INVITED EDITORIAL COMMENT: Order from Chaos: *How Data Science Is Revolutionizing Investment Practice*

Joseph Simonian, Marcos López de Prado and Frank J. Fabozzi

JPM 2018, 45 (1) 1-4
doi: https://doi.org/10.3905/jpm.2018.45.1.001
http://jpm.iijournals.com/content/45/1/1

Institutional Investor Journals
1120 Avenue of the Americas, 6th floor,
New York, NY 10036, Phone: +1 212-224-3589

The Journal of Portfolio Management

INVITED EDITORIAL COMMENT

Order from Chaos: *How Data Science Is Revolutionizing Investment Practice*

JOSEPH SIMONIAN, MARCOS LÓPEZ DE PRADO, AND FRANK J. FABOZZI

Data are fundamental inputs to any applied scientific endeavor. Finance is no different. Yet for almost its entire existence as an organized field of inquiry, much of finance has relied almost exclusively on relatively primitive and rigid forms of data analysis to drive both investment theories and real-world portfolio management decisions. Over the years, it has become evident that the complexity of today's markets, with a seemingly infinite amount of data being generated at rapid speeds, cannot be handled using the blunt mathematical arrows being deployed from the analytical quivers of most investment professionals. In response, the investment industry has begun to recognize the value and importance of the various methodological approaches that constitute what has come to be known as *data science*. Although the basic ideas that undergird much of data science can be traced at least back to Alan Turing (1947, 1948, 1951),[1] it is only in recent decades, with rapid advances in computational power, that it has become possible to actualize many of the ideas of Turing and other computing pioneers.

We believe that there is currently no less than a methodological revolution under way in finance, away from a world exclusively dominated by theory-based models, closed-form solutions, and well-behaved distributions toward a new world in which heuristically driven frameworks, algorithmic methods, and nonparametric statistics play leading analytical roles. We believe that the transition to this new world constitutes a *paradigm shift*, as described by Thomas Kuhn (1970), and stands to change finance as much as the Copernican Revolution changed astronomy or the Darwinian Revolution changed biology. This editorial introduces data science to the wider investment community and highlights some of the advantages it can bring to everyday investment practice.

What is data science? Although there is no universal definition of data science, we can say that it is a field of study that combines the use of statistics and computing to discover or impose order in complex data to enhance informed decision-making. It is thus an inherently practical endeavor, just like finance, and so it is especially suited to investment applications. One branch of data science, machine learning, comprises a family of computational techniques that facilitate the automated learning of patterns and the formation of predictions from data. Although there are many types of machine learning frameworks and algorithms, they share the following three elements: (1) a method for extracting and representing the essential features of the data under consideration; (2) a process and period of model training; and (3) an objective function derived ("learned") during the training period and applied to post-training data. Furthermore, machine learning algorithms are generally designed to solve one of two types of problems: a *classification*-type problem, in which the goal is to categorize data into different types, or a *regression*-type

[1] Turing expounded his proto-machine learning ideas not to describe a new area of applied computing, but largely to advance his multi-machine theory of the mind (Copeland and Shagrir 2013).

problem, in which the goal is to predict a quantity for a variable given the values for a set of predictor variables. Both types of problems are ubiquitous in finance, so machine learning can be viewed as a natural extension to investment practitioners' existing tool set.

Data science algorithms gain much of their problem-solving power from the unique ways in which they process information. For example, decision tree–based forms of data analysis provide a hierarchical approach to analyzing data, whereas neural network algorithms use an inherent parallel processing of data to arrive at solutions. Many data science algorithms use randomization, which allows them to reduce the computation time to solve problems involving large data sets (*big data*). Genetic algorithms, for example, which are inspired by the evolutionary theory of natural selection, use *mutation* operators, which produce random alternations of variable values to promote solution diversity.

