Geoforum 71 (2016) 1–4

Contents lists available at ScienceDirect

Geoforum

journal homepage: www.elsevier.com/locate/geoforum

Critical review

Big data and the historical sciences: A critique

Malte C. Ebach [a,*], Michaelis S. Michael [b], Wendy S. Shaw [a], James Goff [a], Daniel J. Murphy [c], Slade Matthews [d]

[a] Palaeontology, Geobiology and Earth Archives Research Centre, School of Biological, Earth and Environmental Sciences, UNSW, Kensington, NSW 2052, Australia
[b] School of Humanities and Languages, UNSW, Kensington, NSW 2052, Australia
[c] Royal Botanic Gardens Victoria, Melbourne, VIC 3004, Australia
[d] Pharmacoinformatics Laboratory, Discipline of Pharmacology, Bosch Institute and Sydney Medical School, The University of Sydney, NSW, Australia

ARTICLE INFO

Article history:
Received 17 December 2015
Received in revised form 7 February 2016
Accepted 21 February 2016

Keywords:
Big Data Cycle
Big Data hubris
Biogeography
Google Flu Trends
Next Generation Sequencing
Small data

ABSTRACT

Social media, government, industry and science use data in the same way, through the pursuit of correlations in large data sets. As this critique shows, however, there is greater dialogue about the potential pitfalls of Big Data and the Big Data Cycle in non-historical science fields, such as medicine and advertising. Pitfalls, such as the Big Data Hubris, the Filter Bubble and correlation superseding causation, are discussed in relation to the historical sciences.

Contents

1. Introduction

The Big Data Era has brought with it many changes to scientific practise, such as new ways of extracting information (i.e. data mining) and analysing data.[1] These changes have been adopted without scientists engaging in a dialogue about the potential pitfalls of Big Data (for a discussion in the social sciences see Batty, 2013; Barnes, 2013; Creswell, 2013). For example, can Big Data predict causal events in the same way as traditional methods? More importantly, have the problems of traditional scientific practise been addressed in Big Data? We feel that well-meaning scientists have drunk the Kool Aid of the Big Data Hubris.

Big Data is a term that refers to large databases, whereas the Big Data Cycle (BDC) denotes data-driven mathematical models that are used to analyse for trends, correlates as well as patterns and the associated technology or hardware (*sensu* Jagadish et al., 2014). Big Data also encompasses a group of users, in our case scientists, who are dependent on the databases, models and technology to do their work. The impact of Big Data in government policy, economics and the social and political sciences has resulted in terms, such as "Disaster Big Data", "Dictatorship of Big Data", and "Big Data Era". While much media attention is given to the role of Big Data in mass surveillance and consumerism, there has been little criticism of the role of Big Data in the historical sciences (e.g.,

* Corresponding author.
E-mail address: mcebach@gmail.com (M.C. Ebach).

[1] boyd and Crawford argue that "Big Data is less about data that is big than it is about a capacity to search, aggregate, and cross-reference large data sets" (boyd and Crawford, 2012, p. 663).

http://dx.doi.org/10.1016/j.geoforum.2016.02.020
0016-7185/

Earth, biological, geographical and pharmaceutical sciences).[2] DNA barcoding, in initiated in the 2000s, is an early example of Big Data as applied to biological sciences. DNA barcoding was introduced as "a novel system designed to provide rapid, accurate, and automatable species identifications...using short, standardized gene regions as internal species tags....it will make the Linnaean taxonomic system more accessible, with benefits to ecologists, conservationists...agencies charged with the control of pests, invasive species, and food safety" (Hebert and Gregory, 2005, p. 852). Taxonomists, we were told, would benefit, "DNA barcoding will add rigour to the generation and testing of taxonomic hypotheses" (Hebert and Gregory, 2005, p. 855). A decade later, DNA barcoding has had minor impact in taxonomy, with some benefit for phylogeneticists, ecologists and conservationists; it has largely been superseded by technological advances, namely Next Generation Sequencing (NGS), in which whole genomes can now be routinely sequenced. Yet, many biologists had adopted DNA barcoding without discussing the potential pitfalls, particularly in taxonomy (Ebach and Holdrege, 2005). With the onset of NGS, and a new set of potential problems, there still remains no serious dialogue. In stark contrast to the global debates on Big Data and mass surveillance, and Big Data and consumerism, in books, films and social media (e.g., Nekrutenko and Taylor, 2012), there is little critical discussion of Big Data by biologists. While we acknowledge that social issues provoke far more skepticism and concern among the general public, we note a significant risk to science. A lack of caution and skepticism is, ironically, unscientific. As historical science enters the Big Data Era it enters encumbered with "Big Data hubris" (Lazer et al., 2014).

