Applying big data methods to understanding human behavior and health

Moustafa, Ahmed A.; Diallo, Thierno M.O.; Amoroso, Nicola; Zaki, Nazar; Hassan, Mubashir; Alashwal, Hany

Published in:
Frontiers in Computational Neuroscience

DOI:
10.3389/fncom.2018.00084

frontiers in Computational Neuroscience

OPINION
published: 16 October 2018
doi: 10.3389/fncom.2018.00084

Applying Big Data Methods to Understanding Human Behavior and Health

Ahmed A. Moustafa[1,2], *Thierno M. O. Diallo*[1], *Nicola Amoroso*[3,4], *Nazar Zaki*[5], *Mubashir Hassan*[6] *and Hany Alashwal*[5]*

[1] School of Social Sciences and Psychology, MARCS Institute for Brain and Behaviour, Western Sydney University, Sydney, NSW, Australia, [2] Department of Social Sciences, College of Arts and Sciences, Qatar University, Doha, Qatar, [3] Dipartimento Interateneo di Fisica "M. Merlin," Università degli Studi di Bari "A. Moro," Bari, Italy, [4] Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Bari, Italy, [5] College of Information Technology, United Arab Emirates University, Al-Ain, United Arab Emirates, [6] Department of Biological Sciences, College of Natural Sciences, Kongju National University, Gongju, South Korea

Keywords: machine learning, deep learning, psychology, health, human behavior, big data, causality, longitudinal data

INTRODUCTION

While many fields have benefited greatly from the collection and analysis of big data, some health fields and, to a large extent, psychology are still lagging behind (Azmak et al., 2015). Azmak et al. (2015) have shown an example (e.g., Sloan Digital Sky) on how the collection of large datasets has aided researchers to solve difficult problems in astronomy that were not possible in the past. Interestingly, the slow process of applying big data to psychology mirrors the history of development of sciences, as astronomy and other sciences are much older than experimental psychology (which emerged in the nineteenth century). This is related to the fact that while many sciences are data-driven, psychology, to a large degree, is hypothesis-driven (see below discussion on these points).

OPEN ACCESS

Edited by:
Carlo Laing,
Massey University, New Zealand

Reviewed by:
Xiaofeng Zhu,
Massey University Business School, New Zealand

***Correspondence:**
Hany Alashwal
halashwal@uaeu.ac.ae

Received: *26 July 2018*
Accepted: *18 September 2018*
Published: *16 October 2018*

Citation:
Moustafa AA, Diallo TMO, Amoroso N, Zaki N, Hassan M and Alashwal H (2018) Applying Big Data Methods to Understanding Human Behavior and Health. Front. Comput. Neurosci. 12:84. doi: 10.3389/fncom.2018.00084

WHY BIG DATA METHODS HAVE BEEN RARELY APPLIED TO PSYCHOLOGY?

There are several reasons why psychology researchers rarely collect large datasets, and if we do, may not use big data methods for analyses. As pointed out by Cheung and Jak (2016), big data analysis is not considered a core topic in behavioral sciences. Another factor is most psychology research is theory- rather than data-drive (Qiu et al., 2018). Accordingly, most psychology researchers often collect data using a small number of variables to test their theory. Psychology students are often encouraged to have a hypothesis underlying their new experiments. However, there are often many theories that explain a certain behavioral phenomenon, and a theory-driven approach can rarely find best theories. Accordingly, here we argue that it is good to let the data speak for themselves, that is, to take a data-driven approach. However, this requires the collection of large datasets and conducting big data analyses.

Historically and up till recently, most psychological studies collect data using small number of variables (usually under 10) (for discussion see Cheung and Jak, 2016). There are, however, some exceptions including the World Values Survey, Math Garden, Kavli Human Project (Azmak et al., 2015), as well as few recent studies (Kern et al., 2014; Youyou et al., 2015). Most of these studies often analyze big data collected from social media websites, such as Facebook and Twitter. However, even with big datasets, most psychology researchers still divide the data into smaller parts for more standard statistical analyses. This is in contrast to neural (e.g., neuroimaging, EEG, and single-cell recording) that often include 100s of variables.

Further, many of the variables in psychological studies are categorical, such as male/female, lives in Urban vs. rural area, patient or control, Young vs. older, and so on. It is possible that the nature of such data have discouraged researchers from conducting complex analytical tools, as most existing deep learning and big data methods often deal with continuous variables. However, some recent efforts have shown that deep learning methods can also be applied to categorical variables. For example, Zhang et al. (2016) have used several DNNs, including Factorization Machine, supported Neural Network (FNN), and Restricted Boltzmann Machine to understand online advertisement, specifically to predict user responses in a website. While this domain is different from medical and behavioral fields, the data they have used also include categorical variables. Most of these DNNs represent categorical variables as a set of binary values. We argue that these algorithms can be applied to solve complex psychology and health problems.

