# SIMULATION AND BIG DATA: IN SEARCH OF CAUSALITY IN BIG DATA-RELATED MANAGIAL DECSION MAKING

**Maggie M. Cheng, Chenxing Li, Rick D. Hackett**

January 2018

## ABSTRACT

The unprecedented availability of digitized human behavioral data offers new research opportunities for discovering hidden patterns in Big Data that may not be apparent in smaller samples. At the same time, there are potential pitfalls associated with Big Data analytics in the absence of also working to identify causal relationships among the constructs thought to be involved. Indeed, despite the seemingly advanced modeling techniques applied to the analysis of Big Data, they are not well suited to addressing issues of causality. We illustrate the potential issues involved, using the context of human resources selection, in which the relationship between résumé typos and future job performance is of interest. Specifically, using computer simulation methodology, we demonstrate that including résumé typos along with the personality trait of conscientiousness to predict performance is likely to result in adverse impact on job applicants based on their country of birth, without significantly improving prediction. This outcome would leave the employer open to equal employment opportunity lawsuits and raise ethical concerns. In all, we suggest guidelines in which the analytical approaches typically used in the analysis of Big Data be supplemented with experimental and/or statistical approaches better suited to identification of causal relationships.

**Keywords** Big Data, Causality, Simulation, Management Decision-Making

## INTRODUCTION

The unprecedented availability of digitized human behavioral data offers new research opportunities for discovering hidden patterns in Big Data that may not be apparent in smaller samples (George, Haas, & Pentland, 2014; Kosinski, Wang, Lakkaraju, & Leskovec, 2016; Mayer-Schonberger & Cukier, 2014; McAbee, Landis, & Burke, 2017), yet statistical models appropriate to the analysis of Big Data are required, and these often differ from those traditionally used to deal with smaller datasets. Attributes of Big Data include high volume, velocity, and variety (Laney, 2001)—or large and complex data sets with continuous incoming entries from multiple sources. Examples include the growing body of literature that attempts to analyze human behavior using data collected from unconventional sources, including social media (Roth et al., 2016; Jones, Wojcik, Sweeting, & Silver, 2016; McFarland & Ployhart, 2015; Zide, Elman, & Shahani-Denning, 2014), wearable sensors (Chaffin et al., 2017), or facial characteristics (Kosinski & Wang, 2017). In comparison to traditional management literature in which many variables are controlled, and/or control groups are used to strengthen the capability of making causal statements, issues of causality have not been a priority in Big Data research.

We suggest that is it important to supplement the analytical approaches typically used in the analysis of Big Data with those that are better suited to the identification of causal relationships. For example, computer simulation models -- the application of software to model processes, systems, or events (Law & Kelton, 2000) -- provide an opportunity to advance our understanding of the legal and ethical concerns that can arise in dealing with Big Data in management contexts. Specially, Davis et al. (2007) noted that "simulation models can provide superior insight into complex theoretical relationships among constructs, especially when

challenging empirical data limitations exist". These models have the "ability to identify unintended implications [which] accelerates progress in understanding a phenomenon" (Adner et al., 2009; p.204). Well-designed simulation models have both good internal validity (i.e., they adequately capture phenomena observed in management contexts) and external validity (i.e., generalizability to other settings) (Burton & Obel, 2011). The purpose of this paper is to illustrate how simulation models can be a valuable supplement to the types of inferences managers make based purely on the analysis of Big Data. The context for our illustration is human resource (HR) selection.

**Big Data Analysis and Issues of Causality**

Despite the seemingly advanced modeling techniques applied to the analysis of Big Data, including social network analysis, computerized textual data analysis, and deep neural network approaches (Chaffin et al., 2017; Jones, Wojcik, Sweeting, & Silver, 2016; Kosinski & Wang, 2017), these methods alone do not allow for strong causal claims. For instance, Kosinski and Wang (2017) assert that their models have the capability to identify gay men using facial recognition with "91% accuracy", yet there is little consideration of the nature of the causal connection between facial features and sexual orientation. Eichstaedt and colleagues (2015) showed that the Twitter language used in a country, in aggregate, can more accurately predict rates of heart disease than demographic measures from cross-sectional data sources. Jones, Wojcik, Sweeting, and Silver (2016) identified a "pattern" of post-disaster longitudinal Twitter data following traumatic events (e.g. a campus shooting), but did not compare these "patterns" with those following other types of events. This relative lack of concern with regard to issues of causality reflects a larger trend in top-tier behavioral research. For example, Antanokis,

Bendahan, Jacquart, & Lalive (2012) examined a sample of 110 empirical articles concerning leadership with a non-experimental or field experiment design and found that less than 10% of the studies adequately addressed causal threats in their methodology. Some have even suggested that when the dataset is "big enough" to include the whole population as opposed to a small randomized sample, it is possible to "cross the gap between correlation and causation" (McAfee, Brynjolfsson, & Davenport, 2012).

