Big Data, Large-Scale Text Analysis, and Public Health Research

"Big data" is often an amorphous catchall. And lately, it is one that has attracted more criticism than praise. Some have pointed out that a torrent of new data hardly guarantees good insights. Others point to its threats to privacy, often in service of highly targeted and invasive marketing. Then there are those who simply question the hype.

These criticisms are valid but can overshadow the potential of big data. I examine avenues that big data opens for public health researchers, especially those in health promotion. Boiled down to its essence, the term encompasses two revolutions that are undeniably real: more computational horsepower and a greater volume of data.[1] Here, I focus on text, the volume of which has boomed in the digital era, and the analysis of it, often called "natural language processing." Although I provide a necessarily survey-level scan, I offer enough depth so that readers can familiarize themselves with emerging technological trends and methods.

HIGH-PERFORMANCE COMPUTING AND NOVEL DATABASES

New computing infrastructure is required to handle ever-growing amounts of text, including social media chatter, online ramblings, digitized periodicals, Internet ads, and scanned paper documents. Fortunately, the past decade has seen the rise of computing consortiums that allow us to harness any number of computers—from a single server to tens of thousands—for a single task. Commonly referred to as "high-performance computing," this pooled computing power is now accessible to researchers of all sorts, including those without enormous monetary resources. One prime example is the Open Science Grid, a collection of research universities that share their computing resources.[2] The Open Science Grid's staff and campus liaisons provide assistance to researchers with all levels of ability. In practical terms, this means a researcher can open up his or her laptop, distribute a task from a home institution across a grid of servers, and complete it in a fraction of the time it might have taken a couple of decades ago. Concurrently, private firms now sell access to as much computing as a researcher needs, along with technical assistance, all at a relatively low cost. Amazon Web Services has been a leading player for years, with Google Cloud, Microsoft Azure, and others offering similar services.

Paralleling the rise in computing power has been the advent of novel database architectures that can be adapted to increasingly diverse forms of data. Until recently, most researchers have used some version of a relational database: tabular spreadsheet-like databases with individual cells, often queried using the Structured Query Language (SQL). For decades, relational databases have proven useful for holding information about individuals and their various traits, such as those typically recorded in a longitudinal data set, and for data in static form. For more complex and "unstructured" data types, however, SQL databases exhibit severe limits. Single cells in a grid are inefficient ways to hold, for example, the contents of a 200-page legal text. Further, many relational databases are constrictive because of limits on the number of observations and their frequent inability to dynamically add additional fields if needed.

Nonrelational databases, often called "NoSQL," offer a solution to these problems. Most of them contain data conforming to JavaScript Object Notation (JSON), a nontabular standardized format that holds data in labeled fields, such as "title:," "publication date:," and "publisher:."[3] Dynamic updating of the field list is easy and can be done in real time, and there is no limit to the number of fields or total entries in a database. In addition to this flexibility, most nonrelational databases (e.g., the popular MongoDB), are available free of charge with robust online support communities.

RAPID PROCESSING

Tasks that are too overwhelming for a personal computer can now be distributed over high-performance computing grids like those described. Open-source utilities such as Docker, HTCondor, and Hadoop can facilitate the simultaneous operation of an application across any designated number of servers. One recent high-performance computing application is ToxicDocs, a publicly available data set of once-secret corporate documents on industrial poisons that have emerged from the vaults of major multinational corporations.[4] ToxicDocs's creators confronted the task of rendering millions of pages of documents full-text searchable. This typically requires using optical character recognition software that converts images of letters and numbers into actual recognizable characters. Optical character recognition, however, is an inherently slow process. But by deploying the task on a

ABOUT THE AUTHOR

Merlin Chowkwanyun is with the Center for the History and Ethics of Public Health, Department of Sociomedical Sciences, Mailman School of Public Health, Columbia University, New York, NY.