2. Google Flu Trends and the Big Data Hubris

In 2008, researchers at Google and the Centre for Disease Control declared that "[o]ne way to improve early [influenza pandemic] detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day" (Ginsberg et al., 2009, p. 1012). The system, known as Google Flu Trends (GFT), uses non-traditional sources of information to predict flu pandemics, providing fast and fair approximation of a potential flu pandemic separately to traditional source gathering techniques and analysis, which could takes weeks. Put simply, GFT tracks certain Google search terms using a search algorithm. Google launched GFT in November 2008 to much applause (Park, 2009; Helft, 2008; Mayer-Schönberger and Cukier, 2013).

A year later GFT missed the 2009 influenza A-H1N1 pandemic, also known as swine flu (Cook et al., 2011). A new model was constructed. However, as Lazer et al. (2014) later found, the new GFT model "has been persistently overestimating flu prevalence", claiming that "GFT has never documented the 45 search terms used, and the examples that have been released appear misleading" (Lazer et al., 2014:1203–1204). Rather than proving to be an improvement on government, it overestimated the prevalence of flu by 50% for 100 out of 108 weeks from 21 August 2011 to 1 September 2013 (Lazer et al., 2014). GFT was considered an "epic failure" that "turned the poster child of big data into the poster child of the foibles of big data" (Lazer and Kennedy, 2015). By 2015, GFT was "no longer publishing current estimates of Flu and Dengue fever based on search terms" (https://www.google.org/flutrends/about/). The GFT experiment had failed due to Big Data Hubris and the lack of methodological transparency (Lazer et al., 2014).

[2] English polymath William Whewell (1794–1866) referred to these as the palaetiological sciences that "refer to actual past events, but attempt to explain them by laws of causation" (Whewell, 1837, p. 481). In contrast to experimental sciences, in which experiment processes can be observed.

Big Data hubris is the notion that big data replaces, rather than supplements traditional data acquisition and analysis, namely "small data" (Lazer et al., 2014; Kitchin, 2014). The enormity of big data is due to "the output of instruments designed to produce valid and reliable data amenable for scientific analysis" (Lazer et al., 2014, p. 1203). Ginsberg et al. (2009) did state that the GFT was not a replacement for small data or traditional methods but GFT failed due to a lack of methodological transparency, making this a case of a Black Box within Big Data. The problem is not the quantity of information, but rather the methods used to extract, process and analyse data: "Interpretation is at the center of data analysis. Regardless of the size of a data, it is subject to limitation and bias" (boyd and Crawford, 2012, p. 668).

3. "Oh no! My TiVo thinks I'm Gay" and the Filter Bubble

The use of polls, purchasing and internet browsing information to tailor products and services for consumers is perhaps the most prevalent use of Big Data in marketing. The methods used to target individual consumers is made via machine learning, a form of artificial intelligence that is based on a set of collected information about a consumer. In some cases user profiling may not reflect the choices of the individual. Tivo is a good example. TiVo is a digital video-recorder that records some programs, and "assumes [what] its owner will like, based on shows the viewer has chosen to record" (Zaslow, 2002). A user of TiVo felt inappropriately targeted when offered a steady stream of gay programming. The user then recorded war films, so TiVo provided "documentaries on Joseph Goebbels and Adolf Eichmann. It stopped thinking I was gay and decided I was a crazy guy reminiscing about the Third Reich" (Zaslow, 2002).

TiVo is an allegory for many scientific models in which the user's choice, rather than scholarly argumentation, is deciding which scientific models are "better". Models based on popular choice rather scholarly discussion may lead to a group user profile. Not to be confused with a "paradigm shift", a group profile derives from user trends rather actual scientific discovery and debate. In historical sciences, such as biogeography, past biological processes, including dispersal, are unobservable. All that biogeographers are left with are geographical data (i.e., known distributions) and knowledge of evolutionary relationships. Recent developments have seen a shift from debating past processes based on small data, to letting "the data tell us which models are to be preferred" (Matzke, 2014, p. 968). The point here is that Big Data are too large to view objectively and finding a statistical pattern may influence our model choice. Considering biogeographical data are silent about the types of processes that have occurred in the past, it is still up to the user, and not the data, to select which model is best or most appropriate. Given the scale of the datasets, biogeographers who use statistical biogeographic computer software need to, but often do not, carefully consider the theoretical consequences of different models when compared to one another. Instead users often chose the default, that is, the most popular option which may lead to a group user profile, meaning that scientific models may be justified purely through popular or default group user profiles (e.g., long distance dispersal), rather than scholarly argumentation. As a consequence, future computer software programmers may simply discard some model algorithms in favour of others that are simply more popular, through user profiling, thereby excluding effective models.

Modelling based on user profiling also causes an effect called the "Filter bubble", a term coined by Pariser (2011), "in which [political views, news and current affairs] content is selected by algorithms according to a viewer's previous behaviours [...], which are devoid of attitude-challenging content" (Bakshy et al., 2015, p. 1130).