WHY DO WE NEED BIG DATA ANALYSIS IN PSYCHOLOGY?

It is important to note human behavior and health issues are quite complex. For example, Alzheimer's disease, which is the most common neurodegenerative disease in old age (Ballard et al., 2011; Geldmacher, 2012), is associated with several genetic, nutritional, cognitive, and neural changes that amount to 100s of variables. Standard statistical methods as used in most empirical studies are ill-equipped to diagnose and understand AD. Big data methods will allow us to select the most important features that differentiate AD patients from healthy individuals, as this will allow clinicians and neurologists to only test these variables in clinical practice.

As human behavior is extremely complex, it is no surprise that many existing findings in the field of psychology and medicine are conflicting. This is perhaps due to the existence of several factors affecting human behavior as well as the simplicity of theories used to explain human behavior (which is due to theory-driven approach in the field, as we discussed above). However, most psychological experiments mostly focus on measuring 2–7 variables. Most standard statistical methods cannot handle datasets with a large number of variables. Further, many of the small datasets cannot answer questions about causality. To do so, researchers often needs to collect longitudinal and big datasets. Below, we describe how machine learning methods, such as clustering and deep learning, can be applied to big datasets to solve complex psychology problems (see **Figure 1**).

THE APPLICATION OF CLUSTERING METHODS IN PSYCHOLOGY

Clustering is the process of partitioning a set of individuals (or objects) into subgroups. Accordingly, a cluster is a collection of data points that are similar to one another and dissimilar to data points in other clusters (Escudero et al., 2011). Clustering methods seek to segment the entire dataset into relatively homogeneous subgroups or clusters, where the similarity of data points within the cluster is maximized, and the similarity to data points outside this cluster is minimized (Larose, 2005). The clustering problem is defined as follows: Given a set of points in the multidimensional space, find a partition of the points into clusters so that the points within each cluster are similar to one another. Traditional clustering describes clusters using measures of similarity, such as Euclidian distance, and considers data points belonging to one and only one cluster at the time.

However, one type of clustering method known as fuzzy clustering, allows data points to belong to several clusters at the same time but with different membership degrees (Ahmadi et al., 2018). Fuzzy clustering has many applications to health sciences, as some individuals may or may not be diagnosed with a certain disorder, depending on different conditions.

Many psychological studies often start with dividing participants into some clusters. For example, many psychology studies divide participants into patient or control, urban or rural, and so on. However, clustering methods allow us to cluster data based on similarities among members/elements. In doing so, clustering methods can divide data into several clusters, and not necessarily 2 only. For example, instead of human tendency to divide participants into patient or control, clustering algorithms can subgroup these participants into 2, 3 or more clusters, perhaps pointing to several subgroups of patients. It is also possible that the rural/urban data involve several clusters, as urban people may be subdivided into several subclusters. By using clustering methods, we may be able to find more important relationships among participants than often assumed a priori. In one recent study, Crouse et al. (2018) used a clustering method, known as Hierarchical Agglomerate, to subtype psychosis-prone individuals. Unlike standard clustering algorithms, this approach assumes each element has its own cluster and then clusters are merged based on similarity in a hierarchical manner. Results show that there are three clusters, which differ in IQ and social functioning. Future research should use similar methods to subtype participants instead of using a priori (assumed) taxonomy.

THE APPLICATION OF DEEP LEARNING METHODS IN PSYCHOLOGY

Deep neural networks (DNNs) are commonly used to classify data in different fields (LeCun et al., 2015; Amoroso et al., 2018; Wang et al., 2018). DNNs are non-linear methods that allow the learning of complex patterns among features, thus providing a complex non-linear classification of input data (Graepel et al., 2010). With more hidden layers in the network, the data becomes more easily separable due to non-linear transformations along different layers of the network (Plis et al., 2014). Thus, DNNs are able to utilize different feature combinations and thus could potentially improve classification of complex datasets. This has several benefits over linear classification models that often ignore complex interactions among features. In recent years, DNNs have gained great importance, especially because they better manage raw data than classical machine learning algorithms, thus they do not require a strong effort by human experts to