A considerable body of scholarship concerning the requirements that must be fulfilled to establish a causal relationship between two variables x and y, are often brushed aside as evidenced by some of the research we cited earlier. Three classic conditions for establishing causality as noted by Kenny (1979) are that: 1) x must precede y in time; 2) x must be reliably correlated, beyond chance, with y; and 3) other potential causes for the relation between x and y must be ruled out. These criteria help clarify the distinction between correlation and cause; the association between two variables is a necessary but insufficient condition for causation. Others have endeavored to clarify a language of causality in regard to scholarship. For instance, Judea Pearl (2009) stated that it is important for social science researchers to understand the distinction between "A causes B" and "A correlates with B", as well as between "A does not cause B" and "A is independent of B". Further, Wright (1921) and Haavelmo (1943), the founding fathers of simultaneous-equation modeling (SEM; one of the most popular analytical tools in social science research), explicitly noted that SEM only provides a quantitative assessment of causal effects based on known qualitative causal information appraised by other means (e.g. experiment design, a priori studies, theoretical derivation). In all, these well-established basics of causality

are under- referenced in the modern organizational literature, leaving readers to assess whether the association presented in any given study are actually causal, or only seemingly causal.

With the proliferation of statistical models that are being applied to Big Data, the purpose of applying such models needs to be clarified. For example, even though we tend to make causal inferences using regression models, not all of these models are appropriate to establish causation, and inform action. Freeman (2009) has emphasized that statistical models can only inform actions when either design or statistical control is implemented to help ensure causal claims. Relatedly, he identified three purposes of statistical models: 1) data summarization; 2) predicting the future; and 3) predicting the results of interventions. Each of these purposes corresponds to a popular modeling approach in the era of Big Data; i.e., descriptive modeling, predictive modeling, and causal modeling, respectively. Specifically, descriptive modeling summarizes the structure of data in a parsimonious way. Predictive modeling estimates the unknown probability of another variable, using variables that are known. Importantly, only causal modeling aims for causal inferences based on observational studies, natural experiments, and randomized controlled experiments.

Stated most broadly, the issue we are referring to is the problem of endogeneity. Endogeneity occurs when an independent variable is correlated with the error term, which can be a result of measurement error, simultaneous causality, or omitted variables, among other causes. The most common cause of endogeneity is confounding variable(s) that are the true underlying cause of both the independent and the dependent variable. A famous example of the impact of a confounding variable is the observation that the total number of ice cream sales in a city is positively correlated with the rate of drowning in swimming pools. In this example, the

underlying cause of both variables is increases in temperature during the summer; ice cream sales do not cause drowning. While an example of this kind is easy to disprove, if causal analysis is omitted from the modeling process, associations between variables will be subject to potential endogeneity problems. Challenges in correctly interpreting relationships between variables abound across business disciplines as well. The classic example from the field of finance is that the highest correlation with performance of the S&P 500 stock market index over a 10-year period in the 1990s was butter production in Bangladesh (Hope, 2017). Without a theory-based, scientific effort to identify and interpret the actual cause for observed correlations, we can only speak to the concurrence of events. Importantly, for the purpose of informing actions, knowledge of an association provides very little in the way of underlying understanding of the observation.

**An Illustration of the Endogeneity Problem in HR Selection**

Given the increasing availability of big data and related analytical software, a myriad of correlations between variables can easily be "uncovered" using regression models. As noted above, though establishing correlation is an important step toward making causal inferences, it is *only one* of the necessary steps (McAbee, Landis, & Burke, 2017). As such, a belief that correlation is equivalent to cause is dangerous without additional causal analyses. For example, in HR management, the availability of big data has the potential to lead to poor business decisions and legal liability. Employee selection systems resulting in adverse impact against protected classes can result from spurious associations.

Below, we develop and present a computer simulation model for demonstrating the influence of a confounding variable in the context of HR selection. The model explicitly considers the relationship of conscientiousness and résumé typos to employee performance.

