Big Data and the World of Social Media

February 27, 2015

johannes Eichstaedt

“Big Data” eventually will fall out of the hype cycle — but for now, it’s everywhere. In a high-profile manifesto in *Science*, social scientists were asked to step up and “leverage the capacity to collect and analyze data with an unprecedented breadth and depth,” be it through Tweets, Facebook statuses, or cell phone records. We can understand individuals through thousands of data points, and society-scale processes through millions of individuals, institutional review boards permitting. The *Harvard Business Review* famously declared “data scientist” to be the sexiest job of the 21st century, suggesting that data science is the way of the future.

Our research group has certainly contributed to the hype; at the end of last year, we’d shown in a *Journal of Personality and Social Psychology* paper that language-based predictions of personality from social media are about as accurate as friend reports of personality. In January of this year, we published evidence in *Psychological Science* suggesting that Twitter language can predict communities’ heart-disease rates, a finding of which even *The Onion* took note. In 4 years of working with computer scientists, what have we learned about the process? And are psychologists still needed in this world of Big Data?

The World Well-Being Project

In 2009, a team of researchers at Google demonstrated that Internet search queries could be used to estimate the prevalence of flu: People searching for a limited number of terms relating to flu symptoms and remedies not only tracked variation of influenza in time and space as reported by the US Centers for Disease Control, but did so instantaneously, without the reporting lag associated with aggregating hospital reports. After Martin Seligman gave a talk at Google in 2010, a number of researchers at the

Positive Psychology Center at the University of Pennsylvania and at Google.org decided to see whether a similar method could help them estimate the variation in well-being in time and space. After one collaboration, we — the members of the Penn research team — decided to develop the infrastructure to tackle the problem ourselves. We founded the ambitiously titled World Well-Being Project (WWBP) in 2011 to measure the psychological states of large populations via natural-language processing and machine learning over large social-media data sets.

In the beginning, we focused on building our own infrastructure code base and calibrating our methods. Our first publication on using Twitter to measure geographical variation in well-being appeared in 2013, as did a study identifying the language that characterized age, gender, and personality on Facebook across 75,000 users. Both received a fair amount of media attention — more than we had expected.

We were on to something, and in 2014, we received a $3.8 million grant from the Templeton Religion Trust to further this line of research. Our 3-year grant objectives are to use social media to develop the measurement of mental and physical well-being, to understand its determinants and correlates, and to share our methods with the social-science research community, making them widely accessible.

Since then, WWBP has grown to our current team of 11 full-time members, including two computer science postdocs and two psychology postdocs.

Algorithms Versus Humans

Over the years, we have learned a lot about the joys and difficulties of working on a highly interdisciplinary project. To say it up front, working across the boundaries of computer science and psychological science is challenging. The two fields have a very different sense of what makes good research. We value different things; we publish differently.

We like to say that we spent the first 2 years of our collaboration just learning to talk to one another, and in many ways that’s not an exaggeration. Few psychological scientists understand why you would care about cross-validating your results or why you would lock part of your data away from the beginning to test your methods at the end. Computational linguists sometimes write sentences like “psychologists study author attributes,” because to them that’s what humans are: the makers of text.

In computer science, having humans involved in any part of the generation of study results (as curators of a specific set of results) often goes against a professional code: If you need humans, you’ve failed to develop the right algorithms; data should tell their own story. Methods that lead to insight should be adopted bottom-up, not as the result of previous theory. If humans are meddling in the process, computer scientists don’t trust the results as much. Better to show the reader an unsorted list of correlations and have them draw their own conclusions (lest you meddle).

When it comes to interpretation, psychologists are often inclined to do things computer scientists feel uneasy about. For example, the use of the words “apparently” and “actually” are correlated with higher scores in neuroticism. Think about that for a moment — if you are anything like the author, you’ve just formed a theory in your head about why that may or may not make sense in light of what you know about people. You probably did not take into account the ambiguity of words and the many roles they can play in language. We’ve often found psychologists, when presented with such correlations, to be

eager to generate rich and entirely underdetermined hypotheses post hoc — there are many reasons why use of any one word may be correlated with some outcome, including word sense ambiguities or broader language-use confounds tied to third variables. Often the reluctance of computer scientists to interpret single language results has proven to be wise counsel.

