ECR Forum

BIG DATA & SOCIETY

Big Data & Society
April–June 2014: 1–2
© The Author(s) 2014
DOI: 10.1177/2053951714540613
bds.sagepub.com
SAGE

On minorities and outliers: The case for making Big Data small

Brooke Foucault Welles

Abstract

In this essay, I make the case for choosing to examine small subsets of Big Data datasets—making big data small. Big Data allows us to produce summaries of human behavior at a scale never before possible. But in the push to produce these summaries, we risk losing sight of a secondary but equally important advantage of Big Data—the plentiful representation of minorities. Women, minorities and statistical outliers have historically been omitted from the scientific record, with problematic consequences. Big Data affords the opportunity to remedy those omissions. However, to do so, Big Data researchers must choose to examine very small subsets of otherwise large datasets. I encourage researchers to embrace an ethical, empirical and epistemological stance on Big Data that includes minorities and outliers as reference categories, rather than the exceptions to statistical norms.

Keywords

Big Data, computational social science, sampling, epistemology, feminism, women

Twenty thousand users, one hundred thousand users, ten million users. In the world of computational social science, Big Data has provoked an analytic arms race to work with more data, better data, *bigger* data in pursuit of discovering so-called truths about the social world. At meetings, it is not uncommon for computational social scientists to boast about the size of their datasets, as if millions of users are universally and self-evidently better than thousands or hundreds. Like many assistant professors, I struggle with the ''imposter phenomenon,''—a feeling that my intellectual and technical skills do not quite measure up to those of my peers (Clance and Imes, 1978). So, it can be hard to suppress the anxiety that I feel when those questions come my way. ''How big is your dataset?'' they ask. ''1500,'' I say, ''no bigger than a modest survey, but different in an important way.''

I study women—most recently, older women who play online games with such intensity that they distinguish themselves not only from their gender- and age-mates in the offline world but also from their game-playing peers in the online world as well. I have long been interested in how women's lives shape and are shaped by technology, so when I began working with online game datasets in graduate school, it seemed natural to me to focus on women's experiences. Like a growing number of my colleagues in Computational Social Science (Lazer et al., 2009), I am motivated by theoretical questions, and Big Data is the tool to answer those questions. In my case, against a backdrop of discussions about sexism and misogyny in online gaming (Fox and Tang, 2013), Big Data from online games promises to reveal patterns of behavior that can help women resist gendered aggression in male-dominated gaming communities and on the internet more broadly.

However, a large dataset quickly becomes small when you focus on a minority population. In my dataset of 10 million players from the virtual world *Second Life*, about a third are women. Of those, only one in 20 is over the age of 50. Among those, just the tiniest

Northeastern University, USA

Corresponding author:
Brooke Foucault Welles, Northeastern University, Boston, MA 02115, USA.
Email: b.welles@neu.edu

Creative Commons Non Commercial CC-BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 3.0 License (http://www.creativecommons.org/licenses/by-nc/3.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (http://www.uk.sagepub.com/aboutus/openaccess.htm).

Downloaded from by guest on October 25, 2015

statistical minority—1%—has played for 1000 hours or more. So, what started as a dataset of 10 million players is reduced to just 1500 players with novel characteristics. This extreme minority would normally get lost in Big Data analytics, wiped away as noise among the statistically average masses. Some Big Data researchers might abandon projects that whittle datasets down so substantially, believing that focusing on 1500 players no longer "counts" as Big Data research. However, I argue that honoring the experiences of extreme statistical minorities represents one of Big Data's most exciting scientific possibilities.

Choosing to work with a small sample drawn from Big Data represents an important empirical stance for Computational Social Science and Big Data analytics. Scholars have long critiqued the omission of women and minorities from the scientific literature. Even the most methodologically and epistemologically conservative of these critiques speaks to the challenges of using majority samples to generalize about minority experiences (Keller, 1995). When women and minorities are excluded as subjects of basic social science research, there is a tendency to identify majority experiences as "normal," and discuss minority experiences in terms of how they deviate from those norms (Gilligan, 1982). In doing so, women, minorities, and the statistically underrepresented are problematically written into the margins of social science, discussed only in terms of their differences, or else excluded altogether (Smith, 1974).

Historically, in the pre-computational era, researchers may have made the case that it was simply too difficult to work with underrepresented populations and statistical outliers. These people are, by definition, less plentiful in the population. So, they can be harder to find, more expensive to recruit, and more time-consuming to work with. Although ethically and empirically inexcusable, researchers working with tight budgets and limited time frames may have felt that it was not viable to work with non-majority populations. However, Big Data changes all of that. In our datasets of millions, the minorities and statistical outliers are just as easy to access as the majorities and statistically average. We simply have to choose to look.

The reasons to make that choice are numerous. Ethically, we have a responsibility to include a diverse range of participants in our work so that the benefits of our scientific practice can be as widely applicable as possible. Empirically, focusing on minority experiences as reference categories, rather than as deviations from the majority reference, enables better, more accurate theory building and data modeling (Gilligan, 1982). And, epistemologically, choosing small foci within Big Data dismantles the problematic ethos emerging within computational social science and Big Data analytics of bigger data being "truer" data (boyd and Crawford, 2012). Big Data are neither inherently true nor inherently comprehensive, but they do contain clues about populations long-overlooked in the social sciences.

As we enter a new age of Big Data-driven computational social science, we are poised to either replicate or remediate the mistakes of the past. One of the greatest advantages of Big Data in computational social science research is the breadth of experiences that it represents. Big Data allows us to produce summaries of human behavior at a scale never before possible. But in the push to produce these summaries, we risk losing sight of a secondary but equally important advantage of Big Data—the plentiful representation of minorities. Those who might otherwise be represented as a single outlier in a more traditional dataset can number hundreds or thousands in a Big Data dataset—hundreds or thousands whose experiences are currently absent from the scientific record. Rather than actively removing these voices through sampling and data cleaning, or passively silencing them through statistical aggregation, I choose to embrace the opportunity to examine the statistical outliers, and I encourage my colleagues to do the same. By choosing to make Big Data small, we can rectify historical omissions and biases in social science research and build better, more comprehensive, *bigger* understandings of human behavior.

Declaration of conflicting interest

The author declares that there is no conflict of interest.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

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.

Clance PR and Imes SA (1978) The imposter phenomenon in high achieving women: Dynamics and therapeutic intervention. *Psychotherapy: Theory, Research & Practice* 15(3): 241.

Fox J and Tang WY (2013) Sexism in online video games: The role of conformity to masculine norms and social dominance orientation. *Computers in Human Behavior* 33: 314–320.

Gilligan C (1982) *In a Different Voice*. Cambridge, MA: Harvard University Press.

Keller EF (1995) *Reflections on Gender and Science*. New Haven, CT: Yale University Press.

Lazer D, Pentland AS, Adamic LA, et al. (2009) Life in the network: The coming age of computational social science. *Science* 323(5915): 721–723.

Smith DE (1974) Women's perspective as a radical critique of sociology. *Sociological Inquiry* 44(1): 7–13.

Downloaded from by guest on October 25, 2015