What does Big Data mean to public affairs research? Understanding the methodological and analytical challenges

blogs.lse.ac.uk/impactofsocialsciences/2016/12/08/what-does-big-data-mean-to-public-affairs-research-understanding-the-method

12/8/2016

The term 'Big Data' is often misunderstood or poorly defined, especially in the public sector. ***Ines Mergel****,* ***R. Karl Rethemeyer****, and* ***Kimberley R. Isett*** *provide a definition that adequately encompasses the scale, collection processes, and sources of Big Data. However, while recognising its immense potential it is also important to consider the limitations when using Big Data as a policymaking tool. Using this data for purposes not previously envisioned can be problematic, researchers may encounter ethical issues, and certain demographics are often not captured or represented.*

In the public sector, the term 'Big Data' is often misused, misunderstood, and poorly defined. Public sector practitioners and researchers frequently use the term to refer to large data sets that were administratively collected by a government agency. Though these data sets are usually quite large and can be used for predictive analytics, administrative data does not include the oceans of information that is created by private citizens through their interactions with each other online (such as social media or business transaction data) or through sensors in buildings, cars, and streets. Moreover, when public sector researchers and practitioners *do* consider broader definitions of Big Data they often overlook key political, ethical, and methodological complexities that may bias the insights gleaned from 'going Big'. In our recent paper we seek to provide a clearer definition that is current and conversant with how other fields define Big Data, before turning to fundamental issues that public sector practitioners and researchers must keep in mind when using Big Data.

Defining Big Data for the public sector

Public affairs research and practice has long profited from dialogue with allied disciplines like management and political science and has more recently incorporated insights from computational and information science. Drawing on all of these fields we define Big Data as:

> *"High volume data that frequently combines highly structured administrative data actively collected by public sector organizations with continuously and automatically collected structured and unstructured real-time data that are often passively created by public and private entities through their internet."*

This definition encompasses the scale of newly emerging data sets (many observations with many variables) while also addressing data collection processes (continuous and automatic), the form of the data collected (structured and unstructured), and the sources of such data (public and private). The definition also suggests the 'granularity' of the data (more variables describing more discrete characteristics of persons, places, events, interactions, and so forth), and the lag between collection and readiness for analysis (ever shorter).

Methodological and analytical challenges

Defined thus Big Data promises access to vast amounts of real-time information from public and private sources that should allow insights into behavioral preferences, policy options, and methods for public service improvement. In the private sector, marketing preferences can be aligned with customer insights gleaned from Big Data. In the public sector however, government agencies are less responsive and agile in their real-time interactions by design – instead using time for deliberation to respond to broader public goods. The responsiveness Big Data promises is a

virtue in the private sector but could be a vice in the public.