Of course, psychological scientists also scratch our heads at the ways of computer science. Computer scientists do publish papers about things we care deeply about, like predicting depression from what people write on social media. In a computer science publication, such findings are recognized as important in a quick opening paragraph, World Health Organization statistics and all, and then it's on to prediction models and accuracies. What depressed people actually write sometimes gets only the most cursory mention. What processes are implicated by their language use, and how can we understand manifestations of depression better in the modern, digital world? In short, what can we learn about depression itself? You often won't read about those questions in a report written by computer scientists, for whom words and phrases are often just fodder for machine-learning models; the accuracies of these models provide the meat of such an article. Increasingly, though, computer scientists in different research groups are reaching out to collaborate with domain experts (like psychologists) to help contextualize and interpret data-driven findings. Semantic structures (like topics) generalize beyond single words and can be acceptable units of interpretation in the hands of psychologists. Seeing these different perspectives and skill sets, you can understand why collaboration between psychological scientists and computer scientists might be a game changer.

Learning New Ways

Psychologists also publish differently than do our colleagues in computer science. Take computer science conferences, for example. In natural-language processing, our academic neck of the woods, the most prestigious outlets for this work are not journals but conferences, such as the annual meeting of the Association for Computational Linguistics (ACL) or its North American Chapter (NAACL), of which few psychological scientists have ever heard. Conferences have submission deadlines, which means that you can't wait to submit new research until you feel it's done, as you can in psychology — unless you'd like to wait a year until the next conference comes around. The peer-review process is fast, and it follows a strict timeline. All papers become immediately and publicly available online after publication; other researchers respond quickly. Data are shared openly, and sharing is often expected upon acceptance into a conference. Failing to share your data is considered poor form. There are big lessons here for psychology, where results are often disseminated at a glacial pace.

In our own collaborative research, we have generally settled into a natural publishing routine: Our computer scientists take the lead on publishing papers about methods and prediction accuracies and about introducing new problems to computer science. We psychological scientists apply the methods that develop insight into psychological problems and processes. When it comes to interdisciplinary papers, we write and quibble over them together.

Learning From One Another

In trying to work together, we've learned a lot. We as psychological scientists have learned to more proactively avoid "overfitting" — the danger of capitalizing on chance not only when the number of

variables exceeds the number of observations, but also when many hypotheses are tested simultaneously. And we've begun to understand the power of a good clustering algorithm to carve nature at its joints — at its best, it seems like magic.

Psychological scientists continue to act as the grand masters of construct validity — triangulating a subtle explanatory construct out of the stuff of our lives, with divergent and convergent validity across language, behavior, and real-life outcomes. We don't mind that absolute truth is unattainable. For computer scientists, absolute truth is presumed axiomatically, and it often doesn't rise beyond the label an "MTurker" gave to a piece of text (indicating, say, how much "optimism" a given Tweet expresses). Psychological scientists are trained to disentangle the pesky complexity of humans; no clustering algorithm can develop nuanced theory. But algorithms can pull out clusters of language that can be mapped onto theory, and algorithms can suggest that something might be missing from a theory. Psychological scientists are uniquely equipped to broker deals between theory and the data.

A World of Data

In a way, psychological scientists have always been data scientists: With regression, *t* tests, and ANOVAs, our methods fit our data. But as we work with larger and larger data sets, our methods need to grow, too. For the data of today, SPSS is not an option. SQL, the most popular database language, takes an afternoon to learn. And once you've learned a little Python (two afternoons), you will realize it can do everything in two human-readable lines — except make you coffee. Once you've figured out these basic Big-Data-handling skills, you can begin to interface with the tools and infrastructure computer scientists have developed for text analysis. It's really not that hard to get started. We, as a research lab, want to help to bring Big Data text-analysis methods to psychological science, so that our field can benefit from a shared understanding in how to apply these methods. We will soon start making tools, introductory resources, and demos for quick exploration available at lexhub.org. On a quiet afternoon, have a look.

If we want to work on the Big Data of today and tomorrow, we have to continue to be data scientists with methods that fit the data. Psychology, the discipline dedicated to those pesky "author attributes," needs to have a voice everywhere decisions about people are made with data — and not just in marketing. As data becomes the backbone of our economy and even our democracy through increasingly targeted campaigns and predictions of individual behavior, psychological scientists need to make their voices heard. In the end, some of the biggest challenges psychologists and computer scientists will need to address together are ethical: As our methods give more and more fine-grained insights into the private lives of populations — even if just using "public" Twitter data — how do we set boundaries for ourselves? How do we honor principles of "informed consent" in the age of massive, public data sets? Psychologists have a critical role to play in these conversations.

Psychological science is where biology was 10 years ago, when it moved from the study of single gene sites to data-driven discoveries across the genome, and from simple statistics to bioinformatics. Psychology too is moving towards *psychoinformatics*, *digital epidemiology*, or *infodemiology*. The field is nascent enough that you still get to pick your favorite term — or make up your own.

Acknowledgements

The author gratefully acknowledges his esteemed colleagues Greg Park, Andy Schwartz, and David Yaden for their kind feedback and thoughtful suggestions.