Biggish Data

written by Jeremy Huggett | 20/05/2016

Big Data ;-)

Big Data is (are?) old hat … Big Data dropped off Gartner's *Emerging Technologies Hype Cycle* altogether in 2015, having slipped into the 'Trough of Disillusionment' in 2014 (Gartner Inc. 2014, 2015a). The reason given for this was simply that it had evolved and had become the new normal – the high-volume, high-velocity, high-variety types of information that classically defined 'big data' were becoming embedded in a range of different practices (e.g. Heudecker 2015).

At the same time, some of the assumptions behind Big Data were being questioned. It was no longer quite so straightforward to claim that 'big data' could overcome 'small data' by throwing computer power at a problem, or that quantity outweighed quality such that the large size of a dataset offset any problems of errors and inaccuracies in the data (e.g. Mayer-Schönberger and Cukier 2013, 33), or that these data could be analysed in the absence of any hypotheses (Anderson 2008).

For instance, boyd and Crawford had highlighted the mythical status of 'big data'; in particular that it somehow provided a higher order of intelligence that could create insights that were otherwise impossible, and assigned them an aura of truth, objectivity and accuracy (2012, 663). Others followed suit. For example, McFarland and McFarland (2015) have recently shown how most Big Data analyses give rise to "precisely inaccurate" results simply because the sample size is so large that they give rise to statistically highly significant results (and hence the debacle over Google Flu Trends – for example, Lazer and Kennedy 2015). Similarly, Pechenick *et al* (2015) showed how, counter-intuitively, results from Google's Books Corpus could easily be distorted by a single prolific author, or by the fact that there was a marked increase in scientific articles included in the corpus after the 1960s. Indeed, Peter Sondergaard, a senior vice president at Gartner and global head of

Research, underlined that data (big or otherwise) are inherently dumb without algorithms to work on them (Gartner Inc. 2015b). In this regard, one might claim Big Data have been superseded by Big Algorithms in many respects.

The apparent blowing over of the hype surrounding Big Data could seem reassuring. We have always had difficulty in characterising archaeological data as 'big' data. Although a dataset might be big to us, it was tiny in the grand scheme of things. Archaeological data never conformed to the characteristics – the three V's – of big data:

- our *volume* is not that great (terabytes rather than petabytes or exabytes);
- its *velocity* in real time is not an issue (rather the reverse!);
- our *variety* of data types is no greater than many.

Other supposedly big data characteristics such as veracity, viability, validity and value are what Seth Grimes (2013) calls 'wanna-Vs' in that they don't extend the essence of big data and unlike the 3-Vs are equally true of 'small' data (and certainly of archaeological data) in general.

But this isn't the end of the affair. Although archaeological datasets may not truly constitute 'big data', much the same issues arise. In the thrill of being able to access large(ish!) datasets, increasingly on a national scale, we can be prone to the same errors of commission and omission in data handling. These have always existed, but what brings them to the fore is the scale, range, combinations, and accessibility of the data that are becoming available to us. For example, Katherine Robbins has commented that relatively few studies incorporating the Portable Antiquity Scheme data have employed methods to address the sampling biases within the data (Robbins 2013, 58). Similarly, Tim Evans has observed that researchers have failed to engage properly with national event databases (2013, 32). This should set off alarm bells as we see more and more large-scale analyses of archaeological data, whether using PAS data or national inventories or, for that matter, drawing together multiple field datasets. There is perhaps a tendency to shrug our shoulders about data – we know it's incomplete, problematic in numerous ways, but essentially it's all we have so we get on and use it. I know I have, and I'm fairly sure I'm not alone in this.

Katherine Robbins has usefully identified a number of biases within the PAS data, as well as some means of addressing them (2013, 58-69; 2014, 37-76). These include issues associated with differential recovery from different land types, differential access to land, the effect of collecting territories, the presence of known sites and other features such as roads, settlements etc., search and retrieval strategies, levels of reporting, and levels of recording – and without the latter, of course, information never makes it into the database in the first place. Such factors are equally relevant to national inventories, regional Historic Environment Records, as well as field data commonly found in grey literature reports and combined into synthetic reports (see, for example, Evans 2013, 2015; Cooper and Green 2016). Since many large-scale studies – at least, in press – skate over matters of 'data cleansing', 'data integration', 'data homogenisation', or focus on 'big data-like' automated processing to combine datasets, it's difficult to judge the extent to which these kinds of biases have been addressed, most of which require the resolution of human practices rather than technical pattern-matching or aggregation.

