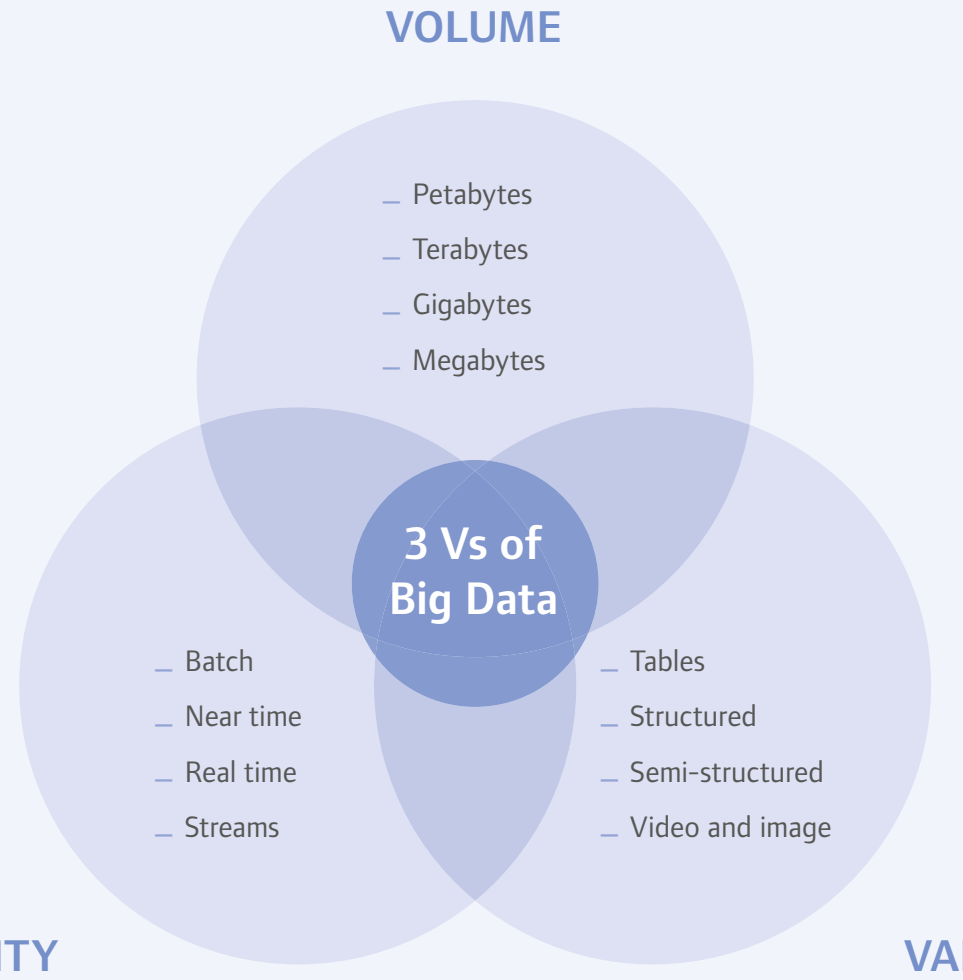
And data is not just stored in 'computers' – it's stored in phones, music players, fridges, microwaves and even Oyster cards.

The reason for the increase in volumes is partly down to the cheapness of storage – you could store all the music ever written in history on a single hard disk that would cost less than $600 – but also down to the ease of producing it. Devices that store data now almost always produce it too, and this data and metadata – data that describes the data – goes on to be stored somewhere else.

And alongside the ease of production comes ease of analysis. Moore's Law, which has held true for almost 40 years, tells us that computing power of processors (technically the number of transistors on a chip, but this is a good approximation for computing power) will double every two years. More computing power allows us to perform analyses through brute force that were impossible even a few years ago.
Many of the computational and statistical techniques beloved by Big Data have their roots in the eighteenth century. One of the most famous is Bayes' Theorem, named after its inventor, Thomas Bayes (1701-1761). In the eighteenth and nineteenth

Moore's Law, which has held true for almost 40 years, tells us that computing power of processors will double every two years. More computing power allows us to perform analyses through brute force that were impossible even a few years ago

**Fig 1 - The 3 Vs of Big Data**



centuries to use the theorem would have required hand calculations, and performing it on data of more than a few dozens of rows would have been extremely difficult. Today I can run a Bayesian analysis on millions or billions of rows of data with ease.

Do charities often have very large datasets? Compared to commercial or scientific organisations probably not. But compared to the size of datasets that they are used to dealing with things are definitely changing. They are also going to have to deal with datasets that are surprisingly wide, for example data from surveys. This is another aspect of the volume of data, and there are a range of advanced data science techniques such as partial least squares and random forests[2] that are ideal for deriving value from such wide data.

Do charities often have very large datasets? Compared to commercial or scientific organisations probably not. But compared to the size of datasets that they are used to dealing with things are definitely changing

Volume isn't just about raw size. If it was, then the British Library would be a good example of Big Data. It's also about the ability to process and use that large amount of data, and do so in a timescale that makes sense for the organisation involved. Which brings us nicely onto the concept of velocity.

### Velocity

In Big Data terms the velocity of data refers to the speed at which data is being generated and the speed at which it needs to be incorporated into actions through analysis.

In the commercial world the speed of data used to be defined by the cycle of accountancy – data's primary use was to keep your books in order, and so data was accumulated on a quarterly or monthly basis. This, in turn, meant that analysis didn't need to be very fast. In theory as long as you finished your calculations before next quarter's data was ready then you were fine.

In the 50s the J Lyons and Co food company decided to use the new science of computing to automate their overnight food production activities – something that was crucial to make their flagship coffee shops operate. Now the speed of data was daily, rather than monthly. This was a huge step change, and required significantly different

**2**
Partial Least Squares is an analytical regression technique that allows effective analysis of data with many variables, but relatively few rows. 'Random forests' are decision tree based models used for classification. Both are techniques that have benefited from the significant improvement in computing power.

**http://en.wikipedia.org/wiki/Partial_least_squares**

**http://en.wikipedia.org/wiki/Random_forest**

technologies. Gone were the banks of computers (people employed to compute totals), and in came a brand new tool – the electronic computer. But because this was the 1950s Lyons had to build the computer, and write the software, themselves – LEO was the first commercial computer.

Walmart, as we saw earlier, moved from daily forecasts to forecasts based on 10-minute slices of data in the 1990s. By the 1990s computers could be purchased rather than built, and the technologies that sat on top of them had advanced significantly too. Walmart put their data into a relational database (RDBMS). This gave them significant advantages over the spreadsheet applications of previous generations. A spreadsheet is essentially a computerised version of the ledgers used in accounting. The computer can add, subtract, multiply and more within the spreadsheet – but they can't effectively link to other spreadsheets. A relational database can, because the system specifies how to link tables together, in a way that can be very flexible.

And as we move into the internet and social media age the velocity rises again. The most Tweets per second recorded to 2013 was just over 143,000 [3]. When you analyse the internet or Twitter you have to go beyond relational database technologies, firmly into the Big Data arena.

However, there is a caveat when it comes to velocity. Although data might be generated increasingly quickly, not all analyses need to be performed in 'real time'. Often the limiting factor is the speed at which you can take action based on the data – and this may be many times slower than the data.

Can the third sector take advantage of the increasing velocity of data? Undoubtedly, although it may require changes to the way that other data is gathered in order to make sense of it. Hampshire County Council's Special Educational Needs team were interested in the possibility of Tweets providing an insight into future needs – as people talk about issues their children are facing. The Tweets were available on a millisecond basis – the school-based information only annually.

Although data might be generated increasingly quickly, not all analyses need to be performed in 'real time'

3 http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats

### Variety

Variety describes the number of different ways that information can be encoded as data – the types, formats and standards that could be used.

When LEO was producing its orders the data it had to deal with was well understood, and relatively simple. It was simply a transcription of the orders that would have been written out by hand, transcribed onto punched cards.

Data was required to be in certain formats for computing – ones that could be easily converted into the bits that the computers could understand. This was a limitation both of the computing techniques and the mathematics of the time. When British Rail wanted to understand the shortest route between any two of its stations in 1955 they had to contend both with writing graph data into the LEO computer and determining the mathematics needed to solve the problem[4]. This was clearly a foreshadowing of the variety issue.

Since then data has evolved. New types, including digital images and video, social media data, sensor data, and mobile application data, have become prevalent. They may have been lurking on the fringes, but until there was the power to process them and the techniques to turn them into understanding, there was no real reason for people to actively generate them.

And even where it looks as if you have a single type of data, you may in fact have many. Imagine you want to take data from a mobile phone – for example to identify where staff move during the day. Assuming we are only going to use data from smartphones we still have four major platforms (iPhone, Android, WinPhone and Blackberry). But within that we have multiple hardware configurations (iPhones 3G, 4, 4S, 5, 5C and 5S, and hundreds of Android handsets!), and multiple operating system versions. You may have to handle multiple discordant formats at the same time! And unlike the simple data of the 1950s to the 1990s this isn't a data format that you can control. Formats can be changed at any time by external organisations.

Variety describes the number of different ways that information can be encoded as data – the types, formats and standards that could be used

4 http://blog.jgc.org/2012/10/the-great-railway-caper-big-data-in-1955.html

## The other Vs

MetaGroup analyst Doug Laney (now with Gartner, the technology evaluation consultancy), kindly provided us with the Three Vs over a decade ago. This, of course, means that other consultants had to explore the limitations of the definition.

Michael Whitehead, CEO of WhereScape Inc, suggests adding Value. This, of course, makes sense. A key feature of Big Data is that organisations can see huge value directly from their data.

Another commonly talked about V is Veracity – proposed by IBM – an understanding of the reliability, or unreliability, of data. This, however, is not a limitation or a description that is in any way unique to Big Data. Indeed it can be argued that the growth of statistics in the eighteenth century was directly linked to the existence of uncertainty in data.

Yet another is Viability. In this case Viability refers to the ability to quickly identify data that actually contains useful information. Again this makes sense. One of the criticisms that is sometimes raised against Big Data is the idea that although data volumes are increasing, most of this is junk DNA – useless noise in which there is little of any real value.

## Towards a common definition – The rise of the data scientist

> "A good rule of thumb to keep in mind is that anything that calls itself a science probably isn't." Professor John Searle

Can we synthesise a better, or more relevant definition for the third sector? It is unlikely that there will ever be a completely agreed definition of Big Data, and perhaps we shouldn't be too worried about this. If data is the 'new oil' as Clive Humby (of data analyst super-firm DunnHumby) put it, then we should instead focus our efforts in maximising the value that it can provide. The three (or four, or five) Vs give us useful

It is unlikely that there will ever be a completely agreed definition of Big Data, and perhaps we shouldn't be too worried about this.

pointers for thinking about when we are in the realm of Big Data. What we need now are guides who can help us navigate this realm, and perhaps bring us to a useful definition in a different way – we need data scientists[5].

With the rise in Big Data came a corresponding need for people who can manage, organise, understand and analyse it. As we'll see later, the technologies involved are sometime esoteric and often less mature than could be hoped. As a result we needed a special breed of person to help us – and they need a method to enable them to act.

The method is data science, and the people are data scientists.

If there is no fully agreed definition of Big Data it will come as no surprise that there is also no fully agreed definition of what a data scientist is. They are often described as part analyst, part coder, part business specialist. Their goal is to find meaning in data and to build data-focused products. Their raw material is Big Data.

Often data scientists have a background in computer science, mathematics, statistics or one of the 'hard' sciences. But they usually have spent some time using data in the commercial world too.

The key traits[6] of the data scientist are:

- statistical or data mining knowledge,
- computing or coding knowledge,
- storytelling skills,
- business expertise.

What we need now are guides who can help us navigate this realm, and perhaps bring us to a useful definition in a different way – we need data scientists

5 http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

6 www.forbes.com/sites/danwoods/2011/11/27/linkedins-monica-rogati-on-what-is-a-data-scientist

Occasionally a fifth trait is added: insatiable (or perhaps, insane) curiosity. Data scientists want to know why.

We will look into the role of the data scientist more later in this document, but for now let's just concentrate on one major trait: business expertise. This is the key to making data science and Big Data work for the third sector, the tight linkage between the data, the analysis, and the goals and objectives of the charity.

Harlan Harris has investigated how data scientists self describe, and has identified different types of data scientists based on a varying set of skills (see Figure 2) - these relate closely to the data science traits described by Monica Rogati: **http://strata.oreilly.com/2013/06/theres-more-than-one-kind-of-data-scientist.html**

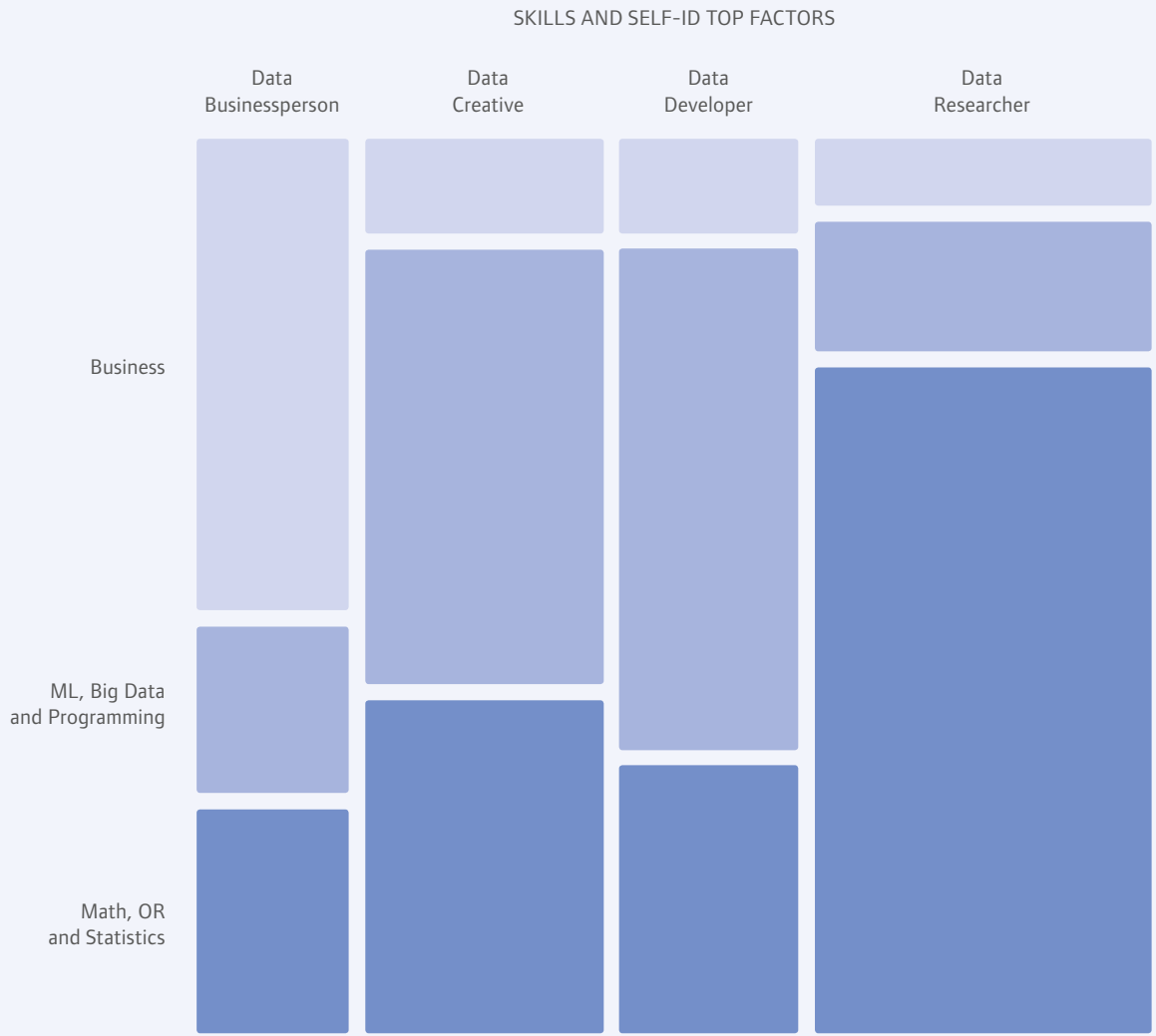
The method – data science – is also vital. It is an analytical and experimental approach to deriving valuable insights and actions from data. It borrows significantly from other techniques, but gives us a way of transforming all of that Big Data into real value.