While there is a small/moderate causal relationship between conscientiousness (as assessed psychometrically as part of the "Big 5") and future performance (Barrick, Mount, & Judge, 2001; Witt & Ferris, 2003), résumé typos also predict performance in certain contexts (Bersin, 2012). However, résumé typos can be influenced *both* by conscientiousness and language proficiency. Importantly, the latter has little correlation with employee performance. As we will show, the simulation reveals that including typos as an independent variable in the selection process has the potential to result in adverse impact on job applicants based on their birth country (whether or not English is their mother tongue). Adverse impact, in turn, can leave employers open to lawsuits tied to violations in equal employment opportunity law, raise ethical concerns, and create bad publicity for the company. Thus, the failure to establish causal relationships in this context may not only compromise the value of analytics, it may create legal, ethical, and reputational problems as well.

In both the HR and legal arenas, it has long been recognized that adverse treatment of individuals based on "prohibited grounds" including age, race, gender, sexual orientation, mental and physical condition, and marital status, is considered unfairly discriminatory unless the employer is unable to demonstrate that the attribute in question is a *bona fide* occupational requirement for the job (Moreau, 2010). Hence, if such prohibited variables are incorporated as independent variables in analytical models for workforce management and lead to adverse impact against protected groups, these models can be found discriminatory and therefore illegal. Importantly, even if the prohibited variables are not directly included in the model, it is still possible that other variables indirectly result in an adverse effect. For example, use of résumé typos as an independent variable may result in an indirect adverse effect on non-native English

speakers. Such indirect discrimination results from management decisions that are not, on the face of it, based on prohibited grounds, but result in an adverse impact on protected groups (Kossek, & Pichler, 2007; Demuijnck, 2009). If a job requirement results in adverse impact, the employer must show that the predictor used has a direct relationship to the content of the target job and to job performance. Hence the independent variables in analytical models must be carefully-selected and reasonably-supported.

## METHODS

To illustrate the consequences of ignoring causality in the era of Big Data we constructed a simulation derived from concepts discussed in Bersin (2012), wherein the objective was to improve the selection of sales people by using the number of typos found on their résumé as a predictor. In designing the simulation, the following procedures were used: 1) assumptions from the literature were employed to define a data generating process (DGP); 2) based on the DGP, we created the simulation dataset; and finally, 3) we re-estimated the model with simulated data to demonstrate the difference between using conscientiousness and résumé errors to predict performance, considering the effects of language proficiency.

### Defining the Data Generating Process: Model Specifications & Parameters

Table 1 contains the two equations used in the study. Equation 1 reflects the fact that conscientiousness is among the most consistent predictors of job performance. The characteristics of this Big 5 variable and rated job performance were based on two studies, Witt & Ferris (2003) and Barrick, Mount, & Judge (2001). For Equation 2, the Dustmann and Fabbri (2003) investigation concerning language proficiency and the labour market performance of immigrants provided the key parameters. Table 1 shows all the parameters used and their

associated values. In all cases the assumptions reflected in our simulation were based on existing literature. Hence, the DGP of our simulation study is defined as in Table 2.

---

Insert Tables 1 and 2 about here

---

Mapping the DGP as in Table 2, we used the parameters in Table 1 to create a simulated dataset consisting of four variables using the R statistical software. We assumed that conscientiousness and language proficiency are independent from each other, and that language proficiency is independent of performance. To minimize the standard error of the estimate, a population size of 10,000 job candidates was generated.

## RESULTS

The means, standard deviations, and the correlation matrix associated with the dataset are shown in Table 3.

---

Insert Table 3 about here

---

### Comparison of Selection Criteria

We ranked the 10,000 simulated job candidates according to two varying selection criteria: 1) conscientiousness and 2) résumé error. The use of each criterion generates a list of ranked candidates with either conscientiousness from high to low, or résumé errors from low to high. Assuming a yield ratio of 10%, we picked the top 1,000 candidates from each list. Based on the simulated dataset and the assumed model, the average value of performance based on

conscientiousness is 4.00 (*SD* = .45), which is 18% higher than the average performance based on résumé errors (*M* = 3.38). Table 4 shows the mean and standard deviation of the performance scores from two selected pools.