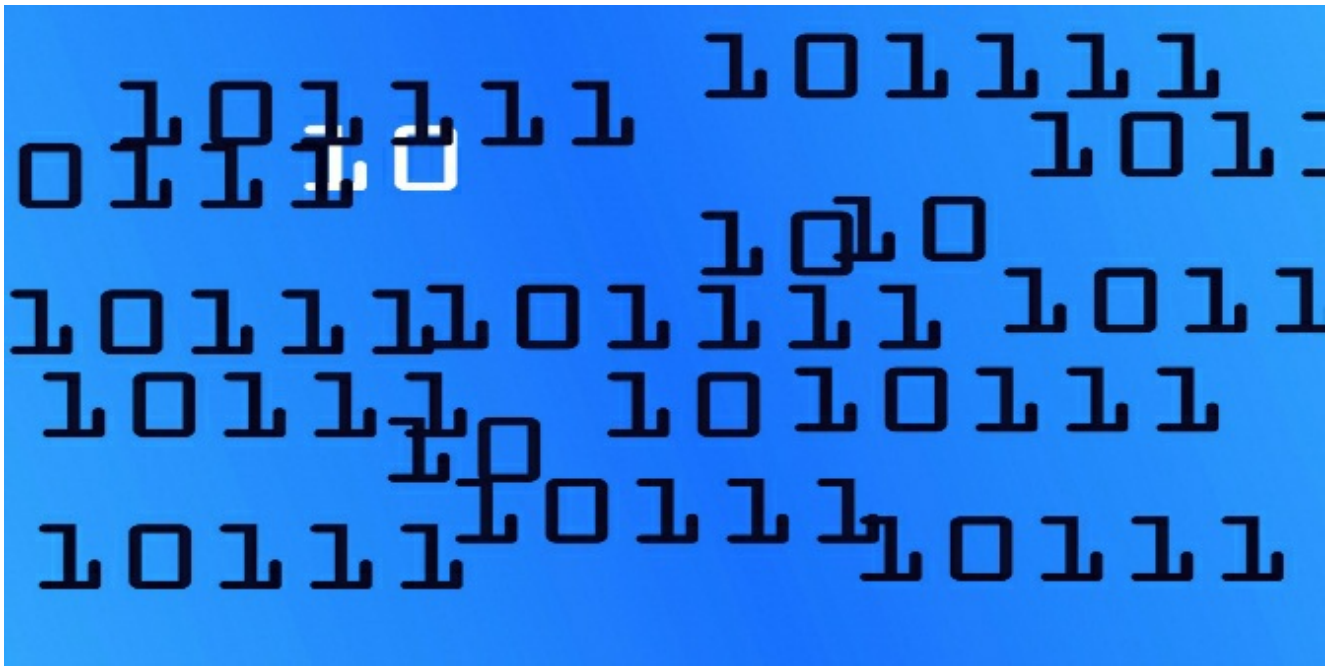

Image credit: Binary World (CC0 public domain)

Moreover, we raise several important concerns with respect to relying on Big Data as a decision and policymaking tool. While in the abstract Big Data is comprehensive and complete, in practice *today's* version of Big Data has several features that should give public sector practitioners and scholars pause. First, most of what we think of as Big Data is really 'digital exhaust' – that is, data collected for purposes other than public sector operations or research. Data sets that might be publicly available from social networking sites such as Facebook or Twitter were designed for purely technical reasons. The degree to which this data lines up conceptually and operationally with public sector questions is purely coincidental. Use of digital exhaust for purposes not previously envisioned can go awry. A good example is Google's attempt to predict the flu based on search terms.

Second, we believe there are ethical issues that may arise when researchers use data that was created as a byproduct of citizens' interactions with each other or with a government social media account. Citizens are not able to understand or control how their data is used and have not given consent for storage and re-use of their data. We believe that research institutions need to examine their institutional review board processes to help researchers and their subjects understand important privacy issues that may arise. Too often it is possible to infer individual-level insights about private citizens from a combination of data points and thus predict their behaviors or choices.

Lastly, Big Data can only represent those that spend some part of their life online. Yet we know that certain segments of society opt in to life online (by using social media or network-connected devices), opt out (either knowingly or passively), or lack the resources to participate at all. The demography of the internet matters. For instance, researchers tend to use Twitter data because its API allows data collection for research purposes, but many forget that Twitter users are not representative of the overall population. Instead, as a recent Pew Social Media 2016 update shows, only 24% of all online adults use Twitter. Internet participation generally is biased in terms of age, educational attainment, and income – all of which correlate with gender, race, and ethnicity. We believe therefore that predictive insights are potentially biased toward certain parts of the population, making generalisations highly problematic at this time.

In summary, we see the immense potential of Big Data use in the public sector, but we also believe that it is context-specific and must be meaningfully combined with administratively collected data and purpose-built 'small data' to

have value in improving public programmes. Increasingly, public managers must know how to collect, manage, and analyse Big Data, but they must also be fully conversant with the limitations and potential for misuse.

This blog post is based on the authors' article, 'Big Data in Public Affairs', published in Public Administration Review (DOI: 10.1111/puar.12625).

Note: This article gives the views of the author, and not the position of the LSE Impact Blog, nor of the London School of Economics. Please review our comments policy if you have any concerns on posting a comment below.

About the authors

***Ines Mergel** is full professor of public administration at the University of Konstanz's Department of Politics and Public Administration. Mergel focuses her research and teaching activities on topics such as digital transformation and adoption of new technologies in the public sector. Her ORCID id is 0000-0003-0285-4758 and she may be contacted at ines.mergel@uni-konstanz.de.*

***Karl Rethemeyer** is Interim Dean of the Rockefeller College of Public Affairs & Policy, University at Albany, State University of New York. Rethemeyer's primary research interest is in social networks and their impact on political and policy processes. His ORCID iD is 0000-0002-5673-8026 and he may be contacted at kretheme@albany.edu.*

***Kimberley R. Isett** is Associate Professor of Public Policy at the Georgia institute of Technology. Her research is centred on the organisation and financing of government services, particularly in health. Her ORCID id is 0000-0002-7584-0181 and she may be contacted at isett@gatech.edu.*