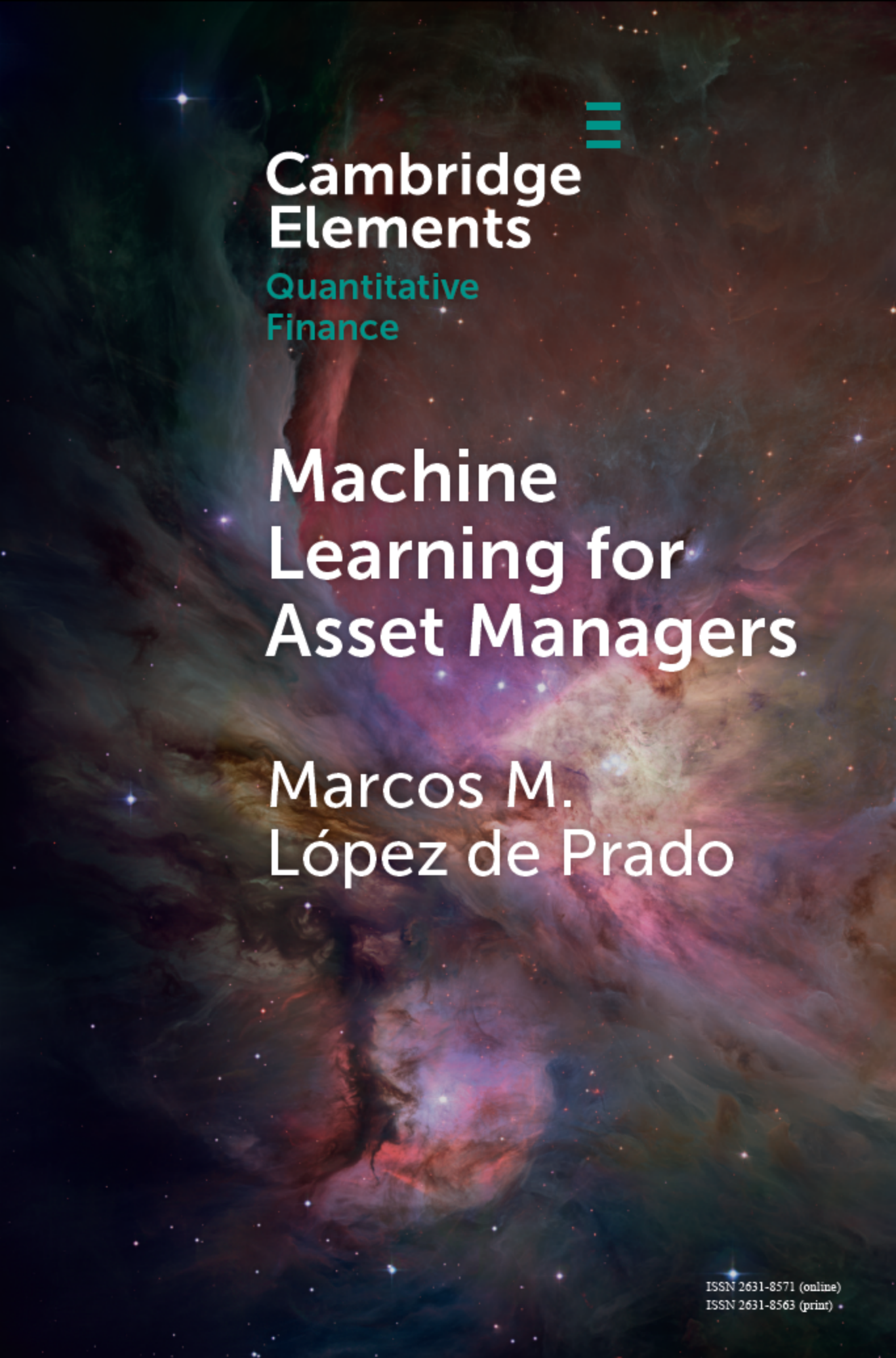

ISSN 2631-8571 (online)
ISSN 2631-8563 (print)

Cambridge Elements

Elements in Quantitative Finance
edited by
Riccardo Rebonato
EDHEC Business School

MACHINE LEARNING FOR ASSET MANAGERS

Marcos M. López de Prado
Cornell University

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108792899
DOI: 10.1017/9781108883658

First published 2020

A catalogue record for this publication is available from the British Library.

ISBN 978-1-108-79289-9 Paperback
ISSN 2631-8571 (online)
ISSN 2631-8563 (print)

Machine Learning for Asset Managers

Elements In Quantitative Finance

DOI: 10.1017/9781108883658
First published online: April 2020

Marcos M. López de Prado
Cornell University

Author for correspondence: ml863@cornell.edu

Abstract: Successful investment strategies are specific implementations of general theories. An investment strategy that lacks a theoretical justification is likely to be false. Hence, an asset manager should concentrate her efforts on developing a theory rather than on backtesting potential trading rules. The purpose of this Element is to introduce machine learning (ML) tools that can help asset managers discover economic and financial theories. ML is not a black box, and it does not necessarily overfit. ML tools complement rather than replace the classical statistical methods. Some of ML's strengths include: (1) a focus on out-of-sample predictability instead of in-sample variance adjudication; (2) the use of computational methods to avoid relying on (potentially unrealistic) assumptions; (3) the ability to "learn" complex specifications, including nonlinear, hierarchical, and noncontinuous interaction effects in a high-dimensional space; and (4) the ability to disentangle the variable search from the specification search, in a manner that is robust to multicollinearity and other substitution effects.

