LETTERS

Edited by **Jennifer Sills**

Twitter: Big data opportunities

Response

WE THANK BRONIATOWSKI, Paul, and Dredze for giving us the opportunity to reemphasize the potential of big data and make the more obvious point that not all big data projects have the problems currently plaguing Google Flu Trends (GFT), nor are these problems inherent to the field in general.

Our Policy Forum is meant to provide a constructive critique by highlighting possible pitfalls of big data analysis. These pitfalls are not the same for all big data sets, but are certainly not unique to GFT. We do agree that Twitter has substantial scientific potential and is distinctive in the public availability of its data. Indeed, one of us (A.V.) is using Twitter data for influenza surveillance in the context of the recent Center for Disease Control (CDC) "Predict the Flu Challenge" (*1*).

Twitter data provide an excellent representation of those who choose to express an opinion publicly, which can be of tremendous value for many research purposes. However, these data may be manipulated by both the service provider (such as Google) and the user (such as companies marketing a product), as we explain in our Policy Forum. In light of these trends, whether these data can be used to represent the entire United States population remains an open question.

Who uses Twitter and how they use it have changed markedly over the past several years. The algorithmic underpinning of Twitter (which identifies "what's trending") is subject to constant and invisible tinkering. The system is under constant attack, with armies of bots ready to produce content for the highest bidder (*2*, *3*). The norms of expression on Twitter are heterogeneous and still rapidly evolving—who feels the need to publicly express that they have flu symptoms on Twitter, and are these predispositions evenly distributed throughout the population (*4*)? Bodnar and Salathé's cautionary tale (*5*) on Twitter-based influenza surveillance clearly shows that seemingly irrelevant tweets (such as those about zombies) are moderately indicative of influenza prevalence, and that the choice of validation methods has a large effect on reported success.

It is possible that one day we will have reliable prediction of flu prevalence from social media. Certainly, this would require a careful evaluation and recalibration of methodologies, public and independent replication of results, and the explicit evaluation of error processes. Clearly, all big data projects do not have the same syndromes as GFT presently does, but by building strong collaborations and adhering to rigorous standards, we should be able to extract considerably more information from these highly informative new data sources.

David Lazer,[1,2]**** Ryan Kennedy,***[1,3,4]
Gary King,[3] ***Alessandro Vespignani***[5,6,3]

[1]Lazer Laboratory, Northeastern University, Boston, MA 02115, USA. [2]Harvard Kennedy School, Harvard University, Cambridge, MA 02138, USA. [3]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA. [4]University of Houston, Houston, TX 77204, USA. [5]Laboratory for the Modeling of Biological and Sociotechnical Systems, Northeastern University, Boston, MA 02115, USA. [6]Institute for Scientific Interchange Foundation, Turin, Italy.

*Corresponding author. E-mail: d.lazer@neu.edu

REFERENCES

1. CDC, CDC Competition Encourages Use of Social Media to Predict Flu (www.cdc.gov/flu/news/predict-flu-challenge.htm).
2. J. Zhang, R. Zhang, Y. Zhang, G. Yan, "On the impact of social botnets for spam distribution and digital-influence manipulation" [2013 IEEE Conference on Communications and Network Security (CNS), 2013].
3. N. Bilton, "Friends, and influence, for sale online" (20 April

PHOTO: MARK OLEKSIY/THINKSTOCK

Published by AAAS

2014); http://bits.blogs.nytimes.com/2014/04/20/friends-and-influence-for-sale-online/.

4. Y. Liu, C. Kliman-Silver, A. Mislove, in Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM'14) (Ann Arbor, MI, 2014).
5. T. J. Bodnar, M. Salathé, "Validating models for disease detection using Twitter," 1st International Workshop on Public Health in the Digital Age: Social Media, Crowdsourcing and Participatory Systems, WWW2013 (2013).

PHOTO: HENRIK JONSSON GRAPHIC DESIGN/ISTOCKPHOTO

Twitter: Big data opportunities—Response

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani

Science, 345 (6193), • DOI: 10.1126/science.345.6193.148-b

Science (ISSN 1095-9203) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.