

“Hello Surveys, meet Data Science”

What it is, potential benefits & how to get started

Afternoon. My name is Andrew Marritt and I am the founder of a Swiss HR Analytics practice called OrganizationView.

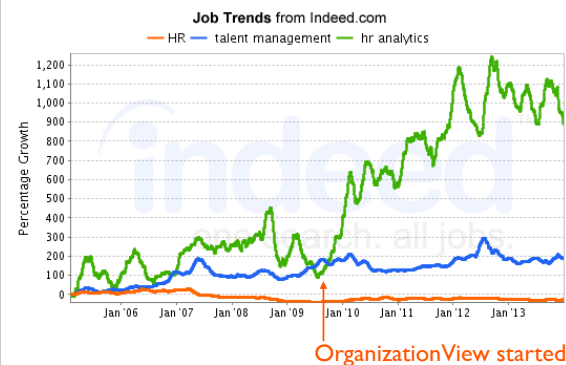
“I was looking for something like baseball, where there's a lot of data and the competition was pretty low.”

Nate Silver, Guardian interview

This quote from Nate Silver resonates with me.

My career has 4 main components. I was a bit of a geek as a child, programming computers from the late seventies. I went to uni in 1990 to read Maths and ended up doing the maths parts of economics. During the 90s I spent most of the time in the management consultancy sector and the noughties running big technology & data enabled projects in corporate HR.

Growth in analytics jobs in HR



I set up OrganizationView 5 years ago to help businesses make better decisions about their employees using data after watching HR struggling. I had seen how data had started to take over in marketing and thought it was only a matter of time before it hit HR.

Our work comprises 3 joined elements: Measurement, Modelling and Visualisation.

We're typically applying techniques that have worked in other areas - Operations Research, Marketing, Risk - to HR problems

In the last 4 years much of our work has involved surveys as a part and it's that that I wanted to talk about today.

I have to be honest. Survey clients usually come to us because our Data-Science perspective allows us to shift the economics of the survey industry - we can typically be less than half the price of an existing supplier, often much less.

Technology shifting economics of survey industry



Start getting the most from People Intelligence today.

Culture Amp is offered as an annual subscription for software and support — Capture People Intelligence all year round. All of our plans come with the great features listed below.

Monthly*	\$400	\$700	\$1,000	\$1,300	\$1,500	Contact Us
Up To	200 employees	400 employees	600 employees	800 employees	1,000 employees	1,000+

*With annual commitment. Includes 10% discount for annual payment.

Ready for People Intelligence?

Sign Up

cultureamp pricing as of 10 Sept 2014. [Source](#)

We're not the only firm doing this. There are some great engagement 'app' providers like Culture Amp with similar price levels. Our differentiation is we do 100% custom solutions for each client.

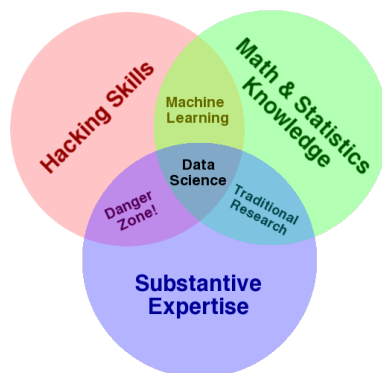
What is data science?

A side effect of what we do is getting the data foundations right. With most clients we're now doing far more sophisticated analysis than they initially envisaged.

Sometimes it's easiest to communicate what you do by providing a simple label. I suppose Data Science for HR is as good a label as any for us. Realistically I feel its a modern adaption of statistics but it neatly differentiates what we do from the sort of statistics that my clients may have learnt at university.

What do I mean by this?