---

Insert Table 4 about here

---

The distributions of the two selected pools are plotted in Figure 1. Plot (a) and plot (b) are notched box and violin plots respectively, demonstrating the difference between the performance outcomes based on the two predictors. Though there is considerable spread in the data in both cases, the majority of the performance values using conscientiousness as a selection criterion lie above those from using résumé errors. The hinges of the box plots represent the first and third quartiles of the data distribution and overlap only slightly between the two data sets. The notched sections, determined using a Student's *t*-test based on a normal distribution, indicate that a significant difference in the median value, within a 95% confidence interval; they do not overlap (Chambers, 1983). The violin plots additionally show the full probability density distribution for each data set, mirrored for clarity (Hintz & Nelson, 1998); though both are unimodal, the standard deviation for the performance as predicted by résumé errors is significantly larger, thus covering a much wider range of performance values.

This is shown further in plots (c) and (d) which compare the performance histograms for each selection criterion, and the kernel probability density distribution of the whole population. In plot (c), we can observe a distinctive right-skewed distribution of the selected group, demonstrating that selection using conscientiousness is effective in finding the best candidates.

In contrast, plot (d) shows that the distribution of the selected group is similar to that of the population, which importantly, suggests that selection based on résumé errors is not a significant improvement over random selection.

---

Insert Figure 1 about here

---

## DISCUSSION

In order to maximize the value of statistical modeling for decision making while ensuring the legality of management practices, organizations should work to identify the true underlying nature of the relationships that emerge from the high volume, high variety, and high velocity data sets. In this regard, we have illustrated the application of computer simulation models. Antonakis et al. (2010) also suggest six methods for inferring causality categorized across two non-experimental designs—statistical adjustment and quasi-experiment. In management practice, finding and controlling for all possible causes of a dependent variable (statistical adjustment) is usually neither possible nor necessary (Cheng, 2017). Therefore, a quasi-experiment design to approximate causality is most often appropriate to management settings. Antonakis and his colleagues further suggest that causal relationships can be identified through SEMs, regression discontinuity, difference-in-differences models, and Heckman selection models, *provided* the underlying assumptions are met.

With regard to the use of data analytics in the context of HR decisions, from an ethical perspective, we call for transparency of all model components, including independent, dependent and control variables, modeling techniques and the other parameters that are applied.

Transparency allows all stakeholders to be involved in debating the legitimacy of the models for legal compliance purposes. At the same time, stakeholders (software solution providers, HR managers, line managers, employees, unions, etc.) should receive training to ensure that they understand fully the models being applied to inform decisions. Finally, data analytic specialists should be included in the multiple stakeholder review process for adopting decision models for operational use.

## REFERENCES


Adner, R., Polos, L., Ryall, M., Sorenson, O. (2001). The Case for Formal Theory. *Academy of Management Review*, 34(2), 201-208.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*(6), 1086-1120.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next?. *International Journal of Selection and Assessment*, 9(1-2), 9-30.

Bersin, J. (2012). How BigData Tools Helps HR Understand You. *Forbes*. Retrieved August 20, 2014, from http://www.forbes.com/sites/joshbersin/2012/02/29/how-bigdata-tools-helps-hr-understand-you/

Burton, R. M., & Obel, B. (2011). Computational modeling for what-is, what-might-be, and what-should-be studies-And triangulation. *Organization Science*, 22(5), 1195-1202.

Chamber, J. (1983). *Graphical Methods for Data Analysis*. Springer: Belmont, CA.

Chaffin, D., Heidl, R., Hollenbeck, J. R., Howe, M., Yu, A., Voorhees, C., & Calantone, R. (2017). The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, *20*(1), 3-31.

Cheng, M. (2017, January). Causal Modeling in HR Analytics: A Practical Guide to Models, Pitfalls, and Suggestions. In *Academy of Management Proceedings* (Vol. 2017, No. 1, p. 17632). Academy of Management.

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing Theory Through Simulation Methods. *Academy of Management Review*, 32(2), 480-499

Demuijnck, G. (2009). Non-discrimination in human resources management as a moral obligation. *Journal of Business Ethics,* 88(1), 83-101.

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., ... Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, 26, 159–169.

Freedman, D. A. (2009). *Statistical models: theory and practice.* Cambridge University Press.

George, G., Haas, M. R., & Pentland, A. (2014). Big data and management. *Academy of Management Journal*, *57*(2), 321-326.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, 1-12.

Hintze, J. L. & Nelson R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2): 181-4

Hope, B. (2017). Inside a quant 'Alpha Factory'. *The Wall Street Journal*, Friday, April 7, 2017, B1-B2.

Jones, N. M., Wojcik, S. P., Sweeting, J., & Silver, R. C. (2016). Tweeting negative emotion: An investigation of Twitter data in the aftermath of violence on college campuses. *Psychological Methods*, 21, 526.