As Anwen Cooper and Chris Green have recently observed, archaeological data

> "… condenses a whole series of other relationships and practices involving people, archaeological materials, organisations, technologies of various kinds (digital and otherwise), and so on, that have sometimes unfolded over a very long time period and that are always evolving." (2016, 279).

This time dimension of archaeological data gives it a particular character – its inherent chronological dimension makes it an example of what Arbesman (2013) has called 'long data', and although this is equally true of others (geology, palaeontology, for instance), archaeology also has a particularly destructive approach to its primary data. But while archaeological data can certainly provide a temporal perspective on our past and hence be valuable in studies of environmental change, for instance, the time dimension also incorporates change within our data practices over what can be two hundred years or more of data collection. This is what makes archaeological data especially challenging: the complexities and practices embedded within them, and the recognition and appropriate handling of these, can make the much-vaunted '3-Vs' of Big Data seem rather simple in comparison.

References

Anderson, C. 2008 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete', *Wired* (23rd June 2008) http://www.wired.com/2008/06/pb-theory/

Arbesman, S. 2013 'Stop Hyping Big Data and Start Paying Attention to 'Long Data'', *Wired* (29th January 2013) http://www.wired.com/2013/01/forget-big-data-think-long-data/

boyd, d. and Crawford, K. 2012 'Critical questions for Big Data: provocations for a cultural, technological, and scholarly phenomenon', *Information, Communication & Society* 15 (5), 662-679. http://dx.doi.org/ 10.1080/1369118X.2012.678878

Cooper, A. and Green, C. 2016 'Embracing the Complexities of 'Big Data' in Archaeology: the Case of the English Landscapes and Identities Project', *Journal of Archaeological Method and Theory* 23, 271-304. http://dx.doi.org/10.1007/s10816-015-9240-4

Evans, T. 2013 'Holes in the Archaeological Record? A Comparison of National Event Databases for the Historic Environment in England', *The Historic Environment: Policy & Practice* 4 (1), 19-34. http://dx.doi.org/10.1179/1756750513Z.00000000023

Evans, T. 2015 'A Reassessment of Archaeological Grey Literature: semantics and paradoxes', *Internet Archaeology* 40. http://dx.doi.org/10.11141/ia.40.6

Gartner Inc. 2014 'Gartner's 2014 Hype Cycle for Emerging Technologies Maps the Journey to Digital Business' (Press Release 11th August 2014) http://www.gartner.com/newsroom/id/2819918

Gartner Inc. 2015a 'Gartner's 2015 Hype Cycle for Emerging Technologies Identifies the Computing Innovations That Organizations Should Monitor' (Press Release 18th August 2015) http://www.gartner.com/newsroom/id/3114217

Gartner Inc. 2015b 'Gartner Says It's Not Just About Big Data; It's What You Do With It: Welcome to

the Algorithmic Economy' (Press Release 5th October 2015)
http://www.gartner.com/newsroom/id/3142917

Grimes, S. 2013 'Big Data: Avoid 'Wanna V' Confusion', *InformationWeek* (7th August 2013).
http://www.informationweek.com/big-data/big-data-analytics/big-data-avoid-wanna-v-confusion/d/d-id/1111077

Heudecker, N. 2015 'Big Data isn't Obsolete. It's Normal.' *Gartner Blog Network* (20th August 2015)
http://blogs.gartner.com/nick-heudecker/big-data-is-now-normal/

Lazer, D. and Kennedy, R. 2015 'What We Can Learn from the Epic Failure of Google Flu Trends', *Wired* (1st October 2015) http://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Mayer-Schönberger, V. and Cukier, K. 2013 *Big data: A revolution that will transform how we live, work, and think* (Houghton Mifflin Harcourt).

McFarland, D. and McFarland, H.R. 2015 'Big Data and the danger of being precisely inaccurate', *Big Data &Society* 2 (2), 1-4. http://dx.doi.org/ 10.1177/2053951715602495

Pechenick, E., Danforth, C. and Dodds, P. 2015 'Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution', *PLoS ONE* 10 (10), e0137041.
http://dx.doi.org/10.1371/journal.pone.0137041

Robbins, K. 2013 'Balancing the Scales: Exploring the Variable Effects of Collection Bias on Data Collected by the Portable Antiquities Scheme', *Landscapes* 14 (1),
54-72. http://dx.doi.org/10.1179/1466203513Z.0000000006

Robbins, K. 2014 *Portable Antiquities Scheme – A Guide for Researchers* (Portable Antiquities Scheme/British Museum/Leverhulme Trust). https://finds.org.uk/research/advice