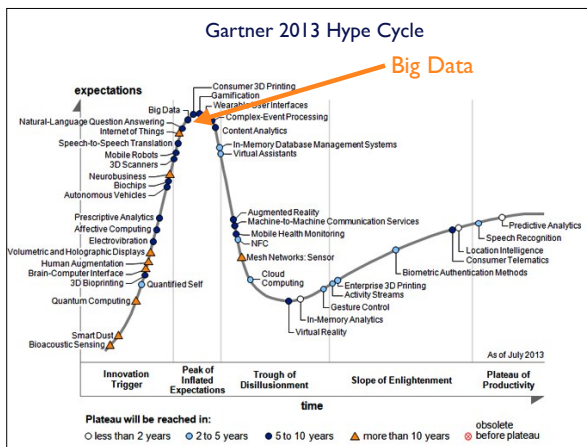
Drew Conway's Data Science Venn Diagram



Source

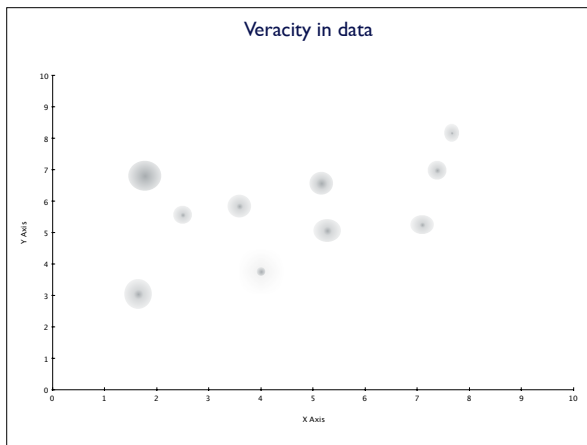
For me it's the switch to using computing power, through writing code, to make sense of the world. This famous image by Drew Conway summarises it neatly.

The other buzzword that people are using is Big Data which is now becoming so widely used that it is losing all real meaning. In fact many of the biggest names in the arena started abandoning the term a few years ago. Some have suggested that it should just be 'Data' or even 'All Data'.

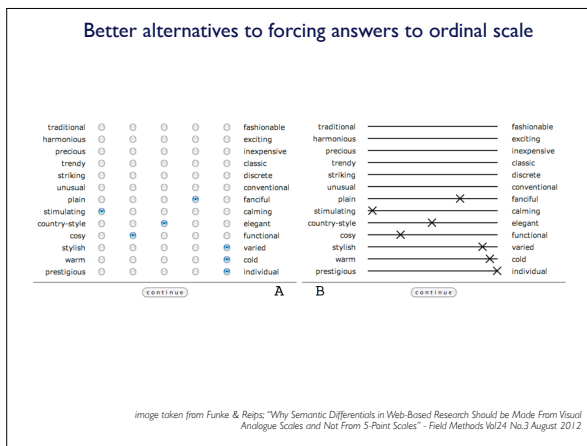


The two aspects that we do see being pain points are Variety and Veracity

Everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims that are doing it.



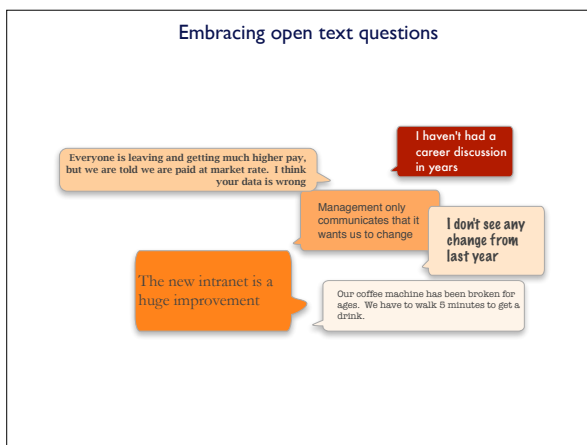
Veracity incorporates a concept which is drummed into every statistician whilst they're learning but is frequently forgotten, that is we must be extra-vigilant and questioning about the data. Veracity means that there we have uncertainty about the data. It could be incomplete, or that the measurement only loosely matches what you're really trying to measure



A good example is the Likert question. The respondent has a true feeling, for example somewhere between Strongly Disagree to Strongly Agree. What the traditional question does is 'force' them to identify the closest point given a few descriptors. When we analyse this data we have uncertainty in the data as not everyone with common feelings rounds it to the same point

Why did we use a 5 point scale? Because it was easy to score with a paper survey. In the world of modern User Interfaces we have new opportunities and challenges. I'll come onto this topic in a moment.