## The data scientist's definition

For the sake of this document then we will start with the business problem and work outwards. This has the merit of allowing us to focus as much on the 'why' of Big Data as the 'what', 'how', or 'who'.

**For the third sector, data science is the way that charities can transform and grow their activities by analysing and using data. The raw material for data science is Big Data – data that because of its size, complexity, or the sophistication of analysis required, is outside the comfort zone of the charity concerned.**

## Fig 2 – Types of data scientist



# What could you do? Where can you find Big Data?



Data science provides huge opportunities for the social sector to transform their activities, but these opportunities are not evenly spread. In this section we'll look at some of the ways that Big Data can be effectively used, identify some of the types of charities who may be able to benefit, and identify some of the key data sources involved.

In the world of Big Data there are often more opportunities to evaluate your hypotheses than there were before, and you should make the most of them.

## What could you do?

The two great goals of data analysis – whether your data is big or not – are prediction and classification (and by definition identifying outliers that aren't in a class). This section will look at some of the types of analysis that can be performed, some of which are general, and others (social media and social network analysis) that wouldn't even be possible without Big Data.

In reality both approaches have significant crossover: a predictive model also provides insight into the causes of the prediction, whilst a classification system also allows you to take the types of actions you would with a prediction.

A third, simpler, but equally important idea is the ability to confirm or reject a hypothesis. This technique is also used to assess the success of predictive models when used in the real world – statistical testing.

For example, you have a hypothesis that a specific training course reduces the rate of reoffending, and you want to test your theory. Statistical tests will tell you if your hypothesis is correct or not.

In the world of Big Data there are often more opportunities to evaluate your hypotheses than there were before, and you should make the most of them.

### Prediction

At its heart, Big Data prediction involves taking examples from the past, and learning

how to predict behaviours in the future. This is an incredibly powerful ability.

What can be predicted? Almost anything where there is a causal link between data elements and outcomes, and where you have historical information.

Examples of the types of prediction that could be useful for social organisations include:

- identifying supporters likely to donate,
- predicting responses to appeals,
- finding out the impact of behaviour changing campaigns,
- predicting where services will be needed,
- forecasting trends and changes.

Prediction is rarely the end of the process: predictive models are actively used in two ways.

The first, and most obvious, is to predict what will happen, often at an individual level. Who is most likely to benefit from an intervention? Our predictive model can give a likelihood (I'm carefully avoiding the use of the word probability) for each person, and we can then rank people to identify the best candidates.

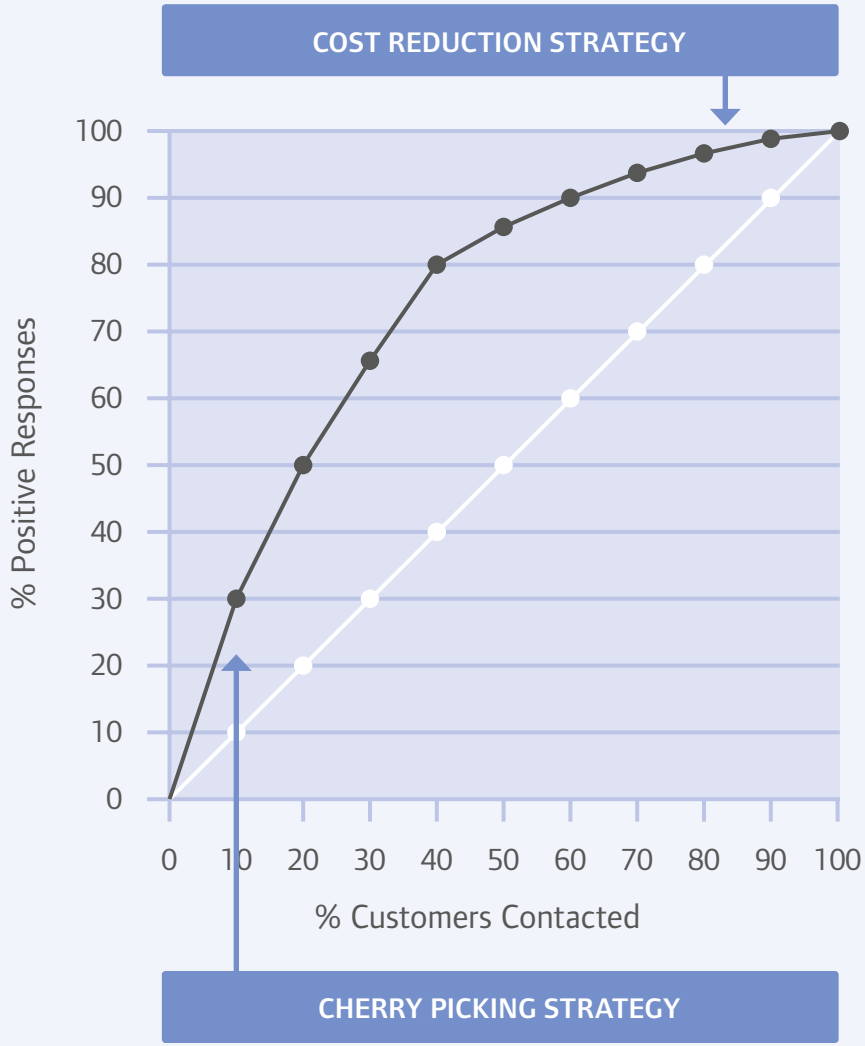
It's important to note that the model will not be perfect, but it doesn't need to be – it just needs to be better than current practice.

This type of use is often important where you have a resource constraint – number of staff or sessions, budget to spend, or time to take actions – and you want to ensure the best use of these resources (see Figure 3).

## Fig 3 - Cumulative gains chart

The gains chart shows a way of understanding how to use predictive models

- The chart assumes that you score data and order it from highest to lowest.
- You assume that you take action based on this order.
- The 'lift' is the degree to which the line extends above the result you'd get with random selection
- Two strategies are obvious
- Reduce the number of actions whilst maintaining almost 100% response (cost reduction)
- Target the best responders (cherry picking strategy)



The second way to use the model is as a means of understanding behaviour. We could look at the variables and find out which ones had the most impact on the prediction. This, in turn enables us to develop new services or improve existing ones.

### Classification

Classification looks to put examples into groups that have more in common within the group than they do to other examples outside the group. The powerful thing about machine learning classification is that it does not bring into play any external biases.

Grouping your data can be done to reveal features that you weren't previously aware of, or to help you build better predictive models.

### Finding the unusual – detecting 'outliers'

When we said that data scientists need insatiable curiosity, outliers are the data that feeds it. An outlier is an example in a dataset that is very unusual. In traditional data analysis we tried hard to remove outliers, as they caused issues with the predictive or classification models.

Although they can still do this with Big Data analytics, the potential size of the datasets means that we often have more of them than we would have had previously, and they are now a big enough group to be analysed by themselves.

Analysis of outliers can lead to insights that we would not otherwise have had. Why did goat prices spike in Africa? The data point is an outlier, and may indicate a problem with our data sourcing, or something truly interesting.

### Social media analysis

Often when people think of Big Data the first things they will think of are Facebook and Twitter. These are certainly interesting and powerful sources of data, providing insights into the thoughts and networks of people that are difficult to gain from other sources.

Often when people think of Big Data the first things they will think of are Facebook and Twitter

There are many things you could do with social media data, depending on the depth of data that is available to you. In order of increasing complexity these include:

- **Brand understanding: 'Do people like our charity?'**
  This can be achieved by simple sentiment analysis of text – classifying statements into groups of either positive or negative sentiment.

- **Issue understanding: 'What are people talking about?'**
  Again this uses text analysis techniques (which can equally be applied to other documents), and word counts may be sufficient. In order to help visualise the data other techniques such as word cloud can be useful. (See 'Word clouds' box p.34.)

- **Influencer analysis: 'Who is important?'**
  This approach uses social network analysis (see below). It can be used to identify people who have strong influence, or who are vital links in communities.

- **Add context to customer information: 'What drives actions?'**
  In order to perform this type of analysis you need to be able to link social media accounts to individuals. When you have these links you are able to determine not just if something happened on Facebook, but if that had an impact in the real world. For example, someone may 'like' a campaign message on Facebook, but it is much more valuable if they then make a donation or write to their MP.

- **Service-led social media strategy: 'Help me!'**
  When a charity is providing a service that is aimed at a geographically separated community then social media can be directly used as an active channel. The most straightforward way to do this is very labour intensive – have a team of people following social media looking for activities of interest. In order to filter and detect these events a data scientist can create a set of scripts to harvest, analyse and prioritise the messages. Obviously these may also require reactions to be in 'near real time' – within an appropriate response time. The length of time that is appropriate will depend heavily on the nature of the issue and the resources available to act on it.

## Word clouds

Word clouds: although these are frequently looked down on by graphics specialists, when used with care they can be powerful and simple to understand. Word clouds take word lists, exclude common words like 'the', then show the most frequently used words in a pattern where the size of the font used for a word indicates how frequently that word was found.


Keyfund were analysing applications for funding. Although there were a number of discreet answers on the applications (applicant age, school name etc), many of the answers were provided as free text entry. In order to help analyse the data wordles were created of the answers, allowing the charity to see the contrast between successful and unsuccessful proposals.

**Organisational social media strategy**
Creating an interaction framework for your organisation.

## Social network analysis

Another useful class of analysis is understanding the connections between people and things – social network analysis. This was first proposed by researchers from the Manchester School, looking at how small groups of people interacted – the new concept was to describe these relationships using mathematics.

In social network analysis there are nodes (things of interest), and edges (the connections between nodes). When using social network analysis to interpret social media the nodes are people and the edges are posts or tweets. By investigating the flow of messages along edges we can understand things about the importance of the nodes. Often we are looking for people who exert influence, or for people who act as bridges between groups of people (see Figure 4).

Social network analysis can also be used to understand organisations. One example of this analysis in the third sector is to understand the sector itself and its funders. In this case both people and organisations (and potentially geographical locations) can act as nodes.

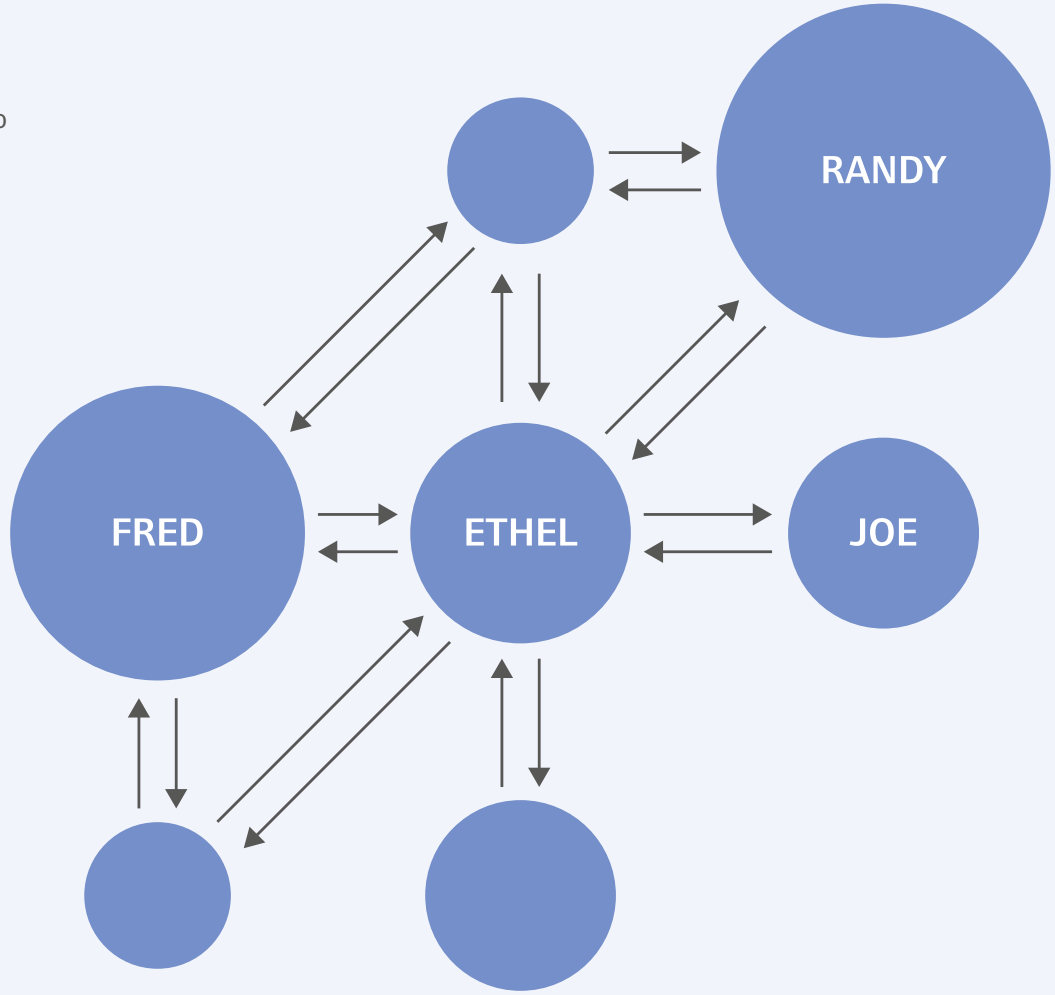
Another example of the use of social network analysis for social good is in the tracking and prediction of disease. Research by Rochester University[7] has looked at how social network analysis can be used to track infectious disease, using a combination of Twitter and geographical data. In many ways this is a modern update of the approach used by John Snow in 1855 to determine the source of Cholera outbreaks in London[8]. In that case the network nodes were water pumps and people, and the edges were caused by people drinking the water.

7 www.cs.rochester.edu/~kautz/papers/Sadilek-Kautz-Silenzio_Modeling-Spread-of-Disease-from-Social-Interactions_ICWSM-12.pdf

8 www.bl.uk/learning/histcitizen/21cc/publichealth/sources/source13/snow2.html

## Fig 4 - Social networks

- Social networks can be directed or undirected and give us an ability to understand the importance of people within their social context
- Networks are constructed from nodes (people) and edges (relationships)
- Relationships could include social media (Facebook, Twitter), membership of organisations, telephone calls etc…



## Who is best placed to benefit?

Most, if not all, social organisations could benefit from the analysis of Big Data. But there are some areas where there are clearly greater opportunities.

### Fundraising

Fundraising is a long-term issue for all charities, and there are areas where Big Data can make significant impact. Although it may be looked down on as being removed from the core objectives of organisations, it is a very necessary task.

Fortunately the direct parallels with the activities of commercial companies means that there are many examples of how to utilise Big Data analytics in this space:

- better targeting of donors,
- evidence base for major funders,
- optimisation of marketing spend,
- multi-channel understanding of responders.

### Development

The example of Oxfam GB's food price work shows some of the potential in the world of international development and aid. This can be a very interesting field, with the use of diverse datasets to try and identify issues in a timely and actionable way.

Across the world there are projects looking at land use, mobility of people, crisis relief, and more. Ban Ki-Moon, Secretary General of the United Nations launched UN Global Pulse[9] to take advantage of these opportunities. "The private sector is analyzing this new data to understand its customers in real-time. The United Nations must do the same for its constituents – people around the world who are losing jobs, getting sick and having difficulty feeding themselves and their families."

Fundraising is a long-term issue for all charities, and there are areas where Big Data can make significant impact

9 www.unglobalpulse.org

### Influencing policy

Although most of us don't have the connections with policy makers that the UN General Secretary has we are often called upon to advocate for our causes with local or national policy makers. Even within our own organisations there are debates to be had about the right way to utilise limited resources – and this is another key area where Big Data can provide insight.