No matter what their specific form, machine learning algorithms are capable of processing and analyzing data, both structured and unstructured, in ways that traditional financial models are not. Structured data, perhaps best exemplified by an econometric time series, are defined as data that have readily observable patterns and regularities—more simply put, they are easily expressible in rows and columns. A time series is clearly indexed chronologically, with each observation expressed in the same unit, scale, and denomination as every other observation in the series. Although time series analysis has borne much fruit for investment analysis over the years, the near-exclusive focus on analyzing structured data has prevented investment practice from harvesting the rich informational resources in unstructured data. What are unstructured data? As the name implies, unstructured data are data that are not discernibly organized in any defined way (e.g., social media postings, high-frequency macroeconomic data, or credit card transactions), often combining numerical with categorical variables. For a long time, unstructured data were either ignored or were processed using computationally primitive and time-consuming means. In recent years, however, the rapid development and dissemination of data science methodologies has given new life to investment research by giving practitioners the formal tools to harness an expanded universe of data types and thereby enhance their potential for discovering profitable information sources.

To see how data science can help advance investing practice, consider the following example: An equity research analyst is interested in forecasting the prospects of a particular stock, Hungry Inc., a U.S.-based chain of restaurants. The 12-month price momentum of this particular security is strong, and the analyst would like to combine the momentum signal with an additional piece of support to make a final decision on whether or not to recommend the purchase of the company's shares. One potentially useful piece of information is the number of patrons who have been frequenting Hungry Inc. restaurants nationwide over the last 12 months. Has the number of patrons been increasing or decreasing coincident with its strong price momentum over the same period? How would one go about finding out this information? Well, one avenue to this information is satellite imagery, which, with some treatment, can be used to count the total number of cars in the parking lots of Hungry Inc. restaurants across the country. If the number of cars has been increasing over the last 12 months, that would seem to justify the strong price momentum observed in the market. Conversely, if the number of cars has been decreasing significantly over the same period, that could be an indication that the stock's momentum is not supported by the most fundamental of factors, paying customers.

Of course, satellite imagery does not have an inherent structure, so we need a way of imbuing it with one. One versatile approach that can help us with our problem is provided by neural network algorithms, mentioned earlier. As the name implies, neural networks are motivated by the functioning of the human brain.[2] Their basic design consists of a collection of data processors organized in layers, called *neurons* (or *nodes*). Information is processed via the responses of neurons to external inputs. These responses are then passed on to the next layer, eventually ending up as final output. The interconnectedness of neurons and their ability to pass information back and forth to each other facilitates the efficient solution of problems. Neural networks learn

[2] There are notable differences between neural networks and biological brains. Among other things, human brains do not use a procedure akin to backpropagation, which is an essential feature of many neural network algorithms.

via a set of training data, and through different types of error correction mechanisms they gradually develop the ability to produce correct answers to new data outside the training set. Generally speaking, neural networks work by initially positing random solutions to problems, solutions that are expectedly inaccurate to some degree. Once an initial solution is produced, the information relating to the prediction or classification error of the initial solution is fed back into the network so that adjustments can be made in the next iteration. As the process is repeated, the network begins to discover more precise adjustments, eventually developing an "understanding" of how to solve the problem.

In our example, our research analyst needs to train the neural network to distinguish between cars and non-cars, so here we have a classification problem. Training in this case would consist of the neural network being fed various images of cars and non-cars, images that have been translated (encoded) into numerical form, so that over time the network would be able to demarcate discrete bundles of pixels on a digital map as individual cars and proceed to count them at each Hungry Inc. location. The features of the cars that would play a role in the training of the network would be things such as the possible spectrum of colors that cars could be, the minimum and maximum dimensions of cars, and the like. The relevant non-cars in this case would be things such as people in the restaurant parking lots, the asphalt in the parking lot, and the roofs of the individual restaurants, among other objects. Once the training is complete, our research analyst can proceed to build a visual signal. We assume that the research analyst has obtained pixelated maps of the United States, on which the location of each Hungry Inc. restaurant has been clearly and accurately identified. Post-training, all that remains is to feed the neural network encoded images of the restaurant parking lots through time and then wait for the running count of the cars over the preceding 12 months.