Variety refers to the different types of data. For engagement surveys this includes the different question types. Likert data is ordinal, other forms like checklists are probably being treated as interval.



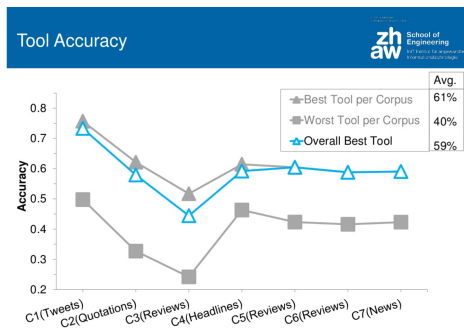
The area here where we see the most opportunity for surveys is unstructured data - mostly free text.

Historically I think a lot of people shied away from open text because it was hard to deal with. It's still not easy but text analysis is starting to come of age. Because we know we can deal with it we tend to use it more than before.

With unstructured data what you are really doing is converting from qualitative data to quantitative. For text this includes counting words or topics - Categorisation - or Classifying the text on some score like sentiment.

Sentiment analysis is a series of techniques to try and ascertain whether the person was positive or negative in their statement. It's a technically difficult subject. For example computers can't easily differentiate between a straight and sarcastic statement

Sentiment analysis has a long way to go



taken from "Bliss or Hell - How Can Computers Tell the Difference", Mark Gellieck, Associate Professor at Zurich University of Applied Sciences. Presented at Zurich Machine Learning Meetup March 27 2014.

Earlier in the year, at one of the Zurich Machine Learnings meetings I got to we had a professor who had studied the accuracy of the leading commercial tools. The maximum accuracy was about 60%. Whether this is good enough depends on what you plan to do with the data.

The joy of survey data is that we typically have other questions that should suggest how positive the respondent was. What we typically find however is that at the top end there is an inverse relationship between engagement and sentiment. Engaged employees will often write the most critical comments, arguably because they care the most.

Automated process for categorising comments



Categorisation is probably an easier problem to solve. Is this a statement about Leadership, Career Development or the Coffee Machine?

At the moment we're using a mixed approach which combines text analytics and humans. All machine learning approaches need a human element to develop what is known as a training set.

The process we use is mostly automated but has an important human role. Our R code goes off to the survey tool via their API and grabs the latest response data. It takes the open text, identifies the language and creates a mini survey with the question name, answer and a set of the most common categories. This then gets uploaded to the survey tool via the API and an email sent to a few reviewers asking them to categorise the text. When they've done this the data comes back to us. We use what is called an ensemble method - in this case voting - to finish the categorisation. You probably wouldn't believe how often smart people disagree what a paragraph is about!

We can also add in automatic machine translation of the comment into a common language - typically English. This is most useful at a global level.

With this data we're starting to train our algorithms to categorise the comments. In future only if they are not confident will we send it to the humans.

The most useful part of open text is that it enables you to answer the 'why?' something is happening. In this way it is closer to the more qualitative approaches such as focus groups.

One of our clients recently described the open text part of their survey as 'the most useful question'. I know their CEO reads every comment each time. At his last town hall he referred to someone who had started their comment with 'you probably won't read this'.

Another useful benefit is to identify areas that might need quantitative questions. This is much easier to spot if you're categorising the comments. It

John W. Tukey

EXPLORATORY DATA ANALYSIS



This brings us into another important topic in Data Science - the combination and balance between Exploratory and Confirmatory Data Analysis. Identifying a pattern in the data, the stuff that gets talked about in the press when Big Data is discussed is all as well but it rarely replaces doing some more traditional statistical testing - the confirmatory data analysis.