### Health

One of the most fruitful areas for Big Data is in the arena of health. The opening up of previously unavailable datasets through the open data movement, together with the increasing availability of sensor data is providing the opportunity for massive innovations in health.

### Community involvement

Big Data is not a one-way street. Approaches such as crowdsourcing mean that we can start actively and ambiently involve the community, in decisions and actions.

To do this we need effective ways to share data, as well as to listen, and some intriguing steps were taken in this direction with the Civic Dashboard[10] prototype built by Mudlark for Birmingham City Council. This took the data from the customer contact system and plotted it geographically, giving citizens a way of seeing where other people were identifying problems.

### Children and young people

Young people are growing up in a world of Big Data, and the current generation is undoubtedly the most connected in history. This provides both a rich source of data for analysis and challenges in how to use it responsibly and without breaking the barriers of privacy.

Keyfund work extensively with young people in the north east of England, and gather data about their activities in order to better tailor activities and programmes.

10 http://civicdashboard.org.uk

## Using data to understand the health of the voluntary sector and to help direct policy.

**Ben Gilchrist - Community and Voluntary Action Tameside (CVAT)[11]**

CVAT are the support organisation for voluntary, community and faith groups in Tameside. Our work is about helping all of those groups, from really small local groups up to national charities, to develop and grow, sustain that work and understand the area in which they're working. So we work with about 1,000 different groups, ranging from core work around funding, how to get themselves set up, and then to the work that I'm responsible for, which is around the policy environment and having influence on local decision makers. And the use of data is something we're really interested in championing as a way of influencing policy.

The data needs of the entire voluntary sector are really big. Even with the more developed charities we can get together and articulate the challenges and opportunities, but when it comes down to saying "what is our evidence for that?" we sometimes come up short. I can see a lot of room for improvement.

Funders are asking for more data and more evidence, and it's part of our role to be supporting organisations with that. And definitely not just to be driven by a funding cycle, but to recognise how vital that data can be for getting that work you do right and developing it further. To ask the questions: "who are these people we're working with, and how effective is what we're doing at making a difference in their lives?"

A really good example is some of the work around food banks – the Trussell Trust has one of the biggest national networks of those, and they've done a really good job of collecting the data. Not just the number of food parcels, but also the geography, the ethnic backgrounds and tracking the trends. It's really, really powerful.

11
www.cvat.org.uk

We focus our work in certain thematic areas, and at the local level any themed area could harness data better. But the ones that stand out for me are work with children and young people, work around health – we know that there are a real range of voluntary groups in health, where there are a huge range of local organisations that may not have health as their primary mission, but they're doing a lot to promote the health of their clients. Also work with older people, where there are huge impacts on isolation. But can we prove that?

Generally the messages we've got from our core data reinforce our understanding, but we're starting to dig into what we don't know. For example how do things differ thematically between different types of services? We're starting to be able to investigate things further now.

For the majority of the voluntary sector, which is really small organisations, it's hard for them to see the relevance of Big Data. But we need to show them that there are small parts of that Big Data that they can see and use. For the minority, including ourselves, we can absolutely see the value of Big Data.

## Using data to understand the limits of coverage in the health sector.

**Ian Carey - former CEO of Barnsley Hospice[12]**

Data is about demonstrating one's impact and the effect of your services, which is complicated to do in health especially for small to medium sized organisations like Barnsley Hospice. So like many organisations we were reasonably OK with data about finance, understanding it not just in terms of basic accounting, but also how

12 www.barnsleyhospice.org

to use it. But when it came to translate this to what we do with our patients, fundraising, retail services, people had more problems in understanding what it was about.

So the Hospice has ten inpatient beds, although it's about more than just that, and Barnsley has a population of about 300,000. We were only seeing about 100 deaths a year, but there were 2,500 deaths each year in Barnsley – what happened to all the people we didn't see? We looked at understanding our internal length of stay, and how that compared on a benchmark, and trying to work out why we weren't inundated with referrals.

In a simple way we had created our own internal analysis and estimates of the need for specialist palliative care. And although we were providing good care we were only seeing a small proportion of the people who needed it, and this made us focus on the ones not getting the service, because it turned out they weren't getting any care at all. We were able to use that from a campaigning point of view to try and rectify that.

We started looking at some of the open data sources, from GPs for example, and compared the number of people on their end-of-life list against the size of their practice. We also noticed that there were some practices with a high number of people with end-of-life care needs where we weren't getting any referrals. And by looking at the referrals to other providers we saw that they weren't getting the referrals either. It was pretty good evidence from a triangulation point of view that there were practices that weren't referring people on to anywhere. Demonstrating that to the Health and Wellbeing Board and Clinical Commissioning Groups was very powerful.

Data analysis creates more questions than answers, but that leads to better decisions.

## Where can you find data?

Although we've indirectly identified some data sources above, we'll quickly look at some of the more prominent Big Data sources. Again, this isn't an exhaustive list, but it should provide a good starting point.

### Internet data

Data from the internet is becoming more and more detailed. Originally websites were just passive documents, but today they are interactive forums that can handle sophisticated code. As a result the data that you can retrieve from website visits has expanded rapidly. It is straightforward to tell if someone has come from a referral site, what browser they are using (and therefore what operating system they are using), and what pages on a website they looked at.

With slightly more work you can examine every action on the website. Clicks, where the mouse moved, abandoned purchases and so on. You can even find out some geographical information through the use of IP addresses.

### Mobile data

The internet isn't limited to fixed desktops, of course. The proliferation of smartphones has also resulted in a proliferation of mobile data. This includes not just the standard information from the internet, but also potentially GPS positioning data.

This data is also often added as metadata to photographs, adding additional depth to visual analysis.

### Wide data

Data from surveys is frequently wide, rather than deep – it has many columns of data, and not necessarily many rows. A single in depth interview for a client may have hundreds of answers, and more if a branched logic is used.

Wide data presents its own processing challenges, with sparse data and ranching, and is an area where Big Data analysis is often the best way of interpreting it.

### Integrated data

In traditional data environments data was often stored in silos. The data from the finance department was separate from the data from clients. Staffing data was in a different database. The appointments system didn't even store its data. External data was somewhere else.

This is a major roadblock to the effective analysis of data. Many Big Data innovations come from the fortuitous linking of apparently disparate sets of data.

### Sensor data

Most of the data that has been mentioned so far is directly generated by humans. In the future this won't necessarily be the case. Sensors placed on devices will generate more and more of the data present in the 'internet of things' and its volumes may dwarf human generated data.

Mobile phones are already acting in this way – pushing out data even when we are unaware. Cars are starting to do this too. Smart meters will do the same for our energy supply.

Each of these are potential sources of data, either accessed through partnerships with the companies concerned, or through crowdsourcing and scraping approaches.

### Geographic data

Some of the most exciting and straightforward work in Big Data is being done by relating data to geographies. Increasingly people are tracking their movements using GPS devices, which can provide locations to within 8m (unless you work for the security services or military, in which case they are much more accurate – the difference being caused by an additional signal). But at a simpler level it is being used because there are a variety of tools that make geographic mapping easy.

Why do maps make sense? People are very visual, and presenting information in this way can be extremely insightful (see John Snow's map of the 1854 cholera epidemic). However it is important to understand that maps are not neutral! At the most maps don't account for population, so simply plotting the number in each geographical area will almost always result in a map that reflects population density rather than anything useful. This can lead to confusing graphics.

Another problem is that population density means that areas of interest in cities will be hard to see compared to the large areas covered by counties. For example the Highland Council in Scotland has an area of 30,660 $km^2$, whereas Islington Council has an area of 14.9 $km^2$. But the population of Highland is only 221,000 only a few thousand higher than Islington's 206,000. If the thing that the charity is interested in is rural in nature it will probably be easy to spot – if it's urban, much harder.

John Snow's map of the 1854 cholera epidemic

### Social media data

Facebook and Twitter are the major social media giants, with millions of users and providing unique insights into connections and motivations.

The two services have very different approaches to their data. Twitter is an open forum, and all data can be accessed. Facebook, in contrast, has richer data, but requires user authorisation to access the most interesting data.

It's also important to remember that there are other social media channels, including, but not limited to:

- location-based social media – Waze, Foursquare,
- photography and image-based social media – Tumblr, Vimeo and Instagram,
- social media that defy easy categorisation – Pinterest,
- international social media – such as Vkonnekt (Russia),

People are very visual, and presenting information in this way can be extremely insightful.

- music-based social media – Buzznet, Last.fm,
- private, or restricted social media - Yammer,
- dating social media - OKCupid.

## Open data

If there is one set of data that is clearly usable it is open data[13]. The term 'open data' is most often used to refer to government information that is made freely available for use. It can (and already does) include information on population, deprivation, transport, prescribing data and more. It's also worthwhile thinking about all the other elements of government data that should be open: court records, local government spending, Companies House data, Charity Commission data.

This doesn't have to be the limit of open data though. What about data from the private sector? Orange France Telecom has already made steps in this direction by releasing (anonymised) telephone call record data from Ivory Coast under the Data or Development[14] initiative of UN Global Pulse. This is the equivalent of the 'metadata'; that the National Security Agency has stored in the USA – but to be used for clearly beneficial purposes only.

Just from these examples it's easy to think through how some of this data could be used to help inform the decisions about how your organisation works.

Of course open data isn't perfect. If you're going to use it you need to think about how the data will change over time – remembering that unlike internal data you don't control this. You also need to be even more aware of the meaning of data and the methodologies used to collect it. You may also need to think about how you're going to deal with it from a technical stand point.

13 www.nominettrust.org.uk/knowledge-centre/articles/open-data-and-charities

14 www.unglobalpulse.org/D4D-Winning-Research

# Case study: Using Big Data to shift the debate



HyeSook Chung is the Executive Director of DC Action for Children[15], a US non-profit that advocates for children in Washington DC. Washington is a diverse city that suffers from high levels of deprivation despite its role as the nation's capital. In this interview she describes how DC Action for Children transformed their ability to interact with decision makers by taking a traditional set of information and transforming it into an interactive, Big Data project. Her story highlights some of the challenges that are faced when changing directions like this, and highlights the importance of leadership in achieving analytical goals.

**Tell us a little about yourself and your organisation**

In my current job I'm Executive Director of an organisation that advocates for DC children. In DC we have thousands of children in poverty.

The Kids Count project is a long-standing one. The funders knew that there had to be a lobbying effort in parallel to the programmatic effort of raising kids out of poverty, and as part of that work they felt that they had to have evidence from data. So they had a data project for 24 years, and each of the 53 organisations (one in each state plus DC, Puerto Rico and the US Virgin Islands) uses data scorecards to ask "Are kids doing better or worse than in previous years?"

**It sounds as if the project was already very data-led?**

Exactly. The theory was that with data around child wellbeing we could make the case for better services. So there was very much an understanding of the connection between data and advocacy.

When we were asked to take over the DC Kids Count project two and a half years ago I thought "Great – it's a national project with a seal of integrity around data". But I realized that what we'd inherited needed to be overhauled.

When we started reaching out to other stakeholders, the Mayor, and agencies, we

15 www.dcactionforchildren.org/node/58

asked who among them was using it. And the answer was: none of them. It was a static book, about 140 pages. Essentially it was an annual trend analysis.

I was flabbergasted when I saw how undynamic the data was. Even though it was an annual review many of the indicators weren't being changed on a year-to-year basis.

**Were the funding Foundation's representatives nervous about your new direction, and if so how did you calm them down?**

A lot of their questions were around "Do you have the capacity to do the analysis and use the software?" They were also concerned about our access to the raw data. But after we'd got over that their next concern was "Why aren't you doing a PDF book?" So we had to go through the whole concept of open, dynamic data.

When we give a book we are instilling on the audience our notions, theories, but we want the city to see how useful and dynamic data can be when making decisions.

**Do you have information on every child?**

No, we get blind (anonymised) data. We get some of the detailed data, but agencies can be very protective. For example the police department – they will give crime rates, but one year they gave it in Excel, the next year they wouldn't give it to us. Year to year it's very hard to get information to do trends. Another issue is that most of our data is at ward level (there are 8 wards in DC[16]) so we worked with a couple of data experts to break it down to the zip code level, which had never been done before.

**Do you have issues with internal or external boundaries?**

Yes, we definitely have issues. One of the big questions is "Are we still a neighbourhood-based school district?" So one of the things we're trying to provide the city is migration maps. In one area 87% of kids were heading out of their neighbourhood to school. We have some theories – poor schools or economic

16 http://en.wikipedia.org/wiki/Neighborhoods_in_Washington,_D.C.

development… So for this next wave of analysis we want to show what's happening, both inside and outside the district.

Beyond that we want to know, based on the data we have, can we run a regression where we can predict how well each child will do?

**Would you classify your data as Big Data?**

With open data I think of accessibility, public, transparent…Big Data? Well we can't compare with California – but yes, it's Big Data because we have a city that wants to use real evidence! We get plenty of money, but it's the ineffectiveness of how we govern, and execute programmes that causes problems.

For example, we just met with someone responsible for one ward – he wasn't interested in things on a city wide basis, he just wanted to know the impact on his ward, and we were able to give that to him. And he was amazed. So yes, I'm a believer in Big Data!

**Are there plans to extend your approach to the other 52 organisations in your network? What are the barriers?**

I've been trying to map out the true costs of doing this across the US. So that's one end. At the other end, the funder pushed back saying "Could we replicate what we did in rural Mississippi or Louisiana?" Part of it is leadership – if I didn't believe this would work it would never have happened.

# The heart of Big Data



When you have made the decision to put data at the heart of your activities you have started down a path to an interesting, and hopefully better, future. Along the way you will be faced with a series of challenges – if it were easy, you would already have done it.

We will start by looking at purely data related issues, moving into analytically focused issues, and ending with a few notes on some of the legal issues.

Before we start it is worth remembering: you are not the only people to have faced these, there are many examples from both the charity and the commercial space where people have overcome these issues – and a few counter patterns where they didn't.

If you haven't been immersed in the world of data before then some of these issues will be new to you. Even if you have, it's worthwhile remembering what they are, and identifying good practice.

## Focus on data

There are a number of key issues that relate directly to data, and we will address some of them next, starting with the most obvious: storage.

### Data storage

Obviously you need somewhere to store data. And ideally somewhere that is:

- accessible to everyone who needs access,
- not accessible to people who don't need access,
- a good fit for analytical technologies,
- a single location that reduces data duplication.

Data duplication, and data silos, are two of the enemies of Big Data. Much of the advances in analytics evident in Big Data come from the ability to link two (or more) previously unlinked datasets to see what you can find.

Data duplication, and data silos, are two of the enemies of Big Data

What happens if I link data on volunteers with data on programme success? What if I link information on bus timetables to information on session attendance? What if I add in weather data?

Clearly having your data in a single location reduces the effort required to run these analyses. It should be said that this doesn't mean that you start with a single repository – you need to be pragmatic about how you decide to source and add data. Finding which data is useful is another good use of Big Data techniques.

The good news is that data storage costs are falling significantly. Huge amounts of data can be stored in systems such as Hadoop Data File System (HDFS) at low cost. Remember, though that the goal isn't simply to store the data. If it was then tape drives would be an ideal solution. The goal is to be able to learn from the data, so accessibility is critical too.

### Data quality

Data quality is one of the major issues that faces data scientists, and indeed anyone working in data. Data is rarely as clean as people would like it to be. We would like data to be present, to be accurate, and to be well documented. But in reality we often find that it isn't. Values may be missing completely. Values might be wrong. Data that is gathered may be incorrectly processed.

As Phil Harvey, DataShaka's Chief Technical Officer puts it "Poor data quality is limiting in three ways: garbage in garbage out, trust, and speed."

Garbage in garbage out (GIGO), is one of the oldest adages in computing. If the data we provide to a system is incorrect there is every chance that the output will also be incorrect.

Trust is also a key element. When you can't trust your data it destroys trust in the process of data science and the value of Big Data. At the end of any analysis we expect and hope that action will be taken. But the action relies on people to do things, and without trust they simply won't.