The foregoing example shows how data science methods can complement existing investment tools. That is not to say that data science in general and machine learning specifically merely exist side by side with traditional methods. In many practical cases, they are better equipped to address the idiosyncrasies of financial datasets. Consider the case of ordinary least squares (OLS) regression, a staple of modern financial research. OLS regression, developed more than two centuries ago by Carl Friedrich Gauss, possesses a number of shortcomings that make it a less than ideal framework for analyzing financial data. Perhaps most significant is the fact that it is a linear continuous model attempting to decipher markets in which nonlinear threshold relationships abound. Furthermore, OLS regression assumes that predictors are relatively independent of one another. In the actual world, many variables exhibit important dependence relationships. Understanding these dependencies is critical if we are to have a clear view of what is driving global markets. The list of shortcomings does not end here—the normality assumption for the estimator and error terms, the undue influence of outlier observations, the expression of the beta coefficient as a mean sensitivity to the predictors, the inability to combine numerical with categorical variables, and the assumption that interaction effects are nonhierarchical all weaken OLS regression as an accurate descriptor of market reality. Does this mean that OLS regression and other traditional tools should never be used again? Of course not. However, in light of the power that data science brings to the analytical table, traditional methods such as OLS regression most likely will play an increasingly secondary role in investment research, either as complements to data science methodologies or, given their simplicity and familiarity, as useful tools for preliminary studies, to test the waters for new ideas.

Finance has spent much of its history trying to mimic the natural sciences, with mechanistic models and simple causal explanations of market behavior. However, unlike research in the natural sciences, finance does not have the benefit of closed experiments and repeated trials; furthermore, it is burdened by a much more complex phenomenon: human intentionality. Nevertheless, much of the profession has, over time, pursued a path in which the production of ever more elegant mathematics has been prioritized over the development of tools and methodologies that, although perhaps messier, could provide more nuanced empirical insight into capital markets. There appears to be a new dawn on the horizon, however, and many have begun to appreciate the limits to much of existing investment theory in relation to its potential to positively contribute to actual investment problems. Consequently, practitioners

are turning to data science and machine learning to drive their analytics and decision-making, trading in the beauty of high theory for the advantages of practical significance. As the sophistication and power of computing continue to grow, data science will surely continue its march to the forefront of investment research and practice. The journey has just begun.

REFERENCES

Copeland, J., and O. Shagrir. 2013. "Turing versus Gödel on Computability and the Mind." In *Computability: Turing, Gödel, Church, and Beyond*, edited by B. J. Copeland, C. J. Posy, and O. Shagrir, pp. 1–33. Cambridge, MA: MIT Press.

Kuhn, T. 1970. *The Structure of Scientific Revolutions*, 2nd ed., enlarged. Chicago: University of Chicago Press.

Turing, A. 1947. "Lecture to the London Mathematical Society." In *The Essential Turing*, edited by B. J. Copeland, pp. 378–394. Oxford: Oxford University Press, 2004.

——. 1948. "Intelligent Machinery." In *The Essential Turing*, edited by B. J. Copeland, pp. 410–432. Oxford: Oxford University Press, 2004.

——. 1951. "Intelligent Machinery, a Heretical Theory." In *The Essential Turing*, edited by B. J. Copeland, pp. 472–475. Oxford: Oxford University Press, 2004.

Disclaimer
This editorial reflects the current opinions of the authors and not those of their employers.

Joseph Simonian *is the director of quantitative research at Natixis Investment Managers in Boston, MA.*
joseph.simonian@natixis.com

Marcos López de Prado *is a principal and the head of machine learning at AQR Capital Management LLC in Greenwich, CT, and a lecturer at Cornell University in Ithaca, NY.*
marcos.lopezdeprado@aqr.com

Frank J. Fabozzi *is a professor of finance at EDHEC Business School in Nice, France.*
frank.fabozzi@edhec.edu