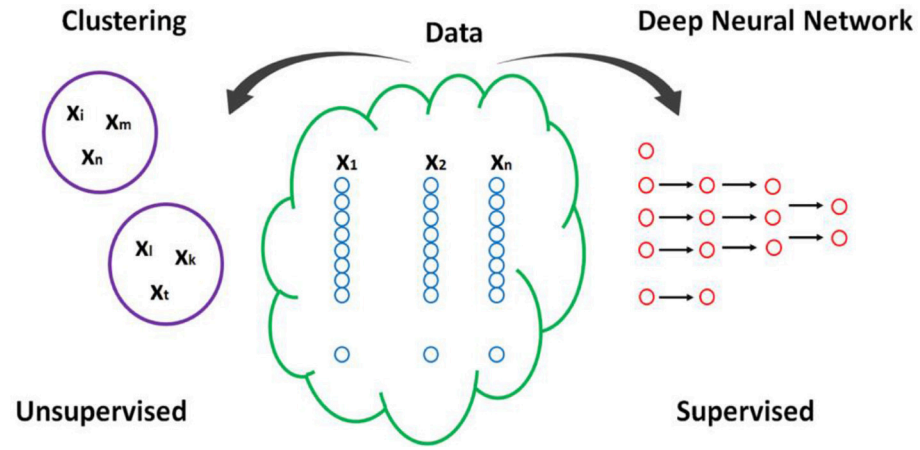

FIGURE 1 | A schematic explaining the clustering and DNN proposed here. X_j refers to feature j, and Y_i refers to individual i. n here is numbers of features, and m to a number of individuals.

denote which variables should be considered and which not on order to detect significant patterns within the data. Besides, the availability of huge computational resources (thanks to cloud technology) allows an intensive use of deep learning algorithms. Importantly, DNNs can be used to help predict who may develop a certain disorder, which is very important for providing an effective treatment for the patients (Choi et al., 2018). DNNs can also be used to predict academic performance of students based on their input data. Importantly deep learning can be used to extract key features underlying category membership (known as feature selection).

Feature Extraction

Different machine learning methods, such as the random forest algorithm, allow researchers to find best features/variables to explain differences among two or groups of participants (Amoroso et al., 2018). There are several ways to conduct feature selection. For example, one study used weight pruning methods in the Input Layer to find relevant features (Roy et al., 2015). Similarly, Munsell et al. (2015) have used the elastic net algorithm to reduce features and network connections. Nezhad et al. (2017) also identified the most relevant features underlying the occurrence of hypertension using an auto-encoder network. Recently, Zhang et al. (2016) have used Discriminant Autoencoder Network with Sparsity Constraint (DANS) to extract most important features that discriminate schizophrenia patients from healthy individuals. They reported some larger weight value in the network for certain features (connectivity of some brain areas, including cortex, basal ganglia, and cerebellum) best differentiate between the two populations.

In contrast, in standard psychology studies, usually experimental scientists test differences usually between one or two variables. As an example, a standard psychology study may investigate differences in quality of life, depression, stress, and so on in urban vs., rural participants. The study will then investigate if each of these variables or perhaps an interaction among two (or more) of them is significantly different. Multivariate classification methods allow the researcher to unveil strategic roles played by a set of variables, weak if considered on their own and which therefore could be disregarded. However, deep learning methods can test the differences among all variables, which can be in the order of 100s. In one recent study, Guo et al. (2015) used deep learning methods to find factors underlying academic performance of students. The networks included background, school-related, past study performance, and personal data, among other variables. The network was able to find a subset of these variables that predict academic performance (network output). Accordingly, psychology researchers can benefit from these findings by focusing on improving scores on variables related to better academic performance. Similarly, selecting key features has clinical importance, as it helps provide neurologists and clinicians with most important features that classify the sample (e.g., patient vs. healthy individual). Based on feature selection algorithms, neurologists can then only focus on collecting and measuring data related to these features in future diagnostic work.

CONCLUSIONS

We here argue that the more data we collect, the better our understanding of human behavior will be. Instead of relying on theory-driven methods as often the case in psychology studies, big data approaches can drive discovery and let new "theories" arise directly from data. In addition, big data methods can provide unexpected results on subtypes of participants as well as better understand the nature of human behavior.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Ahmadi, H., Gholamzadeh, M., Shahmoradi, L., Nilashi, M., and Rashvand, P. (2018). Diseases diagnosis using fuzzy logic methods: a systematic and meta-analysis review. *Comput. Methods Prog. Biomed.* 161, 145–172. doi: 10.1016/j.cmpb.2018.04.013

Amoroso, N., Diacono, D., Fanizzi, A., La Rocca, M., Monaco, A., Lombardi, A., et al. (2018). Deep learning reveals Alzheimer's disease onset in MCI subjects: results from an international challenge. *J. Neurosci. Methods* 302, 3–9. doi: 10.1016/j.jneumeth.2017.12.011

