Empirical Research using Big Data in Canadian Political Science

Mickael Temporão, PhD
Director of Data Science, Delphia

September 8, 2019

The quantity of data generated by contemporary digital systems has been growing exponentially in the past years, creating extraordinarily large and complex datasets also known as Big Data (Groves, 2011; Keller et al., 2017; Lazer and Radford, 2017). These datasets combined with new methods have provided scholars with the ability to study texts, images, video, and audio and have substantially improved the state-of-the-art in many domains (LeCun, Bengio and Hinton, 2015). The past decade has seen Canada pioneering empirical research in the field of Artificial Intelligence fostering advances in Machine Learning, more specifically, establishing Deep Learning as a standard to discover patterns in massive data sets. But, can Big Data be leveraged to conduct valid Canadian Empirical Political Science Research? Canadian Political Science scholars have been slower to adapt to this new *opportunity data*[1]. This can partly be explained by the fact that the knowledge required to leverage those large datasets is broader than classical Statistics and Political Science. Yet, the generation and collection of valid and reliable data to test theories is at the heart of empirical research in Canadian Political Science and remains a major challenge for the empirical tradition. Large datasets alone can't solve social problems but can provide opportunities to explore new phenomena, previously invisible to datasets on a smaller scale (Grimmer, 2015).

Empirical research in Political Science often relies on surveys and random samplings. The continuous growth in popularity of alternative modes of communication makes random sampling increasingly untenable. Compared to classic survey based data, Big Data is characterized by a much lower information to data ratio. However, classic survey instruments are becoming deprecated as the quality of the data available decreases, while at the same time the quantity of data grows exponentially. Big Data can be an alternative, but those data are nearly universally non-random which bring new methodological challenges to the table in order to make generalizable claims. While descriptive inference is often denigrated in Political Science, we need to properly describe this new breadth of data in order to better understand its biases and allow us to have a chance to think about causal inference.

[1] See Keller et al. (2017) for more details on *opportunity data*.

Political Scientists have started make use of this new breadth of data in their research in order to test theories that were previously impossible due to the lack of data (Rheault and Cochrane, 2019; Temporão et al., 2018). For Big Data to help empirical research in Political Science, scholars need to acknowledge that the field is evolving into something broader than Political Science, including machine learning frameworks, that borrow from recent advances in Statistics and Computer Science. This new interdisciplinary research field that is emerging is called *Computational Political Science*. This field of study combines recent advances in Computer Science, Statistics, Linguistics, Information Technology, and many others, to develop new and scalable methods that facilitate the extraction of valuable insights from Big Data and enable the study social problems in Political Science.

Canada can be defined by its diversity (Dufresne et al., 2018). Big Data has the potential to foster empirical research in Canadian Political Science by providing the breadth and depth required to explore and study the heterogeneity among the Canadian mosaic and help improve our understanding of this socially rich context.

Big Data increases the scope of the data available for research, and the challenges associated withe these datasets can lead to increased methodological standards enabling scholars to leverage large-scale heterogeneous non-probabilistic samples (Ruths and Pfeffer, 2014; Shiffrin, 2016). Big Data recently enabled empirical studies at sub-national levels, which have been understudied due to data scarcity, as showcased in recent work on election forecasting allowing daily forecasts to be made at the Provincial or even at Riding level (Temporão et al., 2019). Finally, Big Data also helps to study new types political actors for which data was not previously available and allow the exploration of heterogeneity within a nation-state across time or or even across varying languages (Temporão et al., 2018; Rheault and Cochrane, 2019).

These new opportunities for empirical research in Canada are promising enough to warrant further investigation into the potential of Big Data and can serve as initial interrogations to build and test further theories.

References

Dufresne, Yannick, Charles Tessier, Alexandre Blanchet and Mickael Temporão. 2018. "The symbolic mosaic: an empirical typology of national symbolic webs and its effect on vote choice in Canada." *National Identities* 21(4):329–345.

Grimmer, Justin. 2015. "We are all social scientists now: How big data, machine learning, and causal inference work together." *PS: Political Science & Politics* 48(1):80–83.

Groves, Robert M. 2011. "Three eras of survey research." *Public Opinion Quarterly* 75(5):861–871.

Keller, Sallie, Gizem Korkmaz, Mark Orr, Aaron Schroeder and Stephanie Shipp. 2017. "The evolution of data quality: Understanding the transdisciplinary origins of data quality concepts and approaches.".

Lazer, David and Jason Radford. 2017. "Data ex machina: Introduction to big data." *Annual Review of Sociology* 43:19–39.

LeCun, Yann, Yoshua Bengio and Geoffrey Hinton. 2015. "Deep learning." *Nature* 521(7553):436–444.

Rheault, Ludovic and Christopher Cochrane. 2019. "Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora." *Political Analysis* pp. 1–22.

Ruths, Derek and Jürgen Pfeffer. 2014. "Social media for large studies of behavior." *Science* 346(6213):1063–1064.

Shiffrin, Richard M. 2016. "Drawing causal inference from Big Data." *Proceedings of the National Academy of Sciences* 113(27):7308–7309.
URL: *http://www.pnas.org/content/113/27/7308*

Temporão, Mickael, Corentin Vande Kerckhove, Clifton van der Linden, Yannick Dufresne and Julien M. Hendrickx. 2018. "Ideological Scaling of Social Media Users: A Dynamic Lexicon Approach." *Political Analysis* 26(4):457–473.

Temporão, Mickael, Yannick Dufresne, Justin Savoie and Clifton van der Linden. 2019. "Crowdsourcing the vote: New horizons in citizen forecasting." *International Journal of Forecasting* 35(1):1–10.