Keywords: machine learning, unsupervised learning, supervised learning, clustering, classification, labeling, portfolio construction

JEL classifications: G0, G1, G2, G15, G24, E44
AMS classifications: 91G10, 91G60, 91G70, 62C, 60E

ISBNs: 9781108792899 (PB), 9781108883658 (OC)
ISSNs: 2631-8571 (online), 2631-8563 (print)

Contents

1 Introduction

1.1 Motivation

To a greater extent than other mathematical disciplines, statistics is a product of its time. If Francis Galton, Karl Pearson, Ronald Fisher, and Jerzy Neyman had had access to computers, they may have created an entirely different field. Classical statistics relies on simplistic assumptions (linearity, independence), in-sample analysis, analytical solutions, and asymptotic properties partly because its founders had access to limited computing power. Today, many of these legacy methods continue to be taught at university courses and in professional certification programs, even though computational methods, such as cross-validation, ensemble estimators, regularization, bootstrapping, and Monte Carlo, deliver demonstrably better solutions. In the words of Efron and Hastie (2016, 53),

> two words explain the classic preference for parametric models: mathematical tractability. In a world of sliderules and slow mechanical arithmetic, mathematical formulation, by necessity, becomes the computational tool of choice. Our new computation-rich environment has unplugged the mathematical bottleneck, giving us a more realistic, flexible, and far-reaching body of statistical techniques.

Financial problems pose a particular challenge to those legacy methods, because economic systems exhibit a degree of complexity that is beyond the grasp of classical statistical tools (López de Prado 2019b). As a consequence, machine learning (ML) plays an increasingly important role in finance. Only a few years ago, it was rare to find ML applications outside short-term price prediction, trade execution, and setting of credit ratings. Today, it is hard to find a use case where ML is not being deployed in some form. This trend is unlikely to change, as larger data sets, greater computing power, and more efficient algorithms all conspire to unleash a golden age of financial ML. The ML revolution creates opportunities for dynamic firms and challenges for antiquated asset managers. Firms that resist this revolution will likely share Kodak's fate. One motivation of this Element is to demonstrate how modern statistical tools help address many of the deficiencies of classical techniques in the context of asset management.

Most ML algorithms were originally devised for cross-sectional data sets. This limits their direct applicability to financial problems, where modeling the time series properties of data sets is essential. My previous book, *Advances in Financial Machine Learning* (AFML; López de Prado 2018a), addressed the challenge of modeling the time series properties of financial data sets with ML algorithms, from the perspective of an academic who also happens to be a practitioner.

Machine Learning for Asset Managers is concerned with answering a different challenge: how can we use ML to build better financial theories? This is not a philosophical or rhetorical question. Whatever edge you aspire to gain in finance, it can only be justified in terms of someone else making a systematic mistake from which you benefit.[1] Without a testable theory that explains your edge, the odds are that you do not have an edge at all. A historical simulation of an investment strategy's performance (backtest) is not a theory; it is a (likely unrealistic) simulation of a past that never happened (you did not deploy that strategy years ago; that is why you are backtesting it!). Only a theory can pin down the clear cause–effect mechanism that allows you to extract profits against the collective wisdom of the crowds – a testable theory that explains factual evidence as well as counterfactual cases (x implies y, and the absence of y implies the absence of x). Asset managers should focus their efforts on researching theories, not backtesting trading rules. ML is a powerful tool for building financial theories, and the main goal of this Element is to introduce you to essential techniques that you will need in your endeavor.

1.2 Theory Matters

A black swan is typically defined as an extreme event that has not been observed before. Someone once told me that quantitative investment strategies are useless. Puzzled, I asked why. He replied, "Because the future is filled with black swans, and since historical data sets by definition cannot contain never-seen-before events, ML algorithms cannot be trained to predict them." I counter-argued that, in many cases, black swans have been predicted.

Let me explain this apparent paradox with an anecdote. Back in the year 2010, I was head of high-frequency futures at a large US hedge fund. On May 6, we were running our liquidity provision algorithms as usual, when around 12:30 ET, many of them started to flatten their positions automatically. We did not interfere or override the systems, so within minutes, our market exposure became very small. This system behavior had never happened to us before. My team and I were conducting a forensic analysis of what had caused the systems to shut themselves down when, at around 14:30 ET, we saw the S&P 500 plunge, within minutes, almost 10% relative to the open. Shortly after, the systems started to buy aggressively, profiting from a 5% rally into the market close. The press dubbed this black swan the "flash crash." We were twice surprised by this episode: first, we could not understand how our systems

[1] This is also true in the context of factor investing, where the systematic mistake can be explained in terms of behavioral bias, mismatched investment horizons, risk tolerance, regulatory constraints, and other variables informing investors' decisions.

predicted an event that we, the developers, did not anticipate; second, we could not understand why our systems started to buy shortly after the market bottomed.

About five months later, an official investigation found that the crash was likely caused by an order to sell 75,000 E-mini S&P 500 futures contracts at a high participation rate (CFTC 2010). That large order contributed to a persistent imbalance in the order flow, making it very difficult for market makers to flip their inventory without incurring losses. This toxic order flow triggered stop-out limits across market makers, who ceased to provide liquidity. Market makers became aggressive liquidity takers, and without anyone remaining on the bid, the market inevitably collapsed (Easley et al. 2011).

We could not have forecasted the flash crash by watching CNBC or reading the *Wall Street Journal*. To most observers, the flash crash was indeed an unpredictable black swan. However, the underlying causes of the flash crash are very common. Order flow is almost never perfectly balanced. In fact, imbalanced order flow is the norm, with various degrees of persistency (e.g., measured in terms of serial correlation). Our systems had been trained to reduce positions under extreme conditions of order flow imbalance. In doing so, they were trained to avoid the conditions that shortly after caused the black swan. Once the market collapsed, our systems recognized that the opportunity to buy at a 10% discount offset previous concerns from extreme order flow imbalance, and they