Teach an I-O To Fish: Integrating Data Science Into I-O Graduate Education

Juliet R. Aiken and Paul J. Hanges
University of Maryland

Big data is becoming a buzzword in today's corporate language and lay discussions. From individually targeting advertising based on previous consumer behavior or Internet searches to debates by Congress concerning National Security Agency (NSA) access to phone metadata, the era of big data has arrived. Thus, the Guzzo, Fink, King, Tonidandel, and Landis (2015) discussion of the challenges (e.g., confidentiality, informed consent) that big data projects present to industrial and organizational (I-O) psychologists is timely. If the hype associated with these techniques is warranted, then our field has a clear imperative to debate the ethics and best practices surrounding use of these techniques. We believe that Guzzo et al. have done our field a service by starting this discussion.

We would like to continue their discussion by considering what changes are needed to train future I-O psychologists to be savvy researchers, consumers, and practitioners of big data and data mining results. We start our discussion with an overview of potential modifications in PhD and master's programs. In this discussion, we focus on one technique, neural network analysis, which might prove useful as a gateway for introducing students to big data techniques.

Competencies Needed in the Era of Big Data

Integrating big data research and analytics into I-O education is not as simple as it might initially seem. As Guzzo et al. point out, big data is not merely large datasets. Characteristics of big data—including volume, variety, and velocity (McAfee & Brynjolfsson, 2012)—require a skillset not currently taught in typical graduate-level statistics and research methods classes. Clearly, the probability that our future graduate students have already learned some of these skills is rather low. In fact, the 1999 Society for Industrial and Organizational Psychology's (1994, 1999) guidelines for master's and PhD graduate education do not address a number of skills we consider critical for capitalizing on big data. Because these guidelines indicate that master's students often go into data analysis, it is critical that both master's and PhD students

Juliet R. Aiken and Paul J. Hanges, Department of Psychology, University of Maryland.
Correspondence concerning this article should be addressed to Juliet R. Aiken, Department of Psychology, University of Maryland, College Park, MD 20742. E-mail: jraiken@umd.edu

https://doi.org/10.1017/iop.2015.80 Published online by Cambridge University Press

develop greater breadth and depth in data analytic skills—and especially in how to interpret and leverage the results of big data analysis. Currently, the guidelines specify that students should be skilled in descriptive statistics, regression, correlation, and estimates of mean differences and familiar with path analysis, meta-analysis, and causal modeling. Although these skills are important, we do not believe they are adequate for handling the volume, variety, and velocity of big data.

Of course, before we can discuss how big data skills might be taught, it is important to identify the necessary competencies required to be able to professionally use big data techniques and results. In addition to the skill set outlined in the Society for Industrial and Organizational Psychology graduate education guidelines, some new skills that are absolutely necessary are computer programming skills (at least until these techniques become integrated into more popular statistical software packages) as well as knowledge about the differential assumptions/limitations/advantages of the various big data techniques (e.g., supervised versus unsupervised learning techniques). Indeed, a forthcoming revision to the graduate education guidelines is expected to list programming as a necessary skill for I-O graduate education (Payne, Morgan, & Bryan, 2015). Simple suggestions to include training in these two areas are as follows.

Programming Skills

Big data is exactly that—big—and often these datasets are too large to deal with using traditional statistical software programs. Skills in programming languages such as Python and SQL (Structured Query Language) will be essential for someone who has to create, access, manage, and/or analyze big data (Provost & Fawcett, 2013). Dabblers can partner with researchers or professionals who already have these skills. Experience with these programs will enable future PhD and master's I-O professionals to have a direct influence on data integrity as well as ethical issues surrounding the data.

Although it is unreasonable to fold these specific programming skills into already full statistics courses, perhaps some general programming skills can be incorporated by changing the kinds of statistical software used in these courses. For example, instead of using statistical programs with drop down menus (e.g., SPSS) in our courses, computer programming logic could be introduced by teaching statistics using R.

Of course, I-O graduate programs might enhance the programming skills of their graduates by allowing students to enroll in computer science programming or data science courses as seminars/electives. For I-O students, faculty, and professionals who want to develop big data programming skills but do not have the bandwidth to enroll in a course, online MOOCs (massive open online courses) provide an alternate training ground for these

https://doi.org/10.1017/iop.2015.80 Published online by Cambridge University Press

techniques. The webpage Coursera, for example, hosts a number of data science courses, including two separate course sequences out of Johns Hopkins, which include instruction on programming as well as data science analytic techniques.