When you can't trust your data it destroys trust in the process of data science and the value of Big Data

Speed will be crucially impacted by data quality issues, because to address GIGO and trust we will need to take corrective action, and that action, even if incomplete, will take time. If we have analyses that are time critical then data quality issues may prevent us from taking the actions we want.

If this sounds like a story of doom and gloom, then it needn't be. Although data quality is important we mustn't fetishise it. Data quality needs to be fit for purpose, not perfect. When we perform Big Data analysis we should keep this in mind and test our models for their sensitivity to data quality issues.

### Data curation

If data quality is important, then the related field of data curation is the next logical step. The information contained in your data should be documented in a data dictionary, which should be a living, breathing document.

The data dictionary provides the bridge between your current business and analytical teams, but also provides a bridge for future teams. As over 70% of the time on a Big Data project can be spent in data preparation we need to ensure that lessons learnt in this iteration of analysis are not forgotten.

Data curation extends beyond simply documenting data though, and into improving and maintaining it. With the velocity of change in data this is becoming increasingly important. From the perspective of charities reliant on external data – information from partners or government – it is clear that the change of format from year to year requires careful data curation. Where a group of charities are working in the same field working together to curate this type of external data would be an excellent opportunity.

Although data quality is important we mustn't fetishise it. Data quality needs to be fit for purpose, not perfect

**Differing data types and sources: external and internal data**

As organisations move into the world of data science the balance between internally generated and externally generated data is likely to shift. When you are dealing with financial reporting almost all of your data will be generated by your own systems.

When you are dealing with sensor data, or internet data, most of your data will come from external sources.

Because you don't control these sources yourself you are at risk when formats change. For Chromaroma, which accesses Transport for London travel data through a scraping approach, there was always a risk that the data format will change without warning, requiring them to review and rewrite their application.

**Differing data types and sources: structured and unstructured data**

Sometimes Big Data is described as being 'unstructured' in comparison to traditional data's 'structured' formats. In reality it is a misnomer to claim that data is either structured or unstructured, although this can be useful shorthand for different data types. A more useful way of viewing different Big Data types is as follows.

- Data that uses a **stable schema (structured)**
  - A stable schema is one where the same types of data will be provided in the same format on a regular basis
  - This data has been gathered from processes with well-defined and known attributes.
- Data that has an **evolving schema (semi-structured)**
  - Evolving schemas change over time, or by device
  - Data generated by machine processes; with a known but changing set of

attributes. Examples of this type of data would be web logs, call records, sensor logs, JSON, social media profiles, Twitter feeds.

- This could also include open data where the format might change over time, and without warning.

- Data that has a format, but **no schema (unstructured)**

  - Data captured by machines with well-defined format, but no semantics. This is the category that includes images, videos, web pages, PDF documents. The metadata (data describing data) associated with the images and videos would still be either stable or evolving schema.

  - Semantics can be extracted from raw data by interpreting the format and extracting information, for example counting the number of faces in a picture.

## Building an analytical mindset

Although moving to a data-focused approach can give a charity a huge advantage, it is not without risks. These mostly hinge around the idea of making the wrong decisions, and taking the wrong actions, as a result of faulty analysis.

Of course if you would have taken the wrong decision anyway using data to make it is unlikely to make the decision worse. So what are the most common technical issues related to Big Data?

Much of the process issues in Big Data relate to the use of statistical techniques. The terminology here can be confusing – there are subtle differences between statistics, data mining and machine learning, but generally these are for experts and practitioners to worry about.

The most important thing to remember is that all of these techniques focus on the idea of using mathematical techniques to extract new insight from data. And to be clear, the insight we're interested in is that which relates to improving the effectiveness of your charity.

A good guide to the steps needed in this process comes from the CRISP-DM methodology[17]. Set up as part of the Esprit European funding in 1999, CRISP-DM is a problem neutral approach to analysing data.

CRISP-DM identifies a number of crucial steps:

- business understanding,
- data understanding,
- data preparation,
- modelling,
- evaluation,
- deployment.

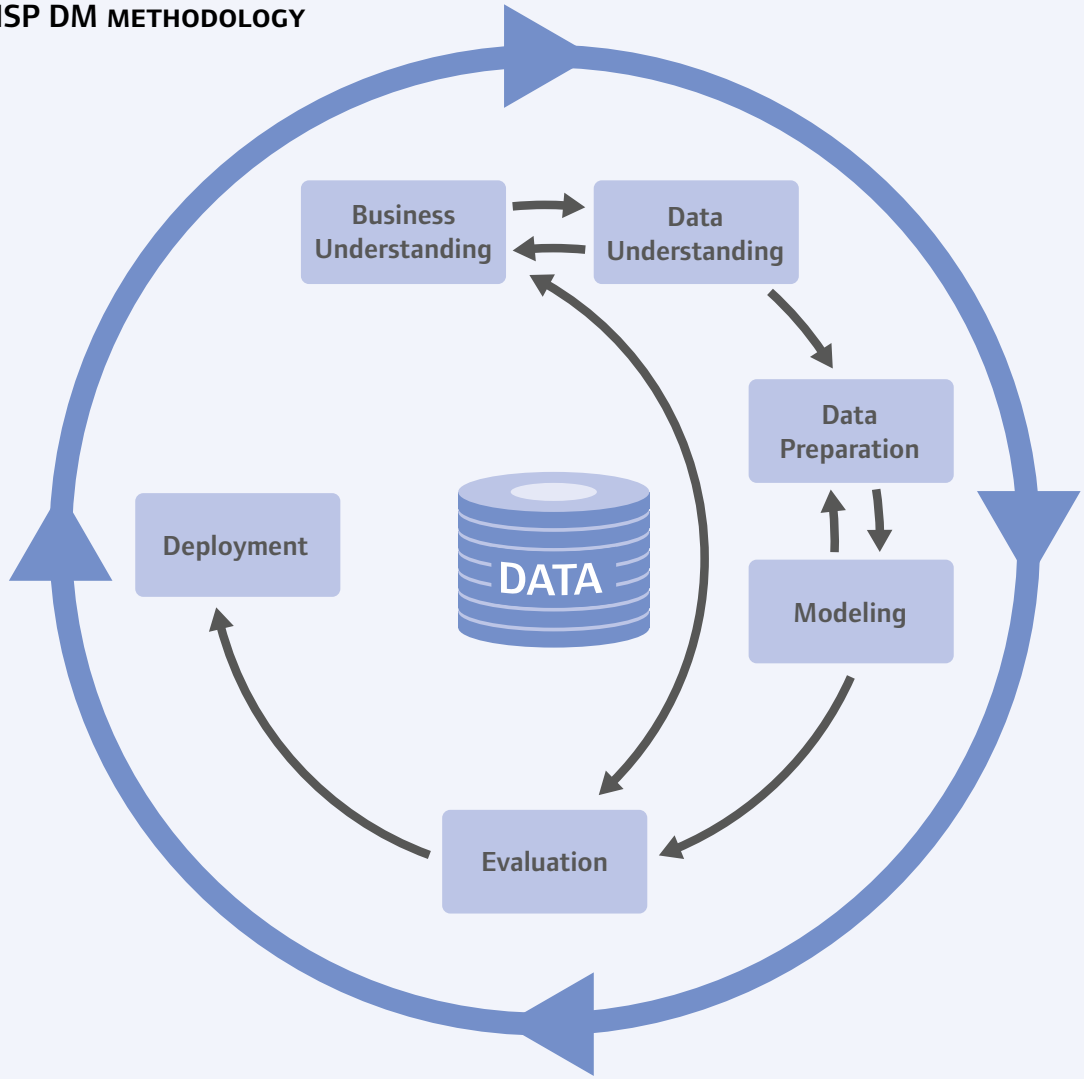
The steps are linked (see Figure 5), but the assumption is that you will have to revisit each one several times during a successful project.

Data science tends to be less rigid in its approach to problems than the older data mining approaches – this is related to the greater emphasis on discovery in data science. CRISP-DM is still a useful reminder of the important objectives of each step, and the need to get the business problem right before launching into analysis.

The most important thing to remember is that all of these techniques focus on the idea of using mathematical techniques to extract new insight from data

17 www.nancygrady.info/CRISP-DM.pdf

**Fig 5 - The CRISP DM methodology**



## Getting the right question

Most Big Data projects that are unsuccessful hit problems because the question that is asked isn't relevant to the charity or business asking it, rather than for purely technical reasons.

Sometimes the right question isn't the most obvious one. For example, if you wanted to know why clients weren't finishing a course designed to improve their employability you might try and predict who was going to leave early. Building a regression model would provide insights into the factors associated with people leaving. But if you wanted to find out who was going to leave, so that you could influence them not to, then the question itself becomes more complex. You should consider:

- what the definition of leaving the course is,
- if there is an earlier indicator that would give you time to react,
- if there are factors that need to be controlled for because they bias the outcome.

An example of a potential factor that could influence the outcome would be people who got jobs before the end of the course, or those who moved out of the area.

Data scientists have been described as being more interested in the questions than in the answers (which is somewhat unfair), and it is certainly true that in the course of a Big Data project it is likely that many more questions will be asked as the project develops.

It's clearly very important that the questions are:

- relevant to the charity,
- relevant to the available data,

- answerable and measurable,
- actionable.

When looking at how actionable they are it is important to understand that predictions may have asymmetric costs. This results from the difference between Type I (false positive) and Type II (false negative) errors. As a simple example, if I was predicting who would respond to a Facebook campaign, I might be wrong in one of two ways: I could predict that customer Y was going to respond when they weren't (Type I error), or I could predict that customer X was not going to respond when they were (Type II error).

In the first case the cost to the charity is the cost of sending out the Facebook campaign to that individual, possibly as low as £0.01.

In the second case the cost to the charity (since I don't contact customer X) is the lost donation. Lets say that this could be £20.

### Willingness to fail

Data science is, to a large degree, an experimental approach to data analysis. Even when using historical data the concept is to create a hypothesis that can then be tested, so that new actions can be taken.

This carries with it a risk of failure. Not every Big Data project will find what you are looking for. Organisations that are very risk averse can find it hard to operate in this way.

Negative results are far from meaningless though. If we have a hypothesis that primary school success is related to first language, for example, then failing to find that link is, itself a very powerful learning.

Not every Big Data project will find what you are looking for. Organisations that are very risk averse can find it hard to operate in this way

## Probability – handling uncertainty

> Rosencrantz repeatedly tosses a coin, which lands heads up…
> Rosencrantz: Seventy-six—love. Heads.
> Guildenstern: A weaker man might be moved to re-examine his faith, if in nothing else at least in the law of probability.
> Rosencrantz: Heads.
>
> Tom Stoppard, Rosencrantz and Guildenstern Are Dead

One of the technical challenges of dealing with data is that you will be moving into a world of probabilities rather than certainties. The joy of Big Data is that as data sets grow in size probabilities become more and more consistent and useful.

This may seem counter-intuitive. But think about tossing a coin. Probability tells us that the chance of heads is 50%. But if I toss a coin ten times what is the chance I will get exactly five heads and five tails? The answer is a surprisingly low 24.6%. As the number of events gets bigger the chance that your split will become closer to 50:50 actually increases (although the chance of exactly 50:50 doesn't) – large datasets make the statistics more certain.

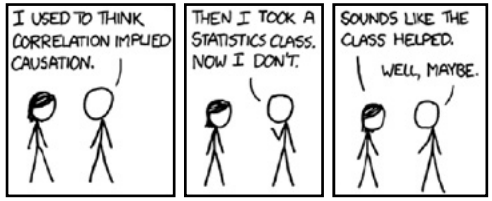
Related to this is the opportunity to spot 'black swans', rare but critical events. As your dataset gets bigger there is more chance that you will see these otherwise rare events.

The downside of this is that standard statistical tests (where we typically say something is statistically significant if the chance of it occurring randomly is less than 5%) will often fire accidentally. In very large data sets there may be many things that co-occur sufficiently to pass this test, but as we've seen correlation does not equal causation. How can we tell the difference between a black swan and a coincidence? In addition to the time factor we can also try to impose a discipline on our analysis. Rather than searching for the things that are statistically significant after we gather our data we should state the things we are going to test in advance.

Rather than searching for the things that are statistically significant after we gather our data we should state the things we are going to test in advance.

## Correlation vs causation

This is probably one of the most important distinctions in Big Data, and yet one that even experts still get wrong. Correlation is a situation when a change in one thing is accompanied by a change in another. For example educational achievement and the number of books in a home are often thought to be correlated (the correlation is a positive one: as the number of books increases, educational achievement increases). You can see that correlation can be a weak relationship, or may even be a chance relationship. This is especially true in very large data sets.



Source XKCD.com Correlation

Causation is when one event causes another. As you might suspect causation is much harder to prove than correlation. In our example above, does poverty cause poor educational achievement? One feature of causation that helps us spot it is that it is time dependent. We can examine when one event happens relevant to another. If Z happens after Y then it can't have caused Y.

We can also run experiments to test if an event is causal or simply correlation. In our first example we can send households books and see if the educational achievement of their children improves. If we were very sophisticated we could send different numbers of books to different households (and even remove books from other households).

## Don't get hung up on accuracy

When people build analytics, especially Big Data analytics, they often focus far too much on the accuracy of the model. Although accuracy can be important, it is not the only measure by which analysis should be judged.

When considering any analysis you should evaluate:

- Required accuracy.
  - This may just be 'better than guesswork'. You need to be realistic about what would be a sufficient level of accuracy to improve the work of the charity. It is important to remember that the first analysis may well not be the last one –

analyses can be reworked and improved over time.

- Robustness.
  - Is the analysis consistent when you run it with slightly different settings? This is the hallmark of a useful, usable model that reflects causal relationships.
- Asymmetric costs and values.
  - See section on Type I vs Type II errors.
- Explicability.
  - Can the analysis be explained sufficiently well to users or interested parties. If you are trying to convince influencers about the impact of a policy it is useful to be able to explain how the results were created. Random forests (a highly accurate advanced machine learning technique) aren't very useful for explaining how the answer was reached. We often find that there is a trade off between accuracy and explicability.

### Avoiding bias

Unnoticed bias is one of the big worries of all data analysis. What if the work we have done has been impacted by bias, and won't work in the real world?

Because of the potential impact of bias on analysis there has been extensive study into the types of bias that can occur, and the ways of detecting and preventing it. Rather than examining all of them we should keep our minds on three of the most important for charities – two that we might be guilty of, and a third that may cause us problems when interpreting other results.

Unnoticed bias is one of the big worries of all data analysis. What if the work we have done has been impacted by bias, and won't work in the real world?

- Sample bias.
  - The first risk is that we build an analysis on a set of cases that isn't really representative of the problem.
- Confirmation bias.
  - The second risk is that we ignore findings or insights that disagree with our initial opinions. This is surprisingly easy to do.
- Publication bias.
  - The third risk is that people will fail to publish results that show them (or their organisation) in a bad light, only publishing results that are positive. If we are using this data for insights we may run into difficulties.

**Putting it into action: deployment and visualisation**

When we've created our analysis, tested it for biases, ensured it's causal, and that it really does relate to a real business problem we face, we then need to find a way to use it.

Sometimes the result we need is a prediction, and a numerical output is ideal. If we want to know which supporters to target for an activity then knowing the likelihood of their response (a percentage) gives us a way of ordering a list for action. If we want to predict who will develop a disease then again a percentage likelihood is a good output.

Sometimes though, we will want to visualise the output we create. This works well because people respond to visual stimuli. A well designed graph can provide more insight than exactly the same information presented as a table. Concepts and facts can leap out of the page.

When we've created our analysis, tested it for biases, ensured it's causal, and that it really does relate to a real business problem we face, we then need to find a way to use it

Beware that the power of visualisation also brings risks. More people have leapt to the wrong conclusion from poor visualisations than from almost any other technique (although Microsoft Excel holds the actual record[18]).

Several visualisation tools have been developed to work interactively with Big Data – Spotfire and Tableau are two examples of specialised data visualisation tools.

## Legal and reputational risks

The use of Big Data will take you into new areas of activity, and there is a strong chance that you aren't fully prepared for the potential legal and reputational issues that might be involved.