The experiments that Facebook were conducting recently was almost certainly them using Confirmatory Data Analysis after spotting patterns using Exploratory Data Analysis.

Identifying areas that might require further analysis by quantitative means - most notably changing questions in future surveys - is a typical Exploratory technique. We might also use data mining to find areas or groups that are behaving in a different way than expected. What you always have to remember is that in a billion row datastore you're very likely to find a one-in-a-billion-chance occurrence.



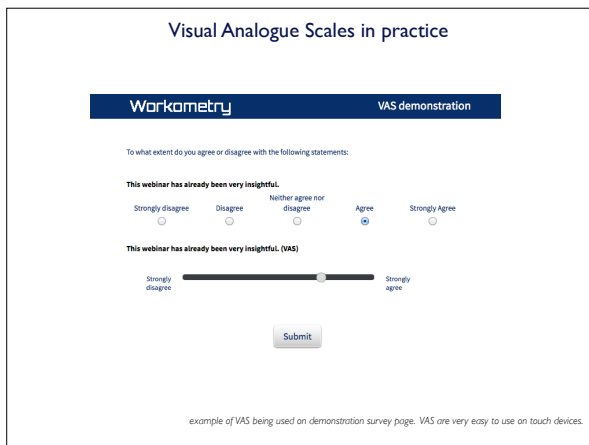
As good analysts we need to constantly challenge the way that we do things. The way that we conduct a survey hasn't changed much over the last 20 years, yet the possibilities that exist are radically different. To give you an idea of the change in user interfaces, this is Microsoft's homepage in 1994.

Many of these possibilities exist because of how technology is evolving. Most of the survey research we conduct these days is done electronically. One client is a mobile phone company and with them we're designing for smartphones first, because that is how most of their staff will access the survey. Earlier this year we did some research with the top 400 managers of a big financial services firm. Many of those managers used an iPad to access the questionnaire. You need to think about how people will use the survey on touch devices.

We're big fans of adaptive questioning, that is that the respondent gets subsequent questions based on previous answers. It's a technique used a lot in computer aided testing. This makes the survey more like a conversation. As such it arguably makes the survey more qualitative, but it enables us to explore concerns in much greater detail.



Another similar approach is to use recommendation algorithms to ask a broad set of questions to the population without asking everybody the same questions. Respondents typically answer questions in patterns. If someone answered question (a) positively they may also be likely to have done so for question (b). This is very much the same way that movie recommendation algorithms work and we've used such algorithms to 'fill the gaps' in survey data. On our tests we can maintain 63% accuracy even if we only show each person 20% of the question set. This of course means it's possible to have large sets of questions and only show respondents each a small semi-randomised selection.



I talked before about the issues when you force a respondent, who's feeling is somewhere on a continuum, to say which of 5 points they feel is the closest. The question is why do we do this, especially as we know that it is adding measurement error?

About a year ago we did some research to see whether marking a point between two extremes on a line was a good alternative, by using a slider. Technically this is called a Visual Analogue Scale.

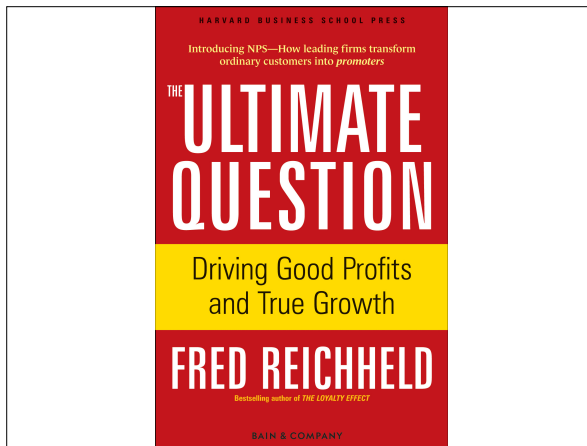
The feedback that we got from our testing was that respondents felt much more able to communicate their views accurately this way. This is what we'd expect from the theory

Furthermore our research suggested that we can count such data as interval rather than ordinal. The respondents saw the line as a linear scale. This gives us huge advantages when analysing, modelling and visualising the data.