Kenny, D. A. (1979). *Correlation and causation.* New York: John Wiley Et Sons.

Kosinski, M., & Wang, Y. (2017, September 24). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Retrieved from psyarxiv.com/hv28a

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological methods*, 21(4), 493.

Kossek, E. E., & Pichler, S. (2007). EEO and the management of diversity. *Oxford Handbook of Human Resource Management*, The, 251.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note*, (February 2001).

Law, A. M., Kelton, W. D. (2000) *Simulation Modeling and Analysis*, McGraw Hill, New York, NY.

Mayer-Schönberger, V., & Cukier, K. (2014). *Learning with big data: The future of education*. Houghton Mifflin Harcourt.

McAbee, S. T., Landis, R. S., & Burke, M. I. (2017). Inductive reasoning: The promise of big data. *Human Resource Management Review*, 27(2), 277-290.

McAfee, A., Brynjolfsson, E., & Davenport, T. H. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60-68.

McFarland, L. A., & Ployhart, R. E. (2015). Social media: A contextual framework to guide research and practice. *Journal of Applied Psychology*, *100*(6), 1653.

Moreau, S. (2010). What is discrimination? Philosophy & Public Affairs, 38(2), 143-179.

Pearl, J. (2009). *Causality*. Cambridge university press.

Roth, P. L., Bobko, P., Van Iddekinge, C. H., & Thatcher, J. B. (2016). Social media in employee-selection-related decisions: A research agenda for uncharted territory. *Journal of Management*, *42*(1), 269-298.

Sinha, V., Subramanian, K. S., Bhattacharya, S., & Chaudhuri, K. (2012). The contemporary framework on social media analytics as an emerging tool for behavior informatics, HR analytics and business process. *Management, 17*(2), 65–84.

Witt, L. A., & Ferris, G. R. (2003). Social skill as moderator of the conscientiousness-performance relationship: Convergent results across four studies. *Journal of Applied Psychology*, 88(5), 809.

Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, *20*(7), 557-585.

Zide, J., Elman, B., & Shahani-Denning, C. (2014). LinkedIn and recruitment: How profiles differ across occupations. *Employee Relations*, *36*(5), 583-604.

## Appendix

Table 1

Specified Model and Parameter Values Used in the Empirical Simulation

| $Performance = a + b\ Conscientiousness + e$ | (1) |
|---|---|
| 159 | |
| $N_{(CP)}$ | |
| $b_{(CP)}$ .25 | |
| $SD_{(C)}$ .7 | |
| $R^2_{(P\text{-}C)}$ .4 | |
| $E_{(C)}$ 3.75 | |
| $ResumeError = \alpha + \beta_1 LanguageProficiency + \beta_2 Conscientiousness + \epsilon$ | (2) |
| $N_{(R\text{-}LC)}$ 839 | |
| $\beta_1$ .4 | |
| $\beta_2$ .12 | |
| $Var_{(L)}$ .45 | |
| $E_{(L)}$ .7 | |
| $R^2_{(R\text{-}LC)}$ .1 | |

Table 2

Data Generating Process of the Empirical Simulation

$$Conscientiousness \sim N(3.68, 0.7^2)$$

$$LanguageProficiency \sim N(0.7, 0.45^2)$$

$$ResumeError \sim N(1 - 0.8LanguageProficiency - 0.2Conscientiousness, 0.1^2(1 - 0.8))$$

$$Performance \sim N(1.1 + 0.6Conscientiousness, 0.7^2(1 - 0.6))$$

Table 3

Means, Standard Deviations, and Correlations of the Simulated Dataset

| Variable | M | SD | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| | 3.67 | .69 | - | | | |
| 1. Conscientiousness | | | | | | |
| 2. Résumé Errors | .04 | .09 | -.30*** | - | | |
| 3. Language Proficiency | .64 | .33 | .02 | -.65*** | - | |
| 4. Performance | 3.3 | .6 | .68*** | -.21*** | .01 | - |

** $p < .005$. *** $p < .001$.

Table 4

Means, Standard Deviations, and Extreme Performance Values of the Two Selection Criteria

| Criterion | Mean of Performance | SD of Performance | Max of Performance | Min of Performance |
|---|---|---|---|---|
| 1. Conscientiousness | 4.00 | .45 | 5 | 2.70 |
| 2. Résumé Errors | 3.38 | .6 | 5 | 1.41 |
| 3. Population | 3.30 | .6 | 5 | 1.12 |

Figure 1

Comparison of the distribution of two selected pools