The Guardian

Big data and the end of theory?

Does big data have the answers? Maybe some, but not all, says Mark Graham

Mark Graham

Fri 9 Mar 2012 14.39 GMT

In 2008, Chris Anderson, then editor of Wired, wrote a provocative piece titled *The End of Theory*. Anderson was referring to the ways that computers, algorithms, and big data can potentially generate more insightful, useful, accurate, or true results than specialists or domain experts who traditionally craft carefully targeted hypotheses and research strategies.

This revolutionary notion has now entered not just the popular imagination, but also the research practices of corporations, states, journalists and academics. The idea being that the data shadows and information trails of people, machines, commodities and even nature can reveal secrets to us that we now have the power and prowess to uncover.

In other words, we no longer need to speculate and hypothesise; we simply need to let machines lead us to the patterns, trends, and relationships in social, economic, political, and environmental relationships.

It is quite likely that you yourself have been the unwitting subject of a big data experiment carried out by Google, Facebook and many other large Web platforms. Google, for instance, has been able to collect extraordinary insights into what specific colours, layouts, rankings, and designs make people more efficient searchers. They do this by slightly tweaking their results and website for a few million searches at a time and then examining the often subtle ways in which people react.

Most large retailers similarly analyse enormous quantities of data from their databases of sales (which are linked to you by credit card numbers and loyalty cards) in order to make uncanny predictions about your future behaviours. In a now famous case, the American retailer, Target, upset a Minneapolis man by knowing more about his teenage daughter's sex life than he did. Target was able to predict his daughter's pregnancy by monitoring her shopping patterns and comparing that information to an enormous database detailing billions of dollars of sales. This ultimately allows the company to make uncanny
predictions about its shoppers.

More significantly, national intelligence agencies are mining vast quantities of non-public Internet data to look for weak signals that might indicate planned threats or attacks.

There can by no denying the significant power and potentials of big data. And the huge resources being invested in both the public and private sectors to study it are a testament to this.

However, crucially important caveats are needed when using such datasets: caveats that, worryingly, seem to be frequently overlooked.

The raw informational material for big data projects is often derived from large user-generated or social media platforms (e.g. Twitter or Wikipedia). Yet, in all such cases we are necessarily only relying on information generated by an incredibly biased or skewed user-base.

Gender, geography, race, income, and a range of other social and economic factors all play a role in how information is produced and reproduced. People from different places and different backgrounds tend to produce different sorts of information. And so we risk ignoring a lot of important nuance if relying on big data as a social/economic/political mirror.

We can of course account for such bias by segmenting our data. Take the case of using Twitter to gain insights into last summer's London riots. About a third of all UK Internet users have a twitter profile; a subset of that group are the active tweeters who produce the bulk of content; and then a tiny subset of that group (about 1%) geocode their tweets (essential information if you want to know about where your information is coming from).

Despite the fact that we have a database of tens of millions of data points, we are necessarily working with subsets of subsets of subsets. Big data no longer seems so big. Such data thus serves to amplify the information produced by a small minority (a point repeatedly made by UCL's Muki Haklay), and skew, or even render invisible, ideas, trends, people, and patterns that aren't mirrored or represented in the datasets that we work with.

Big data is undoubtedly useful for addressing and overcoming many important issues face by society. But we need to ensure that we aren't seduced by the promises of big data to render theory unnecessary.

We may one day get to the point where sufficient quantities of big data can be harvested to answer all of the social questions that most concern us. I doubt it though. There will always be digital

divides; always be uneven data shadows; and always be biases in how information and technology are used and produced.

And so we shouldn't forget the important role of specialists to contextualise and offer insights into what our data do, and maybe more importantly, don't tell us.

Mark Graham is a research fellow at the Oxford Internet Institute and is one of the creators of the Floating Sheep blog

More open data

- Data journalism and data visualisations from the Guardian
- Who's who on the Datablog

World government data

- Search the world's government data with our gateway

Development and aid data

- Search the world's global development data with our gateway

Can you do something with this data?

- **Flickr** Please post your visualisations and mash-ups on our Flickr group
- Contact us at data@guardian.co.uk

- **Get the A-Z of data**
- **More at the Datastore directory**
- **Follow us on Twitter**
- **Like us on Facebook**