Supervised and Unsupervised Learning Data Science Techniques

In addition to programming skills, students of data science will need to learn a number of new analytic techniques. These new techniques can be classified into two broad categories: supervised versus unsupervised. *Supervised* techniques are those analyses focused on predicting criteria variables. For example, regression analysis is a supervised analytic technique in that the criterion variable provides information regarding the choice of optimal weights for the predictors. *Unsupervised* analytic techniques do not include a criterion but still search for optimum weights or a combination of a set of variables. For example, exploratory factor analysis results in a factor structure capturing the intercorrelations among the variables, and it does so without the aid of an external criterion.

In contrast to traditional statistical techniques that use ordinary least squares regression or maximum likelihood rules to mathematically derive an optimal solution, big data techniques typically use iterative machine learning rules that use computational brute force. Traditional statistical tools can be thought of as a subset of big data techniques. However, big data techniques go beyond traditional tools by releasing restrictive statistical assumptions and can solve problems that could not be reasonably approached with traditional tools. This will require that I-O psychologists learn new techniques such as text mining and tree induction, a supervised technique used to classify people/objects into distinct groups.

These techniques also do not have to be folded into existing statistics courses. Data science courses hosted on online MOOCs, such as those mentioned previously, and data science courses in other departments on campus can serve to educate I-O students on these data analytic techniques. Further, students and faculty are urged to supplement online or in-person education by reading accessible data science books, such as *Data Science for Business: What You Need To Know About Data Mining and Data-Analytic Thinking* (Provost & Fawcett, 2013).

Finally, as I-O faculty develop competency in analyzing big data themselves, they can construct I-O-specific big data seminars to offer to PhD and master's students. One analytic technique we suggest faculty prioritize integrating into their classes is artificial neural networks (ANNs). ANNs offer an ideal gateway for I-O psychologists into the world of big data analytics for several reasons. First, options to run ANNs are already available in many of the most popular statistical software packages (e.g., SPSS,

https://doi.org/10.1017/iop.2015.80 Published online by Cambridge University Press

SAS, R) as well as available in graphical user interface format in such programs as MATLAB. Thus, introducing these techniques into current research methods courses does not necessarily require students learning a new program in addition to a new technique. Second, as suggested above, some of the neural networks can be structured to accomplish the goals of traditional statistical techniques (Hanges, Lord, Godfrey, & Raver, 2002)—albeit with fewer statistical assumptions. This overlap in purpose should reduce the initial conceptual confusion about ANNs and how they work. Finally, ANNs already have a foothold in I-O psychology. ANNs have been used to predict job performance (Minbashian, Bright, & Bird, 2010) and employee turnover (Somers, 1999), among other outcomes. Given that ANNs are accessible and already used in I-O research, they are an excellent place for any aspiring I-O data scientist to begin. We will next very briefly introduce ANNs and offer suggestions for interested parties to develop their ANN skills.

Building Blocks of ANNs

The term "artificial neural network" (ANN) actually applies to a family of models rather than one specific model. An ANN is a parallel distributed processor, in that, rather than information being processed step by step, incoming information is sent to and processed by all parts of the network simultaneously. In addition, ANN models iteratively learn and make decisions based on prior learning and stimulus input. Knowledge is stored in synaptic weights that are conceptually equivalent to variable weights in traditional statistical models.

ANNs have three classes of nodes, or variables. Two of these classes resemble the kinds of variables researchers use in a regression: input (predictor) nodes and output (criterion) nodes. The third class, the hidden nodes, loosely corresponds to interaction terms in traditional analyses; however, the full power of what is possible with these kinds of interaction terms has not been tapped in traditional statistics (Hanges et al., 2002). Along with the output variable and the system's learning rule, optimal weights are obtained so that the ANN can accomplish its mission.

Because ANNs are a family of model, the number of levels of the input, output, and hidden nodes and the way learning occurs depend on the architecture of the ANN in question. Multiple different architecture "templates" exist. For example, the multilayer feed-forward network involves one or more layers of hidden nodes. This network is said to be feed forward because information from the input variables flow into and influence the hidden nodes. The hidden nodes in the first layer of this network influence hidden nodes in the second layer, but hidden nodes in the second layer cannot affect those in the first layer. Thus, information and influence flow in one

https://doi.org/10.1017/iop.2015.80 Published online by Cambridge University Press

direction. Finally, in addition to different learning rules, the researcher can set aspects of the ANN architecture. For example, the researcher can set the speed at which the ANN arrives at a solution (i.e., learning rate), the extent to which the ANN explores different values after finding an initial solution (i.e., momentum of learning), and the reduction in learning rate that occurs as the network approaches its goal.