It's worth mentioning that the legal risks may actually be less important than the reputational risks to a charity. If your supporters and volunteers feel that your actions are inappropriate they may take their funds and efforts elsewhere.

### Data and the law

The law requires you to take care with certain types of data. The following section is intended as a rough outline only, if in doubt you should obtain legal advice on your responsibilities. However it is worth saying that the reputational damage that your charity could suffer if it gets things wrong is probably at least as damaging as the legal implications.

If you keep and use personally identifiable data (think of names and addresses as a minimum starting point) then you are almost certainly covered by the UK Data Protection Act 1998. This requires you to register with the Information Commissioner's Office as a Data Controller.

The law requires you to secure personal data, to take action to ensure it is correct, and to only use it for the purposes for which it was collected. However, when it comes to

18 www.theguardian.com/news/datablog/2013/jul/24/why-you-should-never-trust-a-data-visualisation

analysis there are exemptions from the Act, specifically provision 33, which allows for statistical analysis of data.

### What is personal data?

Data is personal if it relates to a living individual who is identifiable. This can include audio and image data.

A higher category, **sensitive personal data**, exists. This is personal data that relates to one of a number of sensitive categories such as race, health data, political views and so on. This data must be treated with extra care[19].

Obviously you should tread carefully, as the Information Commissioner's Office can give you significant fines if you breach the Act, but there are practical steps you can take to ensure that you are behaving ethically and legally.

- Make sure you have a strong and effective data protection policy.
- Anonymise your data before analysis – if individuals cannot be identified then the data is no longer personal.
- Make sure that only appropriate people have access to the data.
- Ensure you ask for permissions before you capture data.

### Anonymity

One way of decreasing the risk of getting on the wrong side of the Data Protection Act, and of ensuring that you keep your reputation intact, is to anonymise data. The easiest way is to strip out names. But of course if you leave the address in the data it is fairly easy to identify the person involved.

One way of decreasing the risk of getting on the wrong side of the Data Protection Act, and of ensuring that you keep your reputation intact, is to anonymise data

19 www.ico.org.uk/for_organisations/sector_guides/charity

So perhaps removing the address, but keeping the postcode would work? That way you would still be able to do geographical analysis.

Unfortunately it is surprisingly easy to break this anonymisation (by which I mean to be able to positively identify an individual). Obviously if someone has a rare condition, for example a particular disability, then knowing that someone in postcode B1 1BB has disability Y will be enough for someone to identify them. However it turns out to be simpler than that. Studies in the US have shown that a combination of three simple variables is enough to uniquely identify 87% of the US population[20].

The variables concerned are gender, birth date, and zipcode (postcode).

So should we abandon anonymity as not worthwhile? No, we shouldn't. Firstly it isn't necessary in most analyses to know names, so removing them won't hinder our work. Secondly it will prevent data scientists from leaping to unwarranted conclusions based on irrelevant data. Thirdly it will make it less likely that we will accidentally cause problems for the individuals concerned if data is lost.

Ian Carey, former CEO of Barnsley Hospice described how they dealt with extremely sensitive information on end-of-life care:

"Information governance was also a key factor for us – making sure you anonymise data properly. You want to be able to backtrack to a unique identifier, but you need to handle it carefully. It's so easy to suddenly realise you've divulged personal information. For instance, we were starting to use postcodes to plot activity on maps. And although postcodes can identify a street, in some rural areas it can identify a house. So we changed to use just the first three or four digits of the postcode. We felt that was a lot safer, because people wouldn't be able to guess 'that's Mrs Jones, down the road'."

### Whose data is it anyway?

When data is sourced from the internet there is often an assumption that it is 'free'. In

20 www.citeulike.org/user/burd/article/5822736

some cases this is true: Twitter are clear that what people write in their public tweets is publically available. However in some cases this may not be true.

Websites may put terms and conditions on the use of data, and if you simply take it you may be breaching these terms.

Sometimes websites will include the robots.txt file (see Glossary) or a 'robots' meta tag to indicate that web pages can or cannot be indexed by web crawlers. Although this doesn't replace understanding the terms and conditions of the website, the presence of a robots.txt file that permits indexing is strongly suggestive of a willingness for that data to be used.

### The world is big

A growing issue with data and the law is a very basic one. Whose law? I may be performing an analysis in London, on data about people who live in France, some of whom were traveling in Germany when the data was collected, the data was collected by a US company, and is sitting in the cloud (we don't know where exactly, but we suspect Iceland). Which laws are relevant?

This is less of an issue in the European Union where data protection legislation has the same legal basis – EU Directive 95/46/EC, which frames the concepts across the Union. However, even here each nation has chosen to enact the directive through legislation in its own way.

Things get even more complex when we think about the USA, whose approach to data is founded on the 18th century interpretation of freedom of expression as much as anything else.

### Reputational damage

What would happen to a charity if it was found to be doing something unethical? The results could seriously impact the ability of the charity to fulfil its objectives. The confidence that clients, funders and supporters need to have could easily be damaged.

In most spheres of activity these risks are well understood, and can be readily mitigated. Financial probity, for example, is secured through appropriate policies and safeguards.

But in the world of analysis and data things are not so straightforward. To a large extent this is because there isn't a consensus on important concepts like privacy and ownership when it comes to data. The law mentioned above, the Data Protection Act dates from 1998 – well before the advent of Big Data. In turn it is based on the EU Data Protection Directive, which dates from 1995 – a time when there were only about 25,000 websites! It shouldn't surprise us that the law lags behind public perceptions.

But it's also true that public perceptions lag behind technology, and that they are far from consistent. Different people think of data privacy issues very differently, and it would be a mistake to assume that everyone wants data to be totally private, or totally free.

Perhaps the most famous case of reputational damage occurred when some technically great analysis was revealed by Target.

### Target

Target, a US retailer, have long had a sophisticated data science team. That team were looking at purchase data to try and identify trends and opportunities to sell additional goods. One thing they noticed was that if purchases of 25 key items were made then there was a high chance that the person making the purchases was pregnant.

This was a great piece of analysis – from a purely technical perspective. Target were not just saying 'We think this person is pregnant', they were creating a score that indicated the likelihood that they were. But what was the point in Target doing the

analysis? It was to send a future mother coupons for purchases of baby related goods. Anyone who has been involved in the birth of a child in a UK hospital may be familiar with the pack of offers that new mothers are given shortly after the birth of their child. This is the same idea, but crucially before the birth[21, 22].

What Target found was that when people became aware of what it was doing, via an excellent New York Times article, the general public didn't like it. It made them feel distinctly uneasy.

Would this have been a good analysis for a charity? Potentially it could be extremely valuable from a health perspective. If you can identify people who are pregnant you can give them valuable advice, and the sooner you know the sooner you can help them. The same is true of other health issues too, and there are a range of data that could lead you to a useful conclusion about someone.

Would a charity have the same issue of reputational damage? That depends on the reason that the work was being done, and how the charity used the information. It's possible that the public would be more tolerant of a charity doing this than a private company.

As Solon Barocas, a Student Fellow of NYU's Information Law Institute put it, "If you wouldn't go up to someone and ask them the question directly you probably shouldn't do the analysis."

21
www.nytimes.com/2012/02/19/magazine/shopping-habits.html

22
www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did

# Ten Big Data concepts to explore



## Get engaged in data mining

Let's start by urging you to move beyond simple analytics. Yes, you can learn a lot from a spreadsheet or a bar graph. Correlation analysis can find out more.

But things begin to get really exciting when you start to employ more sophisticated methods to understand the links in your Big Data. Techniques like regression and decision trees – simple prediction and classification algorithms – are easily within reach of most users, and if you really want to push the boat out then more sophisticated techniques such as Bayesian analysis or random forests (see glossary) can take things further.

Mastodon C were asked to look at prescribing behaviour of NHS GPs, to see what could be learned. As well as their interesting work on prescription of generic statins[23] they also took a look at the factors that make a GP likely to prescribe a new drug.

Rather than using simple charts they decided to use a sophisticated data mining technique – Cox proportional hazard modelling – that allowed them to take the time order of neighbouring GP practices into account. This led to some surprising conclusions that wouldn't have been obvious otherwise – that proximity to other prescribing practices was more of an influence than the age of the GPs.

## Why not nudge behaviours?

Nudging behaviour is the technique of understanding behavioural triggers in order to persuade people to do what you would like them to. Data is at the core of understanding how to do this.

Retailers have known for a long time that data gives us insight into what makes people change behaviours. Price optimisation is the science of adjusting prices to encourage people to purchase what you want them to. Elasticity models tell the retailers which products sell more when the price is lowered, and by how much. Cross sell and canibalisation models take into account the effect on other products.

Nudging behaviour is the technique of understanding behavioural triggers in order to persuade people to do what you would like them to

23 www.economist.com/news/britain/21567980-how-scrutiny-freely-available-data-might-save-nhs-money-beggar-thy-neighbour

So are there triggers that you could discover that might change people's behaviours in ways that would help the goals of your organisation? What are the nudges that you could think of?

A recent study by the Cabinet Office Behavioural Insight Team (known by some as the Nudge Unit) has looked at how charities can increase charitable giving using these techniques[24].

**Become agile and fail fast**

Fail fast is a term that is often heard in association with Big Data. It has been appropriated from the Californian business startup community. The idea, when starting a company, is to identify issues in your concept as quickly as possible, in order to avoid wasting time and money. If you find that your idea is failing then you pivot – identify the core of your idea, or a new one that you learned whilst failing – and head in that direction.

For data science the idea is similar in concept. If your analysis isn't succeeding then you need to find out as quickly as possible, and pivot your analysis! Find a new analysis or pivot your ideas.

A practical example: a team of data scientists were working with Oxfam GB to try and predict future world food prices. If food prices could be predicted then the impact of food poverty can be predicted. One team had a hunch that food prices would be related to world oil prices. Since oil plays a huge part in the costs of agriculture this makes sense. After a few hours analysis they came to the conclusion that although there were weak correlations there were also changes that were impossible to link to food prices. Rather than being disheartened they used their experience to try a different approach.

Fail fast recognises that not every analytical idea is going to fly. It celebrates the concept that a negative answer to a question still provides useful, sometimes critical, information. And it's certainly a far better idea than failing slowly.

If your analysis isn't succeeding then you need to find out as quickly as possible, and pivot your analysis! Find a new analysis or pivot your ideas

24 https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/203286/BIT_Charitable_Giving_Paper.pdf

### Visualise the problem

People are visual creatures. We react very well to visual stimuli, putting those billions of brain cells to work in ways that have been evolving for millennia. So finding effective ways to visualise your data makes good sense.

Visualising Big Data requires a bit more thought that visualising small data sets, but can be equally rewarding. And one of the most effective ways to visualise Big Data is to…

### Put it on a map

Geographical, or location related, data is often classified as Big Data because using it effectively can prove challenging from both a technical and an end-user perspective. Many modern devices (specifically smartphones with GPS capabilities) have the ability to record and track position to within 30m. Other geographical data also exists: postcodes can localise information to within half a street, latitude and longitude (or northings and eastings) can also pinpoint locations.

Maps are wonderful ways of making things stand out, and of enabling people to make leaps of understanding.

Often the problem with geographical data is translating between one type of geographical information and another. If data is in postcode format how would you present it on a map? As the centre of the postcode area (which could, inconveniently, be in the middle of a field), or as a bounded area? How would you deal with issues of coterminosity – when boundaries between different geographical datasets, and the organisations that create them, don't match up?

Although geographical data is ideal for presentation on maps, it requires additional work when used in analytical applications. If you are assessing transport needs for a vulnerable group, for example, the distance between two points needs to be well understood. Is it the shortest straight-line distance, the shortest road distance, or the shortest route on public transport? The correct decision will depend on the nature of the problem, and the degree of accuracy that is necessary to solve it.

Often the problem with geographical data is translating between one type of geographical information and another. If data is in postcode format how would you present it on a map? As the centre of the postcode area, or as a bounded area?

And don't forget the lesson of the London tube map – sometimes the best maps aren't directly related to geography!

### Play games

Gamification is a way of turning processes into games. There is considerable evidence that people are willing to play games when they would not otherwise be willing to interact. Gamification usually has a primary goal beyond the game itself, and is used as a mechanism for generating data and changing behaviour.

One of the opportunities for charities and social organisations is to use gamification to reach otherwise unreachable groups. Although you might initially think that this would primarily target young people, the success of applications such as Foursquare and Waze shows it's ability to reach into older age groups (see Figure 6).

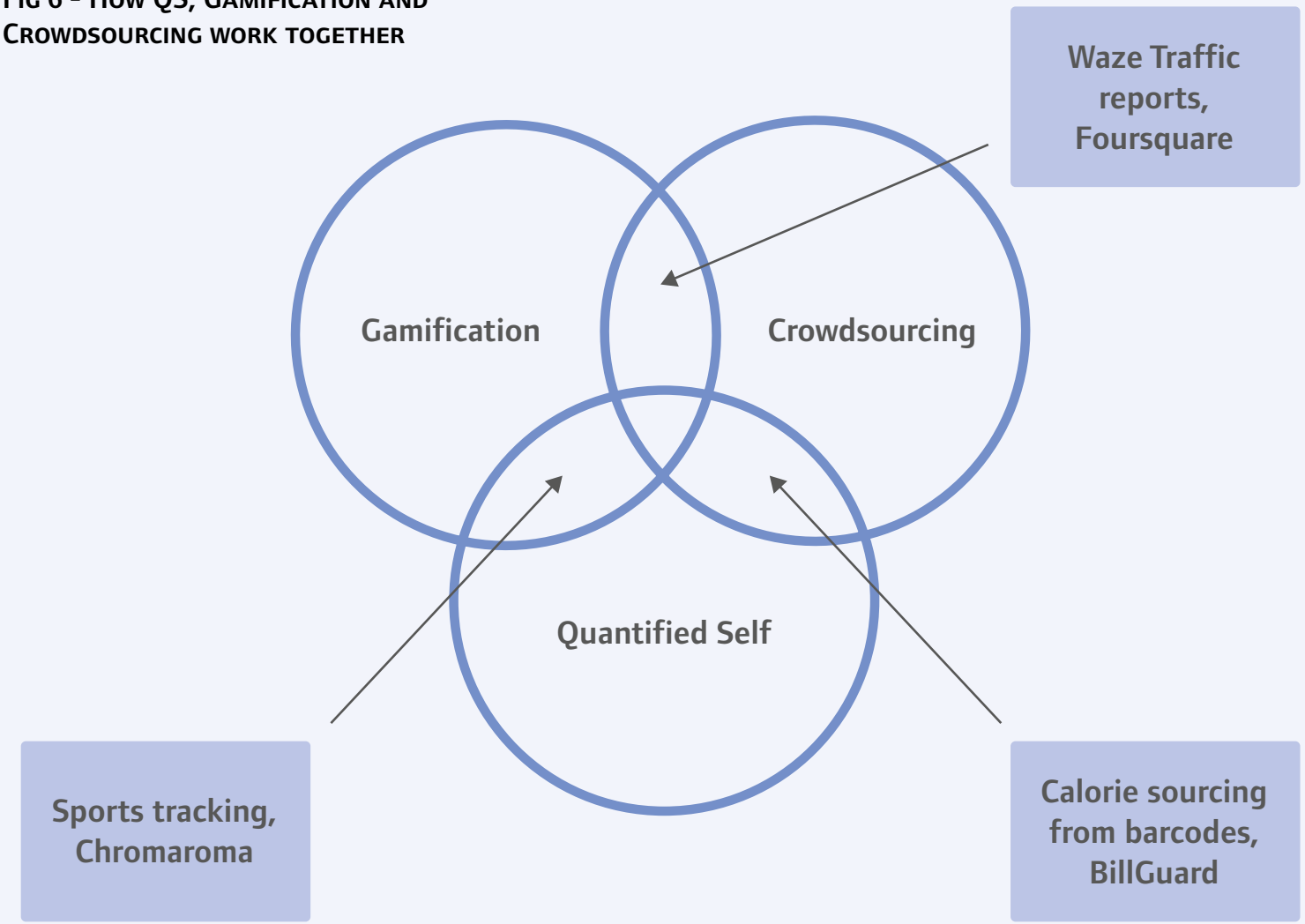
### Chromaroma

Chromaroma[25] is an example of how Big Data can tie together a variety of themes: Gamification, data scraping, and geographic data into a tool that both changes people's behaviour and generates additional data. In the game players provide their Oyster card details, and are put onto one of four colour teams. They then have a variety of tasks to perform, checking into stations, using different routes, and doing a 'goose' (getting off one stop early). All by simply travelling on the London Underground.