Azmak, O., Bayer, H., Caplin, A., Chun, M., Glimcher, P., Koonin, S., et al. (2015). Using big data to understand the human condition: the Kavli HUMAN project. *Big Data* 3, 173–188. doi: 10.1089/big.2015.0012

Ballard, C., Gauthier, S., Corbett, A., Brayne, C., Aarsland, D., and Jones, E. (2011). Alzheimer's disease. *Lancet* 377, 1019–1031. doi: 10.1016/S0140-6736(10)61349-9

Cheung, M. W., and Jak, S. (2016). Analyzing big data in psychology: a split/analyze/meta-analyze approach. *Front. Psychol.* 7:738. doi: 10.3389/fpsyg.2016.00738

Choi, H., Jin, K. H., and Alzheimer's Disease Neuroimaging Initiative. (2018). Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging. *Behav. Brain Res.* 344, 103–109. doi: 10.1016/j.bbr.2018.02.017

Crouse, J. J., Moustafa, A. A., Bogaty, S. E. R., Hickie, I. B., and Hermens, D. F. (2018). Parcellating cognitive heterogeneity in early psychosis-spectrum illnesses: a cluster analysis. *Schizophr. Res.* doi: 10.1016/j.schres.2018.06.060. [Epub ahead of print].

Escudero, J., Zajicek, J. P., and Ifeachor, E. (2011). Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2011:6470–6473. doi: 10.1109/IEMBS.2011.6091597

Geldmacher, D. S. (2012). "Alzheimer's disease," in *Clinical Manual of Alzheimer Disease and Other Dementias*, eds M. F. Weiner and A. M. Lipton (Arlington, TX: American Psychiatric Publishing), 127–158.

Graepel, T., Candela, J. Q., Borchert, T., and Herbrich, R. (2010). "Web-scale bayesian clickthrough rate prediction for sponsored search advertising in microsoft's bing search engine," in *Paper presented at the Proceedings of the 27th International Conference on Machine Learning (ICML-10)* (Haifa).

Guo, B., Zhang, R., Xu, G., Shi, C., and Yang, L. (2015). "Predicting students performance in educational data mining," in *International Symposium on Educational Technology* (Perth, WA).

Kern, M. L., Eichstaedt, J. C., Schwartz, H. A., Park, G., Ungar, L. H., and Stillwell, D. J. (2014). From "sooo excited!!!" to "so proud": using language to study development. *Dev. Psychol.* 50, 178–188. doi: 10.1037/a0035048

Larose, D. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, NJ: John Wiley & Sons.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Munsell, B. C., Wee, C. Y., Keller, S. S., Weber, B., Elger, C., da Silva, L. A., et al. (2015). Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. *Neuroimage* 118, 219–230. doi: 10.1016/j.neuroimage.2015.06.008

Nezhad, M. Z., Zhu, D., Li, X., Yang, K., and Levy, P. (2017). SAFS: a deep feature selection approach for precision medicine. *arXiv.org*. doi: 10.1109/BIBM.2016.7822569

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., et al. (2014). Deep learning for neuroimaging: a validation study. *Front. Neurosci.* 8:229. doi: 10.3389/fnins.2014.00229

Qiu, L., Hian, S., Chan, M., and Chan, D. (2018). Big data in social and psychological science: theoretical and methodological issues. *J. Comput. Soc. Sci.* 1, 59–66. doi: 10.1007/s42001-017-0013-6

Roy, D., Murty, K. S. R., and Mohan, C. K. (2015). "Feature selection using Deep Neural Networks," *in Paper presented at the 2015 International Joint Conference on Neural Networks* (IJCNN), Killarney.

Wang, S. H., Phillips, P., Sui, Y., Liu, B., Yang, M., and Cheng, H. (2018). Classification of Alzheimer's disease based on eight-layer convolutional neural network with leaky rectified linear unit and max pooling. *J. Med. Syst.* 42:85. doi: 10.1007/s10916-018-0932-7

Youyou, W., Kosinski, M., and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proc. Natl. Acad. Sci. U.S.A.* 112, 1036–1040. doi: 10.1073/pnas.1418680112

Zhang, W., Du, T., and Wang, J. (2016). "Deep learning over multi-field categorical data- a case study on user response prediction," in *Advances in Information Retrieval, ECIR 2016, Lecture Notes in Computer Science, vol 9626*, ed N. Ferro (Cham: Springer), 45–57. doi: 10.1007/978-3-319-30671-1

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer XZ and handling Editor declared their shared affiliation at the time of review.