Background

Big data analytics can contribute to fruitful future scientific inquiries as they allow researchers to investigate new underexplored scientific questions that could not be studied before and to dive deeper into existing scientific questions in order to elicit better answers. In this short paper, I outline the opportunities that arise from leveraging existing econometric methods in combination with big data. In particular, I discuss how the four major features that characterize big data (also known as “4 V’s”) can greatly enhance standard econometric techniques and how big data analytics can vastly benefit from these techniques. More importantly, I outline the possibilities of leveraging existing econometric methods in combination with big data in order to make causal inferences from observational data. For data science and big data analytics to become more useful towards examining causal relations nowadays, I argue that we need to draw on the substantial knowledge base created in the economics and social science fields over the years in order to infer interesting causal effects as simply analyzing large amounts of data does not necessarily help us make better data-driven decisions.

Problem Statement

Data science and big data analytics have the potential to contribute to different research paradigms and facilitate inter-disciplinary research thanks to the characteristics of variety (i.e., variety of data sources) and veracity (i.e., quality of the data) of information. Even though big data alone is usually not sufficient to generate causal inferences, it can be particularly useful for leveraging research designs and statistical techniques that allow data-driven analysis to move from correlations to causation. First, big data can greatly enhance our ability to account for previously unobserved confounders as well as to employ alternative controls and proxies. This is particularly important as consideration of confounding is fundamental to the design and analysis of studies of causal effects (Greenland et al. 1999). Hence, big data can alleviate concerns of potential endogeneity that undermine researchers’ efforts to extract causal relationships from non-experimental data. For instance, researchers can leverage unstructured data through big data and machine learning techniques that allow them to control for various factors that remained unobserved in previous research. Furthermore, the variety and veracity of big data can facilitate the identification of exogenous variations (Einav and Levin 2014) and the utilization of natural experiments in order to generate causal inferences from observational data. For instance, Ghose and Todri (2015) exploit the granular information on the viewability of advertising
exposures (i.e., information on whether a user has viewed the display advertisement that was served to him/her) that enables the employed natural experiment research design. The circumstances that affect the viewability of a display advertising impression serve as an exogenous shock to the firm’s targeting and simulate a natural experiment creating two groups of users: those who view the display ad and those who do not, while both groups are automatically targeted by the same marketing campaign fulfilling certain targeting criteria. Such natural experiment frameworks can avoid the self-selection and other treatment selection biases that are important concerns in analyzing non-experimental data for making causal inferences. Similarly, the access to rich information regarding the advertising exposures (i.e., timing of exposures, location of targeted consumer, etc.) allows the authors to use the exogenous shocks of hyper-local granular weather information in order to employ the instrumental variable panel data method and address concerns for time-varying unobserved confounders. In particular, employing a variety of sources they use information on hyper-local weather for very granular time intervals (i.e., 20 minute time intervals) matched with individual-level advertising exposures as an instrument because weather data is correlated with the (potentially endogenous) explanatory variable of a viewable impression (i.e., browsing the Internet is an activity that competes with other outdoors activities the users might be enjoying) but not correlated with the dependent variable under study.

Apart from facilitating scientific inquiry through the variety and veracity of data, big data analytics methods can further contribute to fruitful future scientific inquiries through the other two unique aspects of big data that can lead to more precise and more accurate estimation of various effects are the velocity and volume of big data. The volume of data refers to the amount of data and determines the potential of the data under consideration, while the velocity of big data refers to the speed of data generation. As far as the volume of big data in analytics is concerned, it can enable the identification of especially small effects, when they exist, resolving to a large extent the issue of statistical power. Furthermore, standard econometric methods can benefit from big data analytics due to the velocity and volume of the data. For instance, big data analytics enhance the quality of the various matching methods since they enable researchers to reliably estimate the treatment effect using the observed counterfactuals that exist in big data sets without the need to depend on model-based extrapolations due to lack of sufficient data. Similarly, instrumental variable methods can benefit from big data for the estimation of the effects beyond the local average treatment effect of intervention. In particular, the increased volume and velocity of observational data can be more informative about the causal effect of treatment thanks to the larger number of individuals whose treatment status has been manipulated with the exogenous shock.

Moreover, big data analytics can enable theorizing and hypothesis development at an unprecedented level of granularity (Varian 2014). Utilizing the granularity of big data, researchers can gain a deeper understanding on phenomena that have been previously studied at a more aggregate level. For instance, big data analytics in many cases could allow researchers to further deepen the exploratory analysis by integrating various relevant contextual factors in their analysis and, hence, have the potential to vastly facilitate hypothesis development as well as data-driven theorizing in future research. For instance,
(Ghose and Todri 2015) leveraging big data find that hyperlocal contextual information, such as weather conditions, affect the browsing activity of consumers. This finding could be a topic of future research that could significantly lead to fruitful research in the future. Besides, the granularity of advertising exposures revealed that on average 55% of the display ads are not rendered viewable. This is particularly important as previous academic research on display advertising assumed that an impression was always viewable to the user whenever loaded and might have led prior research to overestimate or underestimate the true effect of display advertising.

Finally, big data analytics allow researchers to investigate phenomena and study research questions that could not be explored before due to the unavailability of or the limited access to certain information. In particular, one could take advantage of specific institutional details and micro-level variation that would be difficult to isolate and exploit with more aggregate level data. For instance, researchers could tap into the granularity of big data and further develop theories that are more granular. Besides, data science and big data analytics generate novel findings, it also the potential to subsequently further spur academic interest leading to novel scientific inquiry and hypotheses generation creating a positive “feedback loop”. For instance, Ghose and Todri (2015) discovered that one of the effects of advertising is that customers engage in both active and passive information gathering processes. Based on this finding that was enabled by big data analytics, future research could investigate and identify the underlying mechanisms that are related to this effect and cause the corresponding user behavior.

**Broader Impacts**

This paper elaborates on the opportunities for fruitful collaboration between classical econometric techniques and big data in order to make causal inferences and estimations that are more accurate. While experimentation is the gold standard for establishing causal effects, it is not always feasible, ethical or cost-efficient and, therefore, researchers across disciplines have to rely on non-experimental data in order to extract knowledge and advance the theory of their respective scientific fields. Hence, the proposed methods of combining econometric techniques with big data in order to extract causal inferences from observational data is of critical importance when cause and effect are the central questions of the scientific research.

**References**