Chromaroma shows how an innovative approach can have direct public benefit. Mudlark founder Charles Hunter talks about his team created the project.
"We had noticed that the Oyster card was a portable digital object that tracked your movements in time and space through a transport system. From a gaming point of view we thought 'Well hang on, what if we got that journey data and we

25 www.chromaroma.com

**Fig 6 - How QS, Gamification and Crowdsourcing work together**



could give it back to you as moves in a game?' It was ambient gaming; you didn't need to do anything to play it other than give your card information to the game. We decided to build a prototype that scraped the data. And that's still how it works – you provide your log in details and Chromaroma scrapes that data every night on your behalf, and that gets delivered back to you as a game move.

Our argument has always been that we are giving people back their travel data. That's a useful and true argument from a data protection viewpoint, but it's more important than that. We all create clouds of data every day and we should have ways of getting that data back. I should be able to ask Tesco for all the data they take from me from my loyalty card, and they should return that in a useful way – with graphs and analyses of my diet and purchasing behaviour.

Behaviour change was built into the game, already you can register your home station and work station, and if you get off one stop early – what we call a goose – you get extra points. But there is also a rewards system for travelling outside rush hour, which could potentially help reduce congestion. Or if there was a bottleneck in a junction we could add bonus points – say if there was some problem at Kings Cross we could encourage people to get out at Euston. The real interest is not just moving from car to bus, but from bus to feet – fewer cars, fewer buses, fewer emissions, better health."

### Find some data exhaust

Data exhaust is a term that has been invented to describe data that is generated as a by-product of a business process. For example the process of completing Companies House returns provides data exhaust on the director level connections between companies (and Charities if they are also companies limited by guarantee).

In the world of connected devices this is becoming more and more prevalent. Smartphones regularly generate and store location data from their GPS systems. One

crowdsourced application of data exhaust has been put together by Openpaths[26]. Volunteers allow their location data exhaust to be collected, and researchers can ask them (via the website) for permission to use the data for specific research purposes.

One risk of using this type of approach is sample bias. Sample bias essentially asks you to beware that the group you are using to build your theory or model on may not fairly represent the whole population. In the case of Openpaths there are two potential biases.

- Firstly the only data available is that created by volunteers, and there is no guarantee that the volunteers are a good reflection of society as a whole. They may move more or less than other people, or be based more in one geography.
- Secondly the data can only be supplied by people who have smartphones. This is also a self-selecting section of society – people with smartphones are more likely to be wealthy, for example.

A significant advantage of data exhaust is that it is unlikely to have been commoditised. It can also provide insights because of its accidental nature.

### Try crowdsourcing your data

Crowdsourcing is utilising a network of people outside your organisation to achieve your aims. It can cover a wide range of activities, with both conscious and unconscious approaches.

Conscious crowdsourcing might encourage people to provide data they had generated, or to participate in activities on your behalf. If people have access to smartphones or computers they could provide a lot of valuable information on data that you would not otherwise have access to. Often this can get you around barriers put in place by companies or government who don't want the data released.

### Help people measure themselves

'Quantified self' is a movement of individuals who are interested in using technology to

If people have access to smartphones or computers they could provide a lot of valuable information on data that you would not otherwise have access to

analyse their own behaviour. By tracking information about themselves, and potentially being able to connect it to information about other similar people, they are able to better manage their lives. Quantified self can be as simple as recording the calories you consume, but its value becomes greater when you link together multiple sources of information and use behavioural data. For example, linking movement data tracked by a smartphone app to diet and exercise data.

### Look at open data from both sides

Open data refers to datasets that have been made publically available. Most often this is government data, and it can be used to add information to existing analyses.

If data isn't open what can you do? Firstly you should ask the government to open up the data[27] (and you should encourage them to do it in a consistent and meaningful way).

If they aren't willing to do this then you should approach the Open Data Institute[28].

But there is another role that you can play – by committing to opening up your own data sets, and working together with other social organisations to provide a common data framework for analysis and understanding.

27
www.data.gov.uk/data

28
www.theodi.org

# Technologies you need to be aware of



## Storing and manipulating Big Data

The earliest drivers of Big Data were the apparent contradictions between the growing volumes of data and the capacity to store and manipulate that data. For social organisations it's likely that the absolute volume of data will not be as important as the ability to analyse it and interpret it in a reasonable timeframe.

When you are deciding on analyses one of your key decisions is in the right place to host the data. For many years there have been three main contenders: files, spreadsheets, and databases – now joined by NoSQL and Hadoop as fourth and fifth.

Each has some advantages and disadvantages. But one of the significant elements of Big Data is the likelihood that you will require your storage technology to also allow you to process data.

Of course there is no need to have a single approach – what works well for the financial team might not be appropriate for the data scientist team.

### Files

The default approach might be to keep your data as a series of files. The main advantage of this approach is interoperability – assuming the files are in a common format then they should be easy to read into a number of different analytical tools. However there are some serious drawbacks to a file only approach. Files don't have any native analytical capability, errors can corrupt the entire dataset, and security relies entirely on your general IT security – there is no traceability.

Common file formats include .csv and .txt (comma separated variable and text format).

### Spreadsheets

Without doubt the most commonly used and abused mechanism for data analysis, spreadsheets were one of the first applications in the modern computer age. Data is presented in tables of rows and columns, and it is possible to perform limited analysis

within the software (including generating graphs and charts).

Spreadsheets are often ideal for smaller data volumes, and where you are interested in analysing a single dimension at a time. They aren't able to cope with complex data models, and have limited security options (although more than files). There is also another concern with spreadsheets – they often contain errors, and these errors can persist for a long time.

The most famous spreadsheet programme is Microsoft Excel.

### Databases

A database is a natural extension of a spreadsheet. Like in a spreadsheet, data is organised in tables of rows and columns. However the key difference is that in databases tables are linked together in a data model. This specifies the relationships between tables: a persons table (that describes people) can be linked to an address table (that describes locations). The link might be one-to-one, or in this case many-to-many (several people can be associated with one address, and one person can be associated with several addresses). This is both the power and weakness of databases. You need to be able to decompose the relationships in order to use the data, and sometimes you need to simplify them in order to get the database to respond in a reasonable time.

Databases have a standardised language for access known as SQL: structured query language. SQL is limited by its method of working. It does not generally permit you to iterate through the same data multiple times in the way that a programmatic approach can. For many applications this is acceptable, but for some, a more complex approach is necessary. This has led to the rise of NoSQL techniques and Hadoop.

Another advantage of databases is that they usually have inbuilt security and robustness. It doesn't mean that errors can't happen, but it should reduce them.

Commercial databases available include: Microsoft Access, Microsoft SQL Server, Oracle, IBM DB/2, Teradata and more.

## NoSQL/NOSQL

NoSQL/NOSQL is an approach to data analysis that tries to break away from the database dominated analysis that requires extensive data modelling. SQL has some significant additional limitations. It is fundamentally a single pass approach to data – it can perform actions as it reads the data, but is poor at iterative analysis. This creates problems for highly processor intensive analyses.

There are two slightly different approaches to this – the original was No SQL, an approach that actively disregarded SQL and sought newer approaches.

NOSQL took a slightly softer approach – the acronym stands for 'Not Only SQL' – and conceded that in certain circumstances SQL could be useful.

NoSQL databases are optimised for throughput , and tend to have less focus on consistency, allowing for more flexible approaches than traditional databases.

Examples of NoSQL systems include:

- MongoDB (**www.mongodb.org**)
- Cassandra (**http://cassandra.apache.org**)
- Couchbase (**www.couchbase.com**)
- Hbase (**http://hbase.apache.org**)
- Voldemort (**www.project-voldemort.com/voldemort**)

## Hadoop

Apache Hadoop (**http://hadoop.apache.org**) is a system that has been defined to handle very large datasets, on cheap commoditised hardware, in a way that avoids some of the pitfalls of databases. Some data is hard to put into tables, and when this

happens you have to make compromises in the database world. Hadoop has also been designed explicitly to allow analysis on the data without removing it from the system. Although some databases can do this they are often limited to SQL functions.

Hadoop has a system that allows analysis to be parallelised (relatively) easily using a technique called MapReduce. Essentially this maps the problem into parallel units, shuffles data (if necessary) and then reduces the parallel answer set to a single answer. It allows you to do things programmatically that are difficult to do in SQL and impossible to do in a spreadsheet. Hadoop is also an open source project.

Hadoop is often described as being part of a software 'stack'. This is simply a set of related pieces of software that are used to extend the usefulness of Hadoop.

- HDFS (Hadoop Distributed File System) – the storage system for Hadoop
- Hadoop MapReduce – the MapReduce implementation
- Pig (**http://pig.apache.org**) – a framework and language for submitting MapReduce jobs
- Hive (**http://hive.apache.org**) – database implementation on top of HDFS
- ZooKeeper (**http://zookeeper.apache.org**) – a coordination service for distributed jobs
- HBase (**http://hbase.apache.org**)– built on top of HDFS and MapReduce this is a column oriented database
- Mahout (**http://mahout.apache.org**) – algorithmic library for machine learning on Hadoop

Hadoop is available in a number of distributions: collections of code that are curated by specific groups or companies. This ensures a degree of consistency and interoperability amongst the stack. It has also allowed these companies to introduce their own proprietary elements to the system.

### Hardware

The separation of software and hardware has allowed most of the storage approaches above to be independent of the underlying hardware and operating systems.

Some database technologies tightly integrate the software and hardware in order to optimise performance, although this is usually at both a price and performance level that is outside the scope of most third sector organisations.

Some database technologies tightly integrate the software and hardware in order to optimise performance, although this is usually at both a price and performance level that is outside the scope of most third sector organisations

The creation of Hadoop and other NoSQL approaches was driven significantly by the need to use large numbers of relatively cheap commodity computers rather than a few purpose built, and therefore more expensive, machines. This has now become standard.

Another approach that could be cost effective is Cloud Computing. Services like Amazon EC2 allow effective short-term capacity that can be ideal for discovery tasks.

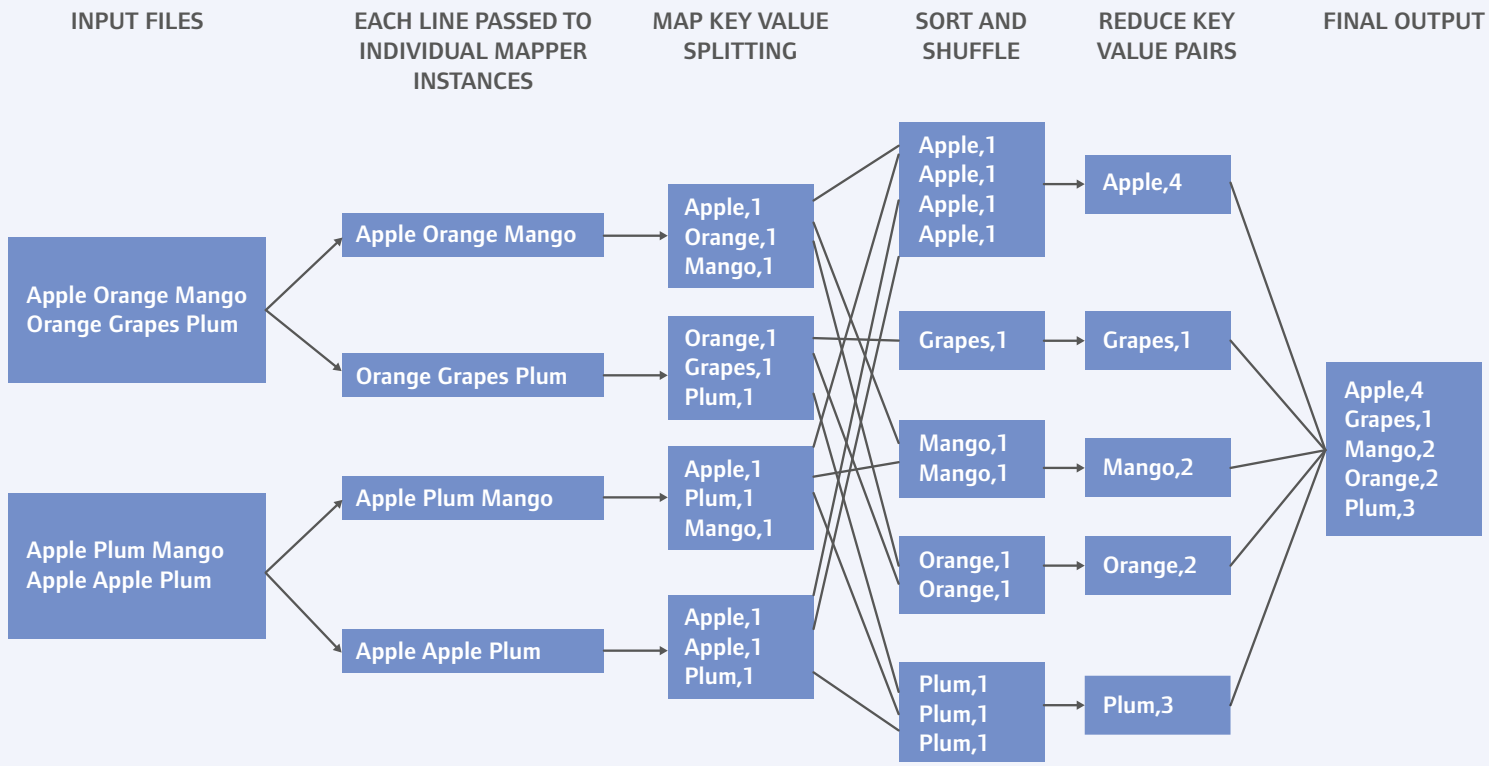
## Accessing and interpreting data

As important as data storage is, we only store Big Data because we want to make use of it. The ways we do that are critical to our success.

### MapReduce

Although we introduced MapReduce as a data storage technology, it's really interesting when used as a way of accessing and interpreting the data stored in Hadoop systems (see Figure 7). Its parallel capabilities make it very powerful when dealing with large data sets that can be distributed.

## Fig 7 - MapReduce process flow



The biggest issue with the MapReduce framework is clearly its reliance on low level programming skills to allow you to perform tasks. This can make tasks more difficult, but the availability of a wide range of publically accessible code in the open source community means that there may be considerably more flexibility than with other systems.

MapReduce is especially useful for analysis of complex data types.

**Visualisation**

Visualisation can be a key way to interpret the data that you have stored. Most analytical toolsets have some ability to visualise data, but often these are limited in their capabilities, or in their beauty.

Most analytical toolsets have some ability to visualise data, but often these are limited in their capabilities, or in their beauty

Traditionally specialist visualisation tools tended to be limited to generating reports and dashboards – the market segment known as Business Intelligence (BI). BI tools usually required considerable work, and often required dedicated data structures known as 'cubes'.

Recently a crop of new tools have come along that offer several advantages. They usually don't require data to be preformatted, and they are interactive. This interactivity supports a key element of the data scientists' work – the need to experiment with the data. Examples of tools in this category are Spotfire and Tableau.

Another approach that is gaining traction is taken by frameworks like D3.js. This is a coding approach that allows you to create powerful visualisations as dynamic web pages directly from Big Data.

**Data mining tools**

The last major technological category are the data mining toolsets.

Data mining approaches are increasingly being built into more general tools. Visualisation tools, BI tools, MapReduce, even Excel can perform some data mining.

But there is still a market for more specialised systems.

The aim of data mining tools is to help you turn data into predictions and classifications.