ANNs are generally a powerful alternative to regression, particularly when predicting complex, nonlinear relationships. An ANN is superior in assessing relationships among variables in large sets of data (Collins & Clark, 1993) and in instances where a great deal of data are missing (Collins & Clark, 1993). Whereas experimental research usually yields complete, fairly compact datasets, the kind of research that yields large—dare we say it, big—datasets tends to involve large, noisy datasets with missing information. Consequently, ANNs are an excellent "starter" big data analysis technique.

However, the benefits of ANN are not limited to statistical data analysis. ANNs can be deliberately designed to provide computational models of theory. These models can be simulated to yield a deeper understanding of the previously unrecognized implications of the original theory. Indeed, ANNs were originally developed to model human neural processes (Friedenberg, 2009). They are uniquely suited to modeling organizational learning and decision-making processes. For example, Lord, Hanges, and Godfrey (2003) tested, updated, and modeled valence–instrumentality–expectancy decision-making theory using ANNs. A full discussion of how to conduct ANN analyses is beyond the scope of the current response. Readers interested in ANNs can start with Collins and Clark (1993) and Hanges et al. (2002); readers interested in big data can start with Provost and Fawcett (2013) and Stanton (2014).

In summary, an ANN provides a starting point for introducing I-O master's and PhD students to big data. Using an ANN enables an easy introduction of programming skills and knowledge about big data analytic techniques into our graduate programs.

Conclusion

In this comment, we have discussed competencies that need to be built into graduate education for I-O as a field to stay on top of the big data trend. In addition to providing actionable recommendations for graduate education, we have provided a snapshot of a tool—ANNs—that will help I-Os enter the world of big data. Finally, as we wrap up this comment, we need to emphasize one final issue. Yes, big data is coming. Yes, big data techniques are important for I-O researchers and practitioners to understand. However, not every I-O professional is going to be (or should be!) a data scientist. Just as I-O students

https://doi.org/10.1017/iop.2015.80 Published online by Cambridge University Press

and faculty currently focus themselves on a handful of areas of interest, big data analysis will likewise be an interest and passion for only a subset of I-O psychologists. However, big data is going to be a reality in tomorrow's organizations, and although not all I-O psychologists need to analyze these data, all I-O psychologists should absolutely know how to interpret and leverage findings from big data analysis in the workplace. This is not just something that would be nice to see; this is an imperative, and our graduate training needs to reflect this imperative immediately.

References

Collins, J. M., & Clark, M. R. (1993). An application of the theory of neural computation to the prediction of workplace behavior: An illustration and assessment of network analysis. *Personnel Psychology, 46*, 503–524.

Friedenberg, J. (2009). *Dynamical psychology: Complexity, self-organizational, and mind*. Litchfield Park, AZ: ICSE.

Guzzo, R. A., Fink, A. A., King, E., Tonidandel, S., & Landis, R. S. (2015). Big data recommendations for industrial–organizational psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 8*(4), 491–508.

Hanges, P. J., Lord, R. G., Godfrey, E. G., & Raver, J. L. (2002). Modeling nonlinear relationships: Neural networks and catastrophe analysis. In S. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology* (pp. 431–455). Malden, MA: Blackwell.

Lord, R. G., Hanges, P. J., & Godfrey, E. G. (2003). Integrating neural networks into decision making and motivational theory: Rethinking VIE theory. *Canadian Psychologist, 44*, 21–38.

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review*, *60*, 60–66.

Minbashian, A., Bright, J. E. H., & Bird, K. D. (2010). A comparison of artificial neural networks and multiple regression in the context of research on personality and work performance. *Organizational Research Methods, 13*, 540–561.

Payne, S. C., Morgan, W. B., & Bryan, L. K. (2015). *Revision of SIOP's guidelines for education and training at the doctoral and master's level in I-O psychology*. Executive Board Special Session at the 30th Annual Conference of the Society for Industrial and Organizational Psychology, Philadelphia, PA.

Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. Sebastopol, CA: O'Reilly Media.

Society for Industrial and Organizational Psychology. (1994). *Guidelines for education and training at the master's level in industrial–organizational psychology*. Arlington Heights, IL: Author.

Society for Industrial and Organizational Psychology. (1999). *Guidelines for education and training at the doctoral level in industrial/organizational psychology*. Bowling Green, OH: Author.

Somers, M. J. (1999). Application of two neural network paradigms to the study of volunteer employee turnover. *Journal of Applied Psychology, 84*, 177–185.

Stanton, J. M. (2014). Data mining: A practical recommendations for organizational researchers. In J. M. Cortina & R. S. Landis (Eds.), *Modern research methods for the study of behavior in organizations* (pp. 199–230). New York, NY: Routledge.

https://doi.org/10.1017/iop.2015.80 Published online by Cambridge University Press