BIG data

Prof Richard Kimbell, Goldsmiths, University of London

I was reading the Metro freebie on the way home the other day and came across a report about the bravery of teachers. Not those facing normal classroom challenges, but rather the bravery of teachers in World War 1. Drawn from 2.8 million service records, the website 'Ancestry' tallied the number of Victoria Cross, Military Cross, Distinguished Conduct Medal and Meritorious Service Medal winners against the number employed in various professions in 1911. And it seems that teachers – with leadership skills in abundance – were awarded (proportionately) more gallantry medals than any other professional group. Whilst this might be of interest historically, culturally and even militarily, it was a quite different perspective on the report that made me sit up and ponder. Because the 'finding' comes from a very simple bit of analysis on a very large data-set. It is another example of the proliferation of BIG data, and what can be done with it.

eBay uses two vast data warehouses for search, consumer recommendations, and merchandising. *Amazon* handles millions of operations every day, as well as queries from more than half a million third-party sellers. And to do this, they have the world's three largest Linux databases. *Walmart* handles more than 1 million customer transactions every hour, which are imported into databases estimated to contain the equivalent of 167 times the information contained in all the books in the US Library of Congress.

I could go on… and on… and on about the data that is now held about us by commercial and other organisations. When Terry Leahy introduced the loyalty card to Tesco in the early 1990s, the idea was put through a series of trials to explore its potential for Tesco marketing. Every time a Clubcard is used, a copy of the store shopped in, products purchased and price paid are stored against the Clubcard account. Applicants are asked to provide personal details such as name, address and children. The data that Tesco holds about me as an individual seems (at one level) trifling, but when cross-referenced to analyse my buying habits over the year and then amalgamated with all customers with a similar post-code, or age, or family size, the data-base is mind-bogglingly big. Reviewing the results of the initial trials, Tesco's then Chairman is reported to have observed "What scares me about this is that you know more about my customers after three months than I know after 30 years."

Mesure, Susie (2003-10-10). "Loyalty card costs Tesco £1bn of profits – but is worth every penny". *The Independent.*

And now there are completely new opportunities for us (the general public) to explore and use BIG data. No longer the preserve of statisticians and big business, teachers and students can now get access to some of these data and use them to inform ourselves and enrich our practices.

One of the most accessible is Google Trends. This is a web-tool that creates a graphic display based on the number of times a particular search term is used in Google. It creates a set of graphs that display the 'search volume index' (a measure of the actual number of hits for a particular search term divided by the average search volume) across various regions of the world and/or across various periods of time. The graph can represent details of a search broken down by countries, regions, cities and languages.

I spent a while exploring what might be gleaned from the UK data in Google Trends and at first what you find is more-or-less what you would expect to find.

- searches on 'smoked salmon' and 'LEGO' reliably peak over the Christmas period;
- searches on 'grass seed' peak in April;
- searches on A&E peak in the week of July 28th (the start of the summer holidays);
- searches on 'curriculum' are at their lowest in August.

But interesting variations arise when the exploration of the search terms becomes global. LEGO in most developed countries peaks at Christmas – but in India there is a HUGE spike in the data for January 2008. Approximately five times the normal volume of searches took place in just that month and similarly for February 2014. What was going on in India that these two peaks – six years apart – should be so extreme? 'Hepatitis' searches for many countries peak in May and dip in December, as regular as clockwork. I imagine that says something about the climate and how it either accentuates or moderates risk factors. 'Cornflakes' has peaked in August for the last three years and especially in Brunei and Malaysia! 'Elvis' produced a really weird graph with a huge spike in August 2007. It turns out of course that – since Elvis died in August 1977 – it was all about a 30th anniversary. But why would the peakest peak be in Tirana, the capital city of Albania? This does not mean that there are *more actual* searches in Tirana than in (say) Memphis. By dividing the

BIG data

real search number by the *average* search number for particular places, the index effectively normalises the data and allows the peaks and troughs to show us useful trends.

Anyhow, what becomes very clear from a bit of digging into these data, is the extent to which the data itself starts to pose interesting questions…and provide real opportunities for designers.

Take one bit of data as an example. I have just watched the Oxford/Cambridge boat-race, and not surprisingly the hot spots in the UK for searches on the 'boat race' are in the week of the race itself (late March or early April) and in the cities of Oxford, Cambridge and London. But there are interesting additional trends in the data. In June there is a noticeable shadow peak. And the cities of Sheffield, Birmingham and Bristol show up as having significant interest. On a worldwide scale so do the Arab Emirates, Australia and Hong Kong. But there is more…. If I choose India for analysis, I find that most States show no interest at all, but it is present in the more southerly States of Andhara Pradesh, Karnataka and Tamil Nadu, and very high in Kerala. If I was a boat race event promoter – or designing a product that related in some way to such an event – or an author about to launch a book about boat races, this is vital information that could shape a product development or promotion strategy.

My colleagues Tony Wheeler and Tristram Shepard have recently been commissioned to develop a KS3 design programme of study – for a whole term – around BIG data. They illustrate how students can use such data to influence the projects they choose to work on. And once embarked on a project, they show how students can use real live BIG data to shape their designing.

Your client has identified 3 areas where BIG data could help you:

1) Targeting where there is a need/demand (context)?
   -mass (something lots of people do - space for new products)
   -niche (something fewer people do - new - with no/few products)
   -event driven (something cyclical - uniquely predictable by your data)
2) Focusing on the sort of things you could design to satisfy this need/demand (outcomes)?
   -products, systems & experiences?
   -more efficient, enjoyable, beautiful, durable, sustainable, desirable, cheaper ... etc..
3) Clarifying what sort of people would be most interested in these things (customer profiles)?
   -age, gender, income, occupation, location (demographics)

(Wheeler T & Shepard T. 2014 '*Designing with BIG data*' for the Creative Academies Trust)

Altogether it provides a fascinating glimpse into the use of simple data (like Google searches) collected on a VAST scale. Taking a slightly different tack, Wheeler and Shepard illustrate how big data can be used live, in real time, to shape our world and our behaviour. Essentially what is going on here is that product developers are enlisting the mass public to create data that either informs the design of the product – or that uses the product to create data from its use. A lovely example is the Copenhagen wheel.

http://www.youtube.com/watch?v=U5k25-hHNrc

The hub of the bicycle wheel collects data as it goes along (e.g. about CO2, NOX and other noxious elements in the atmosphere), so that as one rides through the city it provides a very fine-grained map of the atmospheric make up of the pathways and roads travelled. If that data is shared (anonymously) via smart-phones, then everyone riding a bike with this type of wheel is collaborating in the construction of a real-time, citywide environmental data map. This can be immensely valuable to those responsible for developing transport systems.

But I have another idea. Every secondary school will have a very simple bit of data about every student…their A-C 'passes' at GCSE. So each child can have a GCSE grade index (a simple count of their 'passes' at A-C. And of course the date of birth will be known. So we can construct a BIG data map of performance against birth sign. Are Capricorn's better at chemistry? Do Scorpio's score with music? Are Aquarian's great at art, and Leo's red-hot with Latin? On second thoughts we'd better keep quiet about all that, or Mr Gove will launch another national curriculum review to overcome our perceived underperformance in…something or other.

r.kimbell@gold.ac.uk