Examples of toolsets include:

- SAS (**www.sas.com/technologies/analytics**)
- IBM SPSS (**http://www-01.ibm.com/software/analytics/spss**)
- R (**http://cran.r-project.org**)
- MatLab (**www.mathworks.com/products/matlab**)
- Statistica (**www.statsoft.com**)

The toolset that has made the most impact in the Big Data space is probably R, an open-source project, available from CRAN. R is designed to allow you to perform advanced statistical analysis and data mining tasks.

R has gained a significant following, and is comparable to commercial software from SAS and IBM SPSS in its breadth. There are hundreds of algorithms available; allowing you to do everything from creating simple graphs to generating and scoring sophisticated statistical models.

## What is 'open source'?

Open source software is provided free, as source code, under a licence that allows users to use, research and further develop the software. Often it is generated by open groups – and it is this that is thought to allow it to produce better quality software more rapidly than commercial organisations can.

Open source software is free to acquire, but it isn't necessarily free to support or implement, as this has to be done by the organisation acquiring it. If you plan on using open source software you should plan appropriate resources for this.

# Case study: Building a Big Data organisation



Hannah Underwood is the CEO of Keyfund[29], a charity based in the North East of England that works with young people not in employment, education, or training. They invite young people to propose projects based around twelve key skills, in order to help them realize their inherent talents (see Figure 8).

Keyfund was founded 20 years ago, but is relatively small, and works in coordination with other agencies, as well as directly with the young people it supports.

**Why is Keyfund interested in data?**

Because we have a lot of it. We started our whole impact analysis strategy with the objective of using it to improve what we do, rather than as just a way of finding evidence for funders to see what had worked.

This is what I call the "improve not prove" model.

Most charities that are starting out on the path to using data are focusing on finding evidence for funders. We were quite lucky in that we already had our own approach – by default we were able to prove that what we were doing worked, but it also meant that we were capturing information that was useful for us. It can be uncomfortable sometimes, finding out that bits of what we do isn't very effective. But then why are we spending our money on it? Lets spend our money on what's effective.

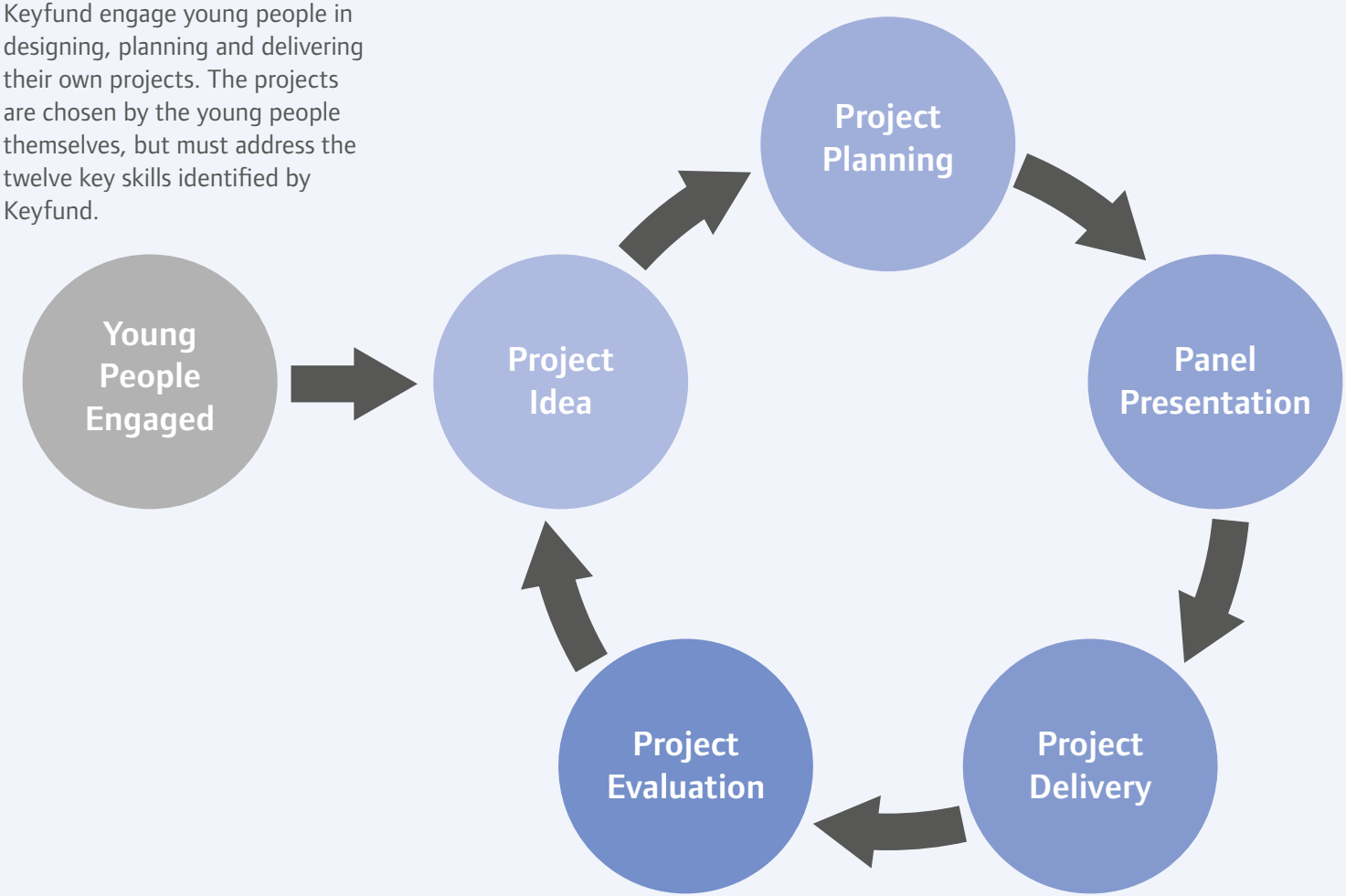
**You had a head start because you were there with a science background – how would you recommend other people start out?**

I think it does take someone interested and passionate about data to drive the agenda. With my science background I wanted to know that what I was doing was making a difference – otherwise I could have spent my time doing something else. I wasn't interested in merely sustaining the charity, or of saying "I know what works because I can see it in the faces of the kids".

29 http://keyfund.org.uk

## Fig 8 - How Keyfund works

Keyfund engage young people in designing, planning and delivering their own projects. The projects are chosen by the young people themselves, but must address the twelve key skills identified by Keyfund.



When we started there were only four of us. I'm not someone who understands databases and so on, but we managed to take the initial steps by getting pro-bono support from other organisations and then absorbing or translating those skills.

We started with process improvements, looking at where we collect the data. We got funding to build a web-based data hub, but it was so challenging because of what we didn't know – we knew what we wanted it to do, and what we wanted it to tell us, but that was it. Even being able to write an effective brief was beyond us.

**What have been the biggest successes in terms of data analytics?**

The biggest success was the opportunity to attend a DataDive (a data-focused hackathon[30]) – that was a huge turning point for us. Massive. At that point although we knew we had lots of data, we were only just scratching the surface on analysis. We've implemented so many things on the back of what came out of that.

One question we addressed was about the four stages we take our kids through. We found that the difference in the learning that took place between stages two and three wasn't enough to justify the money spent on having both stages. But we also learnt that we didn't have the data to investigate it robustly enough to make a decision on which stage should stay and which should go. We needed to expand the recording of input into our skills wheel – we were only collecting information at the very beginning of our intervention and then at the middle of each stage – so we couldn't decipher if the learning was predominately happening at the start of a new stage or the end of the previous stage. As a result we introduced a third new measurement point. In a few months' time, we hope to have enough new data to make a decision.

We also want to get to the point where when facilitators and youth orgs log on they can access impact information on themselves. So if you're a youth organization that has four facilitators we can give them information on their overall impact that they can use for workforce development – we can say you need to offer more support here, or here's an area for development.

30 http://en.wikipedia.org/wiki/Hackathon

**Would you describe your data as 'Big Data', and have you thought about supplementing it with data from other sources?**

Not with a capital B! We have between 4-6,000 people a year passing through our charity, and for each of those we collect data and then add demographic data… so yes, we try to triangulate our data with other datasets. For example I wanted to know if there was a relationship between projects and individual's social backgrounds. Was there a correlation between young people doing leisure projects and deprivation? We looked at taking postcode data, and translated it to latitude and longitude so that we could plot it on Google Maps. From the data visualisation perspective, being able to show stakeholders stuff visually really helps tell the story. Why do these things cluster? Why are the kids all coming from that area?

**You've mentioned visualization – what other analytics would you like to do?**

Well I've got a number of different questions that I want to ask, but not a lot of experience. Certainly I'd like to be able to drill down on specific areas. For example, we realized that young people who assessed themselves as between one and two on the baseline skills assessment were far less likely to finish the course. So the kids who need the help the most feel the least confident, but I can't work out who they are, if there are any similarities… So we'd like to predict where these people are, and see if we can provide more support and advice.

# Key things to make Big Data succeed



There are many different views on what it will take to make Big Data initiatives succeed in organisations[31, 32, 33], but fortunately a number of key themes come up in all of them. Although none are specifically aimed at charities or not-for-profits, there is good reason to pay attention to what they are saying.

## Put business at the heart of Big Data

Almost all of them are united in saying that the first, and most important aspect is that the heart of Big Data has to be the business. In the context of charities this means that your charitable objectives have to be the driver of the data science, not the other way around.

This can be especially difficult when in the middle of a cool, exciting new project. It is always worth asking yourselves 'How does this link to what we're trying to do?' Another aspect of this is the willingness to take different actions if the data suggests that this would be the right thing to do. This can be a very challenging position to take, as it might result in an admission that what you have been doing is not the most effective use of resources.

In the world of analytical startups this type of position is common. As Alistair Croll puts it in Lean Analytics. 'Entrepreneurs are particularly good at lying to themselves… after all, you need to convince others that something is true in the absence of good, hard evidence. You need believers to take a leap of faith with you.' What Alistair goes on to describe are the ways in which entrepreneurs can use Big Data to gather evidence and test their theories, and when the evidence shows that the theory is wrong they 'pivot' – find a new direction for their service or company.

When a charity does this it is far more challenging, as there is a need to keep clients and funders on-board. However, if the charitable objective of the organisation isn't being met (and if the data provides evidence for this) then you need to find another way forward. Experimentation, supported by data, can help.

'Entrepreneurs are particularly good at lying to themselves… after all, you need to convince others that something is true in the absence of good, hard evidence. You need believers to take a leap of faith with you'

31 www.infoworld.com/d/business-intelligence/5-strategic-tips-avoiding-big-data-bust-215296

32 www.mckinsey.com/insights/marketing_sales/putting_big_data_and_advanced_analytics_to_work

33 www.allanalytics.com/author.asp?section_id=1500&doc_id=240439

### Executive leadership

One thing that is common across all the organisations that are succeeding with Big Data, in both the commercial and charity sectors is that they have strong leaders who believe in the power of data. This is a critical part of having business at the heart of your Big Data strategy.

Simply put, without this involvement the chances of a successful implementation of a Big Data project decrease significantly.

One thing that is common across all the organisations that are succeeding with Big Data, in both the commercial and charity sectors is that they have strong leaders who believe in the power of data

### Understanding the problem, and the question

Data science is a discovery mechanism as much as a problem solving mechanism. So it isn't necessary to have a perfect definition of the final analytical question at the start of your Big Data journey. But it is important to know and understand the problem space, and to do so from an analytical perspective.

For charities this means understanding what a successful outcome would look like, and being aware of the potential differential impact of Type I and Type II errors, essentially a mathematical link to your willingness to take action. This also presupposes that there will be a way of measuring the impact of your actions.

### Checklist

- ○ Are my executives committed to using Big Data to change actions?
- ○ Are we willing to change our actions as a result of Big Data learnings?
- ○ Is my organisation capable of measuring the results?
- ○ Will the use of analytics worry or scare funders or clients, and can we mitigate that risk?
- ○ Do I understand the business problem sufficiently from an analytical perspective?

## Data focus

It may seem strange that in a Big Data publication data wasn't the first concern for success. Data is important – vital even – but there is an assumption that virtually every organisation will have access, or will be able to gain access, to some data. The data success factors are therefore somewhat more specific.

### Data collection and discovery

Big Data collection requires that processes are in place to gather the data you need for your initial analysis, any subsequent data discovery steps, and that these processes are repeatable for when you productionise any learnings.

Collection that relies on third parties for the source data (for example when data scraping from the web, or relying on surveys), needs to be well understood, and mechanisms put in place to handle changes in formats. This is likely to be the case for many third sector organisations.

Another consideration is the 'when' of data. There can often be a gap between the time that data refers to and when it becomes available for use. The best known example of this is the Census, where results from the March 2011 survey were first released in December 2012.

### Data governance and metadata

Once data has been sourced it is important that it is well documented, and that it is carefully managed. This will prevent data from becoming unusable, and will also ensure that you have the ability to go back to the data if your results and innovations are challenged.

Metadata is the data that describes data. With Big Data technologies, which tend to be less structured than traditional databases, metadata becomes even more important as a way of automatically documenting data that is being used.

Once data has been sourced it is important that it is well documented, and that it is carefully managed

Another consideration in the third sector is the importance of coming to consistent definitions within your data. This is especially important when working across multiple agencies or presenting results back to funders or to other interested parties.

### Privacy and consent

For many charities this may not be an issue, but the reputational risk of getting it wrong can be especially onerous in the third sector. People may (generally) expect companies to behave in the interests of shareholders first, and only after this to consider others. Charities are held to higher standards of behaviour.

### Accuracy and quality

We have already discussed the importance of data quality, but also emphasised the need to take this in context. Data needs to be sufficiently accurate for the range of problems you are trying to address, and you should not spend time, or money, making it over-accurate.

Data quality is another area where best practices for charities would include documenting both data quality issues and what has been done to address those issues.

### Checklist

- ○ Can I effectively collect Big Data when (and over the timespans) that I require?
- ○ Is my data well curated, or will I put in place ways of doing this?
- ○ Do I have consent for use of personal data?
- ○ Is the data quality appropriate for the actions I need to take?

## People and organisations

The third area for success is in the area of people. We've already identified the importance of executive support, but what about other organisational issues?

In practice our first problem, especially in the third sector, will be identifying the right people.

### Data scientists

One of the effects of the multiple technologies and techniques involved in Big Data and data science is that many people see a shortage in the number of data scientists. The McKinsey Global Institute released a report in June 2011 entitled: 'Big Data: The next frontier for innovation, competition, and productivity'. In it they argued that by 2018 there would be a significant shortfall in data scientists. Talking about the US only they said: '…we project that demand for deep analytical positions in a Big Data world could exceed the supply being produced on current trends by 140,000 to 190,000 positions.'

This represents a 50-60% talent gap. The knock on effect for charities is obvious. These people are rare, and as a result their salaries are significantly higher than most charities would be able to afford. If this is the case then to see the maximum benefits from Big Data charities will have to find other ways of identifying and attracting data scientists.

More recently the O'Reilly group have surveyed data scientists to discover how they would self describe, and come up with four key categories: data businesspeople, data creative, data developers and data researchers[34]. The O'Reilly approach runs the risk of being a self-selecting sample, and with a relatively low sample size (n=250) – but it has, at least, tried to apply data science techniques to the problem.

So what exactly is the role? In fact, as identified by O'Reilly, there are several distinct skillsets and roles that are involved in data science, and although a few people may exist who excel in all of them, in most cases data scientists have a distinct specialism.

34 www.fastcolabs.com/3008621/tracking/what-kind-of-data-scientist-are-you

This means that in identifying data scientists for your organisation you need to be very clear about the nature of the problem you are addressing, and how the various skills relate to that problem.

### Technical data scientist

The technical data scientist is deeply comfortable with both the hardware and software of Big Data. They will be able to understand the intricacies of installing and maintaining the hardware/software stack, and writing and adding code relevant to your infrastructure. At a minimum they will be Linux experts and will often have skills in MapReduce, Python and other code bases, as well as the fundamentals of running Hadoop clusters.

As well as providing the infrastructure for analysis they may be responsible for the implementation of any discoveries in a usable and reusable system.

Where technical data scientists often fall down is in their ability to understand the relative importance of data and how to extract relevant actionable knowledge from it.