We don't see the need for always doing a full company survey - a census - once a year. One client does a pulse survey on a quarterly basis. The other option we're starting to see more often is using a randomised sample on a regular basis - say monthly.

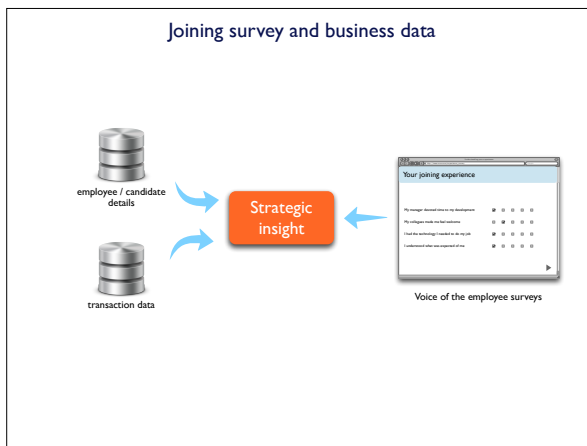
I can't think of any other business data which is only measured once a year. The trend across business at the moment is for more frequent measurement. I see no reason why engagement data should be an exception.

The other thing that I find strange, looking in to this area arguably as an outsider, is that often companies devote large resources and effort to their yearly survey but do very little measurement outside that. If we look at other functions such as marketing the balance has shifted to doing regular-but-smaller research.

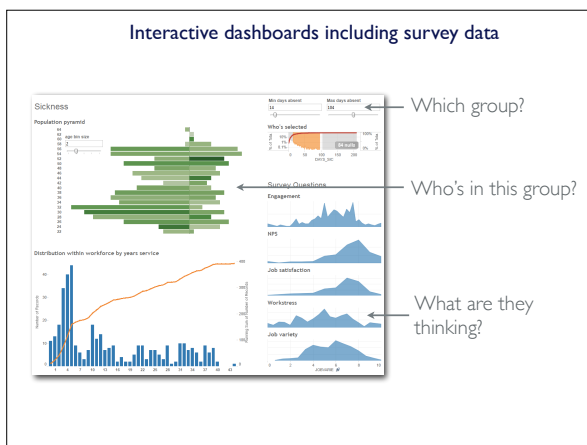


The most obvious form of this is the large Voice-of-the-Customer trend that we've seen over the last 5 to 10 years. Following the hugely influential book by Bain / Fred Reichheld - the Ultimate Question - there has been a strong focus on measuring experience at various touch points. This type of information collecting is based on when a customer has an interaction, rather than at a set time. Several large firms are applying this principle to employees with considerable success.

With one client we're blending the engagement data with the VoC data. The organisation has a large proportion of its staff in call centres, supporting existing clients. Every time after they call they receive an SMS asking them to give feedback. We're linking this with the engagement scores on an individual basis so we can analyse engagement data by levels of customer satisfaction and vice versa.



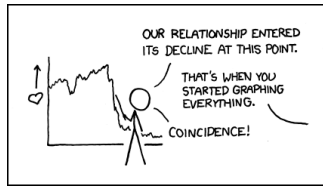
This brings us back to the shift to describe Big Data as Any or All data that I discussed earlier. The data science perspective is to think of this data as just one part of a much larger data ecosystem. Just like in the Voice-of-the-Customer arena, doing so enables us to link perceptions with behaviours. There are two sources of behaviour data. The first is data held within HR systems. This could be things such as sickness, performance, attrition but it could also be things that happens to the employee such as job change, manager change or participation in development programmes.



