SURVIVING THE NEXT AVALANCHE: SKILL DEVELOPMENT AND THE BRAVE NEW WORLD OF ADMINISTRATIVE DATA

MICHAEL HAAN

Michael Haan, PhD, is Associate Professor and Canada Research Chair in Migration and Ethnic Relations at Western University. His interests include migration, immigration and the Canadian labour market. He is also Director of the Collaborative Graduate Specialization in Migration and Ethnic Relations at Western, and Academic Director of Western's Statistics Canada Research Data Centre.

In a 1982 article, Canadian philosopher Ian Hacking documented staggering increases in the amounts of data being collected at the turn of the 19th century, referring to it as an "avalanche of printed numbers." Between roughly 1800 and 1830, he showed, the amount of statistical data increased exponentially, at a time when the written word was only rising in a linear fashion. Undoubtedly, this was a scientific revolution, consistent in nature and magnitude to what Thomas Kuhn described in *The Structure of Scientific Revolutions* (1962). Ways of seeing and thinking about the world changed; it not only became possible to collect and store reams of data, but it became desirable, even essential, for modern and modernizing nation-states to count and measure whatever they could. Consider that the United States (1790), the United Kingdom (1801), and France (1801) all held their first national enumerations around this time.

Unlike other scientific revolutions, however, this one did not initially yield material products. Sure, census results were published in print form, but the outputs were minimal compared to the data being collected. There was nothing like the creation of a printing press or steam engine. All we have from that era are stacks of now dusty books, with countless cross-tabulations. This is because a lot of the tools that we use today to understand statistics didn't yet exist. Although averages existed, medians, standard deviations, variances, correlations and regressions did not; even the word 'statistics' was

not part of the English language until the late 18th century, when it was adapted from the German word *statistik*.

The sheer volume of data being collected, along with a relative absence of techniques for working with these data, created the conditions for the pioneers of statistics to flourish. Brilliant scholars, such as Englishmen Francis Galton and Karl Pearson, developed a statistical infrastructure, much of which remains intact today.

What helped give the 19th century statistical data revolution traction was that it gave us a new way of thinking about the world, a new way of knowing. Gone were the days of indeterminism, as we entered a world of probabilities and likelihoods. Statistics became the ontological engine of the day.

I believe that we are currently at a similar point in history. We now have more data than ever before, and the tools required to analyze these data are either absent entirely, or not commonly used in many disciplines. What this means for social scientists is that we will either need to update our toolkits, or watch as other disciplines speed past us.

THE BIG (OFTEN ADMINISTRATIVE) DATA REVOLUTION

As I argue above, the pace at which data are being collected has once again created a situation in which techniques lag behind data development. The new data regime holds considerable promise, by allowing us to ask and answer questions that were previously impossible. That said, these new opportunities aren't free; big and/or administrative data are generally not collected primarily for research purposes, and as a result are often messy, disparate and poorly documented.

Big data files are not only huge, but also exceedingly complex. The Longitudinal Immigration Database (IMDB), for example, is more than 30 gigabytes once the different tax years have been linked to temporary resident files and permanent resident landing records. Most software packages are simply unable to work with such large files; one does not merely fire up SPSS* and begin running crosstabs on the IMDB. When my students begin working with these files, they often quickly realize that they don't have the skills to work with such data, requiring months of remedial work. This is not a negative reflection on them, but it instead points to how they are not receiving the necessary skills in their coursework.

Users now need to know about complex forms of missing data, probabilistic versus deterministic data linkages, often with less than perfect documentation. Students, at least in my university, don't often receive this training, and are instead forced to learn these things on the job or not at all. Looking forward to the next 100 years, it is difficult to know with certainty what the future will bring, but below are some current trends that I believe will disrupt business as usual for many of us.