Correspondence should be sent to Merlin Chowkwanyun, Center for the History and Ethics of Public Health, Department of Sociomedical Sciences, Mailman School of Public Health, Columbia University, 722 West 168th Street #R931, New York, NY 10034 (e-mail: mc2028@columbia.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This editorial was accepted January 8, 2019.

doi: 10.2105/AJPH.2019.304965

high-performance computing grid, the project shortened what would have taken months on a single desktop computer to a handful of days.[5]

AUTOMATIC DOCUMENT CLASSIFICATION

Large amounts of data often need to be sorted into discrete categories. In the previously described ToxicDocs example, it would be useful not just to possess an enormous cache of documents but also to know whether a certain artifact is a newspaper article, e-mail, scientific article, or internal memorandum. Automatic classification combines older and newer techniques as follows. One typically starts with a team of humans manually analyzing a sample set of documents, known as "training data," while affixing categories to them in a fashion not unlike traditional coding. With more classifications, a probability model then identifies distinctive characteristics in each type of document category (e.g., special characters and formatting in e-mails) and, over time, increases the accuracy of a category guess. Many techniques exist for identifying recurrent characteristics across documents. An increasingly common one is support vector machine classification, in which words and phrases are converted into mathematical representations in a vector space, which are then used to calculate similarity and dissimilarity to other words and phrases. In a health promotion context, a hypothetical researcher might analyze thousands of health department complaints from constituents and automatically slot them into discrete categories, for example, by complaint type (noise, vermin).

NAMED ENTITY RECOGNITION

Numerous robust tools for natural language processing are available to researchers, such as the NLTK (Natural Language Toolkit) and spaCy libraries for the Python computer language.[6,7] They enable rapid analysis of text, ranging from identifying recurring words; detecting their parts of speech; and characterizing nouns as places, organizations, or people. In massive databases, this allows the rapid isolation of commonly used phrases or names of people who might otherwise be an anonymous blur. For those interested in beliefs about certain health practices, named entity recognition could isolate commonly invoked authors on bulletin boards where users regularly swap health information of varying quality, among dozens of other applications.

SOCIAL MEDIA AND METADATA

Finally, social media contain a trove of data, and what is beneath the surface is often as interesting as an utterance itself. Behind a single tweet, for example, is information in a JSON format that contains not only a username and the text of a tweet but also its date and time, the geographical location, the number of times a tweet was liked or retweeted, and its hashtags, among other information. Using the techniques I have described, one can use these data to reconstruct influence networks—for instance, who spreads antivaccination messaging and with what phrases—and identify the relative sway of different purveyors of information.

FUTURE CHALLENGES

For all the potential they carry, these techniques—and big data more broadly—raise critical questions about current institutional structures. One is the question of standards. How will the researchers adjudicate among several ways of doing something? What kind of protocol transparency ought to be mandatory, if any? Then there is the need to update training in both public health and computer science and other related fields. This will facilitate collaboration among future generations, who can use new computational techniques while respecting the longstanding norms of the public health profession. Considering that big data is in its infancy, these issues will occupy much attention in the coming years. But there is no question that big data is here to stay and, with it, an enormous opportunity for all in public health. AJPH

Merlin Chowkwanyun, PhD, MPH

ACKNOWLEDGMENTS

I wish to thank Alex Farrill for clarifying discussions of nonrelational databases with me. Yoka Tomita assisted with preparation of the manuscript, and conversations with Nora Landis-Shack shaped much of my thinking on these issues.

CONFLICTS OF INTEREST

The author has no conflicts of interest to declare.

REFERENCES

1. Barrett MA, Humblet O, Hiatt RA, Adler NE. Big data and disease prevention: from quantified self to quantified communities. *Big Data*. 2013;1(3):168–175.

2. Juve G, Rynge M, Deelman E, Vockler J, Berriman GB. Comparing FutureGrid, Amazon EC2, and Open Science Grid for scientific workflows. *Comput Sci Eng*. 2013;15(4):20–29.

3. Crockford D. Introducing JSON. 2019. Available at: https://www.json.org. Accessed January 24, 2019.

4. Chowkwanyun M, Markowitz G, Rosner D. *ToxicDocs: Version 1.0*. [database]. New York, NY: Columbia University and City University of New York; 2018.

5. Rosner D, Markowitz G, Chowkwanyun M. Toxicdocs (www.toxicdocs.org): from history buried in stacks of paper to open, searchable archives online. *J Public Health Policy*. 2018;39(1):4–11.

6. NLTK Project. Natural language toolkit (NLTK). 2019. Available at: https://www.nltk.org. Accessed January 24, 2019.

7. spaCy. Industrial-strength natural language processing in Python. 2019. Available at: https://spacy.io. Accessed January 24, 2019.