Treating engagement data as just another part of the HR dataset changes how we think of HR reporting. We see engagement data being used on HR dashboards, and especially on interactive dashboards. Of course this needs to be done in a way that protects confidentiality but that is relatively simple to do.

The other source of data that engagement data is typically linked to is business data. Most of our systems these days keep records of how they're being utilised - commonly called log files. These can be linked, often at the individual level. Of course this type of analysis needs to inherit the strongest level of confidentiality requirement of the source datasets, which will typically be the survey. We find that questions such as 'how do the mostly engaged employees behave compared to the rest' are revealing. This type of analysis is much more likely to drive managers to sit up and notice than more general industry findings.

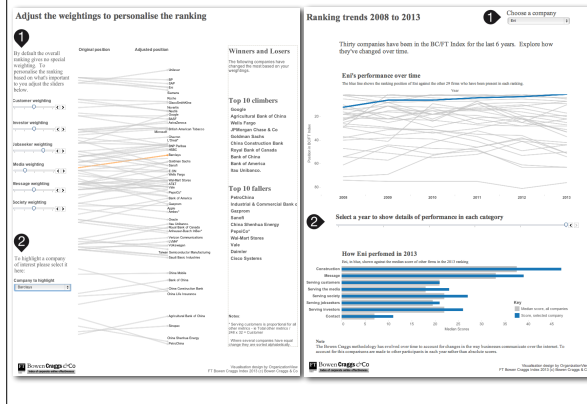
Data visualisation



"Decline" from [XKCD](#)

A key part of the whole data science movement is data visualisation. When done well visualisation can provide much greater insight into complex datasets.

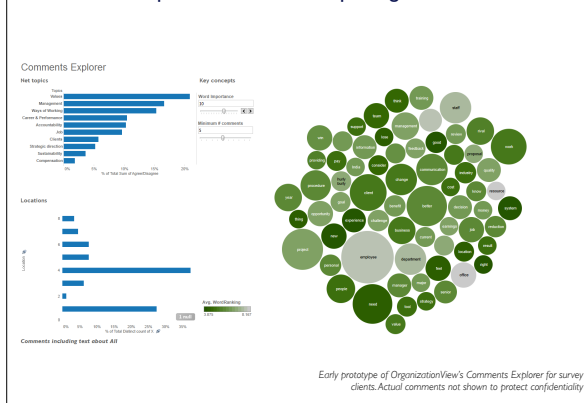
Data visualisation to help people understand and explore



If we look at the places which are doing visualisation well many of them are the online parts of major newspapers like the NYT or the UK's Guardian. This is a piece we did a year ago for the Financial Times online.

A key goal of such interaction is to enable users to personalise the data to match their needs. With engagement surveys, exploring the data interactively is probably only useful for the main analyst team. Where it becomes more useful is when engagement data is part of a larger ecosystem.

Graphical interface for exploring comments



Early prototype of OrganizationView's Comments Explorer for survey clients. Actual comments not shown to protect confidentiality

I think we all probably know word clouds. We tend to resist using them where possible because of the issues with perception - that is long words will look more important than shorter words. For our comment navigator we use bubbles. Clicking on any bubble will filter the other questions or demographics to people who discussed this concept. It will show where they are located, the scores in the other questions and which other concepts were discussed. Suddenly thousands of comments can be analysed for patterns in an insightful manner in real time.