ANSWERING RESEARCH QUESTIONS WITHOUT AN EXPLICIT THEORETICAL FRAMEWORK

Big data often allow us to ask questions that we have no theoretical guidance on. For example, what (eerily accurate) shows will Netflix recommend next, when nearly the only thing they know about you is what you've watched in the past? In the old days, when all we had were the techniques

developed during and after the Avalanche, we would have used the age, sex, place of residence, income, and just about anything else we could get about a viewer to develop a predictive framework for suggesting shows.

Now, most of this information is either used tangentially, or not at all. If Netflix's algorithms do a better job at making suggestions than a well-grounded empirical framework (and, anecdotally, at least for me, they seem to), then numerous disciplines need to fundamentally rethink their curriculum. What role does a theory course play in a social science discipline when there is no obvious link between a theoretical framework and the methods that social scientists will use to get answers?

Although this seems counter-intuitive to how social science research is conducted, it can also be seen as an opportunity to think differently about what research is.

WHITHER INFERENTIAL STATISTICS?

It is increasingly likely that future datasets will represent entire populations, not just samples of interest. What this means is that the results we derive will necessarily be an accurate representation of a population, removing the need for the use of inferential statistics. By definition, any trend derived from population data are statistically significant, because they include everyone that we're trying to generalize from.

MACHINE LEARNING

Simply put, machine learning will likely put some of us, or our students, out of work. Machine learning definitions vary, but they all seem to hold autonomous learning by computers as the centrepiece. This means that humans will no longer need to tell computers what to do, once they've loaded all the necessary algorithms. Reams of data will be analyzed without much input from a human analyst. Without sounding too much like science fiction, computers are beginning to teach themselves, and this will impact how data are collected and analyzed in the future.

There is considerable excitement around machine learning/artificial intelligence, with uses that could range from employing a machine learning algorithm to providing recommendations for law and policy, to finding the best place to live. Judges could make fairer decisions if they are advised on outcomes. Immigrant retention might increase if they can calculate what country they'd be most likely to succeed and be happy in.

COMPETITION FROM OUTSIDE THE SOCIAL SCIENCES

Most of us, as social scientists, have lived fairly peaceful lives within our disciplines. Each of us has a subject area, and everyone else more or less leaves us alone. Individuals? Everybody knows that the study of individuals belongs to psychology. Nations? Political science gets that one. Society? That's sociology; everybody knows that. But what if sociology could do more — take the stage for once?

If you believe that only 0.5 per cent of all data are currently being analyzed (Marr, 2015), and that Google alone handles and saves the parameters of 40,000 searches every second, it would seem that social science analysts will be more outmatched by data than ever before in the big data era. There are

two paths: we can let this opportunity slip away and remain in our individual worlds, or we can engage with this new frontier and use it demonstrate the value of social science.

CONCLUSION

Statistics Canada recently celebrated its 100th birthday, and should be congratulated for its many contributions in the fields of data collection, management and analysis. We in Canada not only have a centenarian statistical agency, but also an internationally recognized one. At the same time, this is hardly the time to pause: private data collection agencies are extremely active, and it's only a matter of time before we render ourselves out of existence. We also have a fairly widespread agreement that statistics are an essential component of modern democracy (just think of the backlash to the Harper government's decision to cancel the 2011 Long Form Census as proof of this). What we now need is a collective increase in the skills required to thrive in this space, to ensure that it and we are also able to survive the next avalanche.

REFERENCES

*SPSS — STATISTICAL PACKAGE FOR THE SOCIAL SCIENCES, a leading software package used for analyzing statistical information.

HACKING, IAN. 1982. "Biopower and the avalanche of printed numbers," *Humanities in Society* 5:279–295.

KUHN, THOMAS. 1962. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

MARR, B. (2015). "Big Data: 20 mind-boggling facts everyone must read," *Forbes* magazine. www.forbes.com/sites/bernardmarr/2015/09/30/big-data-20-mind-boggling-facts-everyone-must-read/#39c4da3a17b1

STATISTICS CANADA. "History of the Census." www150.statcan.gc.ca /n1/edu/power-pouvoir/ch2/history-histoire/5214912-eng.htm. Retrieved May 7, 2019.