### Data miner/researcher

Data mining as a term dates back to the mid 1990s, when computing speeds allowed machine learning techniques to sensibly tackle more sizable datasets. They are at home doing statistical analysis on your data, and should be capable of identifying significant trends and producing predictions. They have to spend considerable time preparing data for analysis, and so may have some of the coding skills of the technical data scientist.

Most of the toolsets used by data miners hide the more extreme coding, allowing the data miner to focus on the analysis. Data miners may need support in documenting and presenting their results to wider audiences, especially where issues stray away from analytical and into the business sphere.

### Visualisers/storytellers

Visualisation is a key aspect of turning a Big Data analysis into actions. Visualisers are

able to take seeming abstract models or discoveries and convert them into patterns or anti-patterns that will make sense in a business context. This is a vital skill when negotiating between funders, service providers and users.

**Business experts**

At least at the beginning of your Big Data journey you will have to find ways of ensuring that the data science team is led by your business, rather than allowing them to proceed in directions that may not be productive. This is really crucial in preventing your ideas getting hung up on the first exciting insight – even if it is something that is already well known by practitioners in the field.

You will have to find ways of ensuring that the data science team is led by your business, rather than allowing them to proceed in directions that may not be productive

## Hiring data scientists

The shortage of data scientists gives us a problem, and you will need to decide how your organisation can address this. The two routes open to commercial organisations are:

- recruit specific data scientists externally,
- identify business experts and encourage them to grow into the role.

But, as Hannah Underwood and HyeSook Chung found, as a charity you don't need to recruit your own data scientists for everything you do. There is a growing field of 'data philanthropy' where volunteers donate their skills to help charities.

There are several organisations that exist to help attract and coordinate these data heroes. One of these is DataKind[35], originally founded in the US by Jake Porway, and its UK equivalent DataKind UK[36]. DataKind acts as a conduit for data scientists to directly donate their skills.

35 http://datakind.org

36 http://datakind.org.uk

Another organisation that works in this area is the Open Knowledge Foundation[37]. With a greater emphasis on open data and education, it is ideal for organisations that are starting out into the open data space.

Timebank[38] has a more general catchment, but a correspondingly large number of potential volunteers – even if many of them aren't data scientists.

### Organisational structure

As a charity it is unlikely that you will have huge numbers of data scientists, so the best model for deploying them is almost certainly as a central team. The most successful data science teams are ones that foster collaboration and where the team is able to work across the whole organisation. Communication is also a key factor – both within the team, and crucially with your practitioners and field staff. A data scientist who understands how results might be used will be far more effective than one who doesn't.

If you are working with volunteers you will also have additional challenges: how to integrate them into your team, how to deal with absences, how to ensure that data protection guidelines are implemented.

### Retention

It's one thing to recruit data scientists, it's another thing to retain them. Hopefully if you have got this far you will be in a good position to do so. You will have integrated data into your business, have appropriate Big Data problems (and the data to go with it), and above all be putting the data scientists skills to use in an area that makes the world a better place.

But sometimes people will move on, and this is even more likely when it comes to volunteers. In these cases it is crucial that the work that the data scientists are doing has been well documented, so that others can pick up where they leave off.

37 http://okfn.org

38 http://timebank.org.uk

Typically this documentation will include metadata, descriptions, outputs (both graphical and data), and code. It is also far easier to document things as a project proceeds rather than waiting until the end.

### Checklist

- ○ Do I have access to relevant data science experience?
- ○ If not, can I recruit or develop it?
- ○ Does my data science team cover the key roles of data science?
- ○ Are procedures in place to retain data scientists and their knowledge?

## Getting the technology right

### Technology follows the problem

> "The answer is Hadoop. Now, what was the problem?" Anon

Whenever the technology leads the business you are storing up problems for later. Ideally you should start with an assessment of what you want (or are likely to want) to do, and from there make your decisions about the correct technologies to use.

Good reasons for choosing a technology include:

- cultural and experience fit with your organisation,
- match to future analytical and data needs,
- evidence of a wide user base in the sector (many people to pool experience),

- long term cost profile.

Bad reasons for choosing a technology can include:

- short term cost profile,
- match to past analytical needs,
- the expertise of a single user (unless you are convinced that they will never leave).

**Understand the advantages and disadvantages of open source**

The open source movement has been a huge factor in the success of the Big Data movement, but open source software may not be the right choice for your organisation. Today open source will require more internal support than commercial software.

This is changing slowly. The open source stacks are being packaged into more accessible and better supported formats – in fact this is a new business model for companies such as Cloudera, Hortonworks, 10gen and others.

Open source advocates will also rightly point to the more rapid development cycles of the open source community – if there is a problem with the software today there may be a fix tomorrow.

**Ask for help**

It's likely that you are not the expert in Big Data technology, which as we've seen can be very confusing. As Keyfund found out, sometimes you may not even know what the right questions are to ask. Fortunately this is another area where volunteers, especially commercial organisations, can provide assistance. In the case of Keyfund part of that advice was in creating a requirements specification in order to be able to identify the right technology.

## Checklist

- ○ Do I have an idea of my future analytical needs?
- ○ Does the technology match my user base needs?
- ○ Will the technology be cost effective in the long term?
- ○ Can I identify external resources that can help me with decisions?

# Appendix I
# Selected reading



**The Signal and the Noise** – Nate Silver. Nate Silver made his name predicting baseball, then made it slightly more well known by predicting the US Presidential Election results in 2008 and 2012. He took on the pundits in the 2012 election and won. This book is a great guide about the why of analytics.

**Lean Analytics** – Alastair Croll and Benjamin Yoskovitz. Lean analytics is really a book for people interested in creating a data-led startup business, or for people who think that these techniques would apply equally well in the social sector.

**Big Data Now** – O'Reilly Media. O'Reilly have made a name for themselves with the Strata conferences focusing on Big Data: **www.strataconf.com**. They also publish this anthology of the state of play in Big Data.

**The Guardian Data Blog:** **www.theguardian.com/news/datablog**. The Guardian almost invented data journalism, an approach that is now widespread. The data blog continues its mission to make data more available.

When **UN Global Pulse** was launched it's Director, Robert Kirkpatrick wrote this article to explain what data philanthropy could mean: **http://online.liebertpub.com/doi/pdf/10.1089/big.2012.1502**

**The Economist** has been following Big Data closely. Here they examine Mastodon C's work on NHS prescription data and statins: **www.economist.com/news/britain/21567980-how-scrutiny-freely-available-data-might-save-nhs-money-beggar-thy-neighbour**

Ken Cuckier of The Economist has also written a book, **Big Data: A revolution that will transform how we live, work, and think**.

**Sloan Review** on data science success: **http://sloanreview.mit.edu/article/organizational-alignment-is-key-to-big-data-success**

**Forbes** on fail fast: **www.forbes.com/sites/nyuentrepreneurschallenge/2012/10/16/fail-fast-succeed-faster**

**NYTimes** on UN Global Pulse: **www.nytimes.com/2013/08/08/technology/development-groups-tap-big-data-to-direct-humanitarian-aid.html**

**Data Science 101: http://datascience101.wordpress.com** A good general data science resource.

The following articles all describe various activities of **DataKind:**

- at the UN: **www.meetup.com/DataKind-NYC/events/99386112**,
- at the World Bank: **http://blogs.worldbank.org/opendata/scenes-from-a-dive-what-s-big-data-got-to-do-with-fighting-poverty-and-fraud**,
- and the project with Oxfam GB: **www.oxfamblogs.org/fp2p**

Finally the review on Big Data from **McKinsey** that woke the world up in June 2011: **www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation**

# Appendix II
# Where can I find data?



The following list of sources should provide a practical starting point to identifying some of the external data sources that are available to you. Some of this data is available as open data, or as free data. Other data sources are private, or have commercial terms associated with its use and reuse.

Of course one of the best sources of data will be your own data, and that of other social organisations in the same or related sectors.

## Geographic and related data

The most detailed UK data on geography is that owned by Ordnance Survey. Unfortunately Ordnance Survey is one of the most difficult organisations to deal with in terms of commercial data, with complex and unwieldy licensing arrangements. However. they have created a limited set of open data that can be used more easily: **www.ordnancesurvey.co.uk/business-and-government/products/opendata-products.html**

If you are interested in weather data then there are a number of options for sourcing it, including Wunderground: **www.wunderground.com**

## Social media data

Most social media sites provide APIs that allow you to directly access data. However there are companies that will provide datasets from social media for you. Two examples are Gnip. **http://gnip.com**, and DataSift, **http://datasift.com.**

## Data markets and aggregated data

Increasingly people are creating single locations where interesting datasets are

gathered for use. Although these tend to have a US flavour they are increasingly being used to source UK specific data.

**www.infochimps.com/marketplace**

**https://datamarket.azure.com**

**http://datamarket.com**

Google also provide data on top of an exploration tool: www.google.co.uk/publicdata/directory

## Open data

The UK Government has been one of the pioneers of open data. Most of the UK Government datasets are gathered together at www.data.gov.uk, although some can be found on the websites of the individual departments (for example English Heritage). For US data see: www.data.gov

The UN have a variety of datasets available, for example information on historical world food prices: www.fao.org/worldfoodsituation/en

The World Bank is also a great source of development data: **http://databank.worldbank.org/data/home.aspx**

## Other data lists

Some data scientists have been very helpful in compiling sources of data sets.

Hilary Mason of Accel put together this list: **https://bitly.com/bundles/hmason/1**

And the BigML group put this one together: **http://bitly.com/bundles/bigmlcom/f**

## Data search

Finally there are now some specialist search engines that can help you look for specific data – one such is Quandl: **http://www.quandl.com**

# Glossary



**Agile analytics:** Fast, iterative, experimental analysis.

**Apache:** Shorthand for the Apache Software Foundation, an open source collection of 100 projects.

**Bayesian statistics:** Named after Thomas Bayes (1701-1761), this form of statistics uses prior and posterior knowledge to improve accuracy.

**Cassandra:** Open source, distributed data management system developed at Facebook.

**Causation:** A relationship between two measurements where one event causes the other.

**Classification:** Grouping things together so that they can be described as a class.

**Clickstream data:** Data logged by client or web server as users navigate a website or software system.

**Cloud:** Method of performing computing on hardware and software that is remote from you. This removes the cost of owning the hardware and software, or of being responsible for the maintenance of it.

**Control group:** Powerful measurement technique that isolates a group for comparison against the test action.

**Correlation:** A weaker relationship than causation, where a change in one measurement is reflected in a change in another measurement.

**Data curation:** Maintaining data sets, including their meaning and data quality.

**Data exhaust:** Data that is created as a by-product of a process.

**Data journalism:** Journalistic technique of data-led writing.

**Data modelling:** A formal way of understanding how a process is reflected in data.

**Data mining:** A business process that discovers patterns, or makes predictions from data sets using machine learning approaches.

**DataKind:** US and UK charities that link social organisations with volunteer data scientists.

**Data science:** The techniques for transforming Big Data into knowledge.

**Decision trees:** Classification or prediction techniques that break down data sets by a series of individually simple splits. Decision tree models have the advantage of being easier to describe than some more complex techniques.

**ETL:** Extract, transform and load (ETL) – software and process used to move data.

**EU 95/46/EC:** The core data directive in European law. Implemented in legislation by each of the member states of the EU.

**Fail fast:** Low-risk, experimental approach to Big Data innovation where failure is seen as an opportunity to learn.

**Frequentist statistics:** Traditional statistical approach that treats events as independent.

**Gamification:** Turning actions into a game.

**GIGO:** Garbage In, Garbage Out. Phrase that warns of the dangers of poor input.

**Hadoop:** Open source software controlled by the Apache Software Foundation that enables the distributed processing of large data sets across clusters of commodity servers.

**Hadoop Hive:** SQL interface to Hadoop MapReduce.

**HDFS:** Hadoop Distributed File System, the data storage layer in the Hadoop ecosystem.

**Java:** Dominant programming language developed in the 90s at Sun Microsystems. It was later used to form Hadoop and other Big Data technologies.

**Key value pairs:** A way of avoiding data modelling.

**MapReduce:** Programming paradigm that enables scalability across multiple servers in Hadoop, essentially making it easier to process information on a vast scale.

**MongoDB:** NoSQL open source document-oriented database system developed and supported by 10gen. Its name derives from the word 'humongous'.

**Moore's Law:** Trend identified by Gordon Moore (1929-) in 1965, that the number of transistors (and hence computing power) on computer processors doubles every 18 months.

**NoSQL:** Data storage approach that avoids the approaches used in relational databases, hence no SQL.

**NOSQL:** Data storage approach that combines SQL and NoSQL techniques.

**Open data:** Data, often from government, made freely available to the public.

**Open source:** Movement for releasing software under licences that allow free use and adaption.

**Personal data:** Data that uniquely identifies a living person, as defined by the Data Protection Act.

**Pig:** High-level platform for creating MapReduce programmes used with Hadoop, also from Apache.

**Prediction:** Using data to forecast future events.

**Python:** Dynamic programming language, first developed in the late 1980s.

**Random forests:** An example of an analytical technique that has become popularised by the Big Data movement. Essentially a model that uses a series of many decision trees that are each built on subsets of variables.

**Regression analysis:** A range of analytical techniques that fit data points to a line or plane in order to allow predictions to be made.

**Relational database management systems (RDBMS):** Traditional data storage technologies that utilise SQL for accessing data.

**Robots.txt:** A standard for allowing web scrapers to understand if web page owners want a page indexed.

**Social Media:** Facebook, Twitter and other internet platforms that allow users to generate content and communicate with each other.

**Social Network:** A collection of people and how they interact. May also be on social media.

**SQL:** Standard, structured query language specifically designed for managing data held in databases.

**Type I error:** A false positive.

**Type II error:** A false negative.

**Variety:** One aspect of Big Data – the number of different data formats and data types that may need to be processed.

**Velocity:** One aspect of Big Data – the speed at which data changes.

**Volume:** One aspect of Big Data – the size of data that needs to be processed.

**Web scraping:** Using software to extract data directly from web pages rather than using a published API.

**XKCD:** Important reference manual for Big Data. www.xkcd.com.

**ZooKeeper:** Another an open source Apache project which provides a centralised infra-structure and services that enable synchronisation across a cluster.

# About Nominet Trust



Digital technology enables us all to think radically differently, to stimulate new forms of collaboration and to mobilise new communities of interest to take action for social good. It offers us phenomenal opportunities to inspire the creativity and compassion of millions of users in addressing social needs.

At Nominet Trust we bring together, thoughtfully invest in and support people who use digital technology in creative ways to make society better.

All of our social investments are informed by research into current thinking and best practice. These investments are, in turn, evaluated to identify good practice. This good practice also feeds into further research on how to advance technology as a tool to mobilise positive social change, which subsequently informs new investments.

To find out more about our work or how you can apply for funding, please visit:

**www.nominettrust.org.uk**

Visit Nominet Trust's website

# About our work



There are many ways in which digital technology can bring about change. To make sure we achieve the greatest impact, our focus is on supporting projects and organisations that are using it in imaginative ways to improve lives of the disadvantaged and vulnerable and to strengthen communities.

It is important to remain open to new ideas that offer a fresh perspective. Our aim is to seek out, galvanise and support innovative, early-stage projects that use digital technology to address big social challenges.

We also invest in a number of programmes that address a specific social group or issue, such as young people, local communities or health and well-being. By clustering our investment in this way we hope to increase our social impact. We regularly review the groups of people and issues we support so please check our website to find out our current focus.

**Do you need support for your idea?**
If you have an idea for a new initiative or would like support for an existing project then please get in touch.

We are particularly interested in projects that develop tools or models that can be replicated or scaled-up to benefit others.

To find out more about how you can apply for funding, visit us at:

**www.nominettrust.org.uk**

nominettrust

Nominet Trust
Minerva House
Edmund Halley Road
Oxford Science Park
Oxford  OX4 4DQ

t  +44 (0)1865 334 000
f  +44 (0)1865 332 314
enquiries@nominettrust.org.uk
www.nominettrust.org.uk