Funnel plots to identify outliers

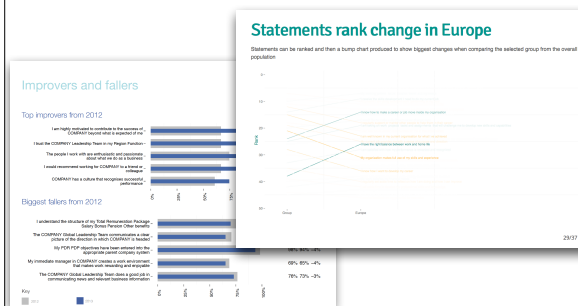


[Paper about funnel plots](#)
[Post on data-visualisation techniques](#)

One visualisation type that I think all analysts should use is the funnel plot. Funnel plots are great way of differentiating between different groups whilst accounting for their differing population sizes. They became popular after David Spiegelhalter, highly influential Cambridge statistician, wrote a great paper recommending them instead of league tables for comparing health performance data in the UK.

There are several other forms of data visualisation, rarely used that are useful which looking at the multivariate data sets such as survey data. I've included a link to a good introductory paper by one of the top academics in this field.

Automated reports



The biggest area many firms would initially realise gains from data scientists in their engagement midst is by automating reporting.

The great thing about computers is that they're really quick, don't demand work-life-balance and never get bored.

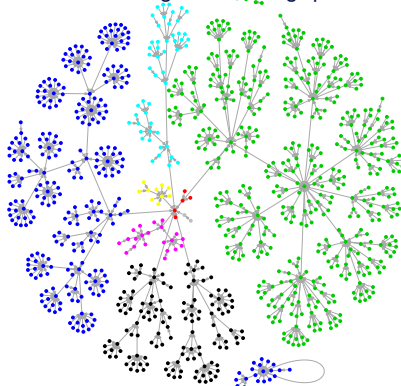
Typically firms are producing manager level reports based on standard templates. This is a perfect application for coding the process.

To give an example, during the last survey we had 110 manager reports ready for a client within an hour of the survey closing. There is no reason we couldn't have got the computer to do this at the middle of the night, emailing the managers the report directly.

We have applied technology that is used in wealth management to produce individual asset reports for clients. The quality we can get far exceeds what is possible from something like SPSS or Excel. A head of HR last week told me that this had increased the trust that the executive team had in the data.

This reduction in time taken, and the associated FTE requirement is what enables us to dramatically reduce the price.

The organisation as a graph

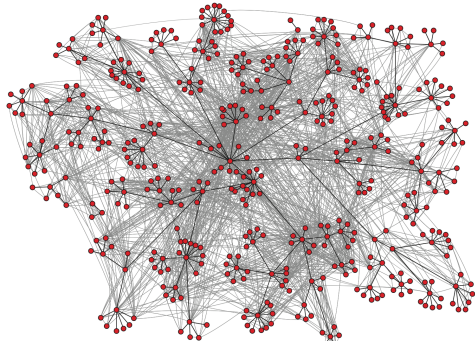


Hand in hand with this automated approach is another classic data-science technique - breaking from the 'standard' approach of using data tables to what is often called 'NoSQL' approaches - in our case storing and querying data in something called a graph.

Graphs are the mathematic name for what is mostly commonly called networks. Graphs, and graph databases underpin some of the biggest information firms such as Twitter, LinkedIn and Deutsche Telekom.

Instead of storing the data in a big table we use networks that link the individuals together. Here you can see the organisation chart of one of our clients plotted using network analysis code in R. This is the simple reporting structure but we can add lots more information without increasing complexity. For example we could add dotted-line reporting, involvement on shared projects or development programmes.

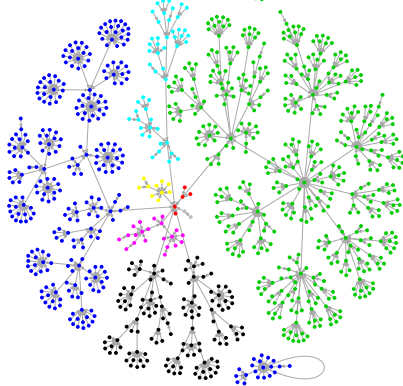
Adding email traffic to an organization graph



We could add in data from communication channels such as the flow of emails between the individuals. All we would do is add new relationships - so called edges - between the people.

Doing this allows us to easily ask more complex questions of the data, mostly by looking at the relationships between employees.

The organisation as a graph

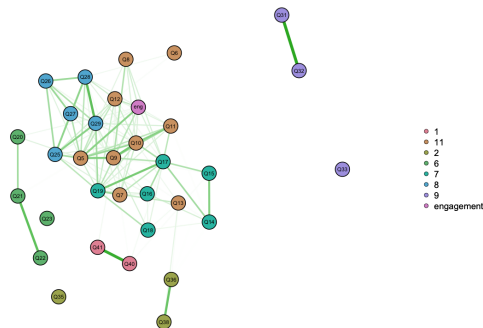


Even the simplest graph as shown enables us to produce some statistics easily that are harder to do otherwise. How many people report into a manager, or people who report in to them? How many reporting lines are between any individual and the CEO.

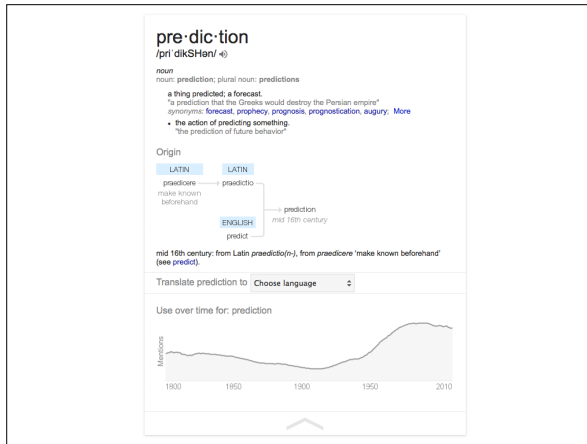
One of our clients has the latter as a good predictor of certain survey items.

We can use also adapt the survey to find relational data using a hybrid social network survey approach. Questions such as 'who do you go to get help to perform your role?' can be added.

Correlation graph of typical survey



Graphs can be used in other areas, not just linking individuals. Here we show the strength of correlation between questions.



Finally I want to talk about benchmarks. As many of you will realise a key issue for benchmarking is finding a reliable and appropriate comparison group. For engagement we all know that certain factors such as time-in-a-job, employment level, or age will alter the probability that any employee is engaged or not.

We're seeing quite a lot of promise at the moment using these demographic factors to build a prediction model of engagement. Where this is useful is when you roll-up the individual predictions of employees in a group to produce an expected level of engagement for the group as a whole. We're able to say 'given the types of people who are in this function this is how the group performed compared to what we'd have expected.'

This is intuitive, and because it is similar to how other business model performance is really appreciated by management.



I hope that I've shown you that there are lots of opportunities to be realised from using a data-science approach to survey data and analysis. As mentioned, the first step should be to integrate survey data with other data, making it more widely available.

Automating report prediction is one of the easiest ways of making an impact, driving down cost and increasing quality.

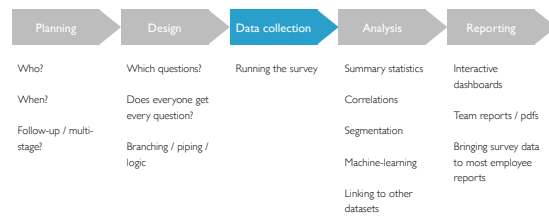
From there it's a question of the questions you ask the data.

Additional slides

Leading Analytic Organizations have 17%
better talent outcomes than other
organizations

source: Corporate Executive Board: The Talent Analytics Era

Current survey process time-consuming



The important part of *communicating data* is
communicating not data.

Analytic Value Chain



Decisions based on experience & gut feel

Information is structured but raw

Ratios, counts:

Correlations & Relationships

Leads to action

Changes in processes and policies

Can show trends but audience gets numb over time

Often misses the "Why?"

Needs to incorporate a narrative

Showing only when needed reduces fatigue

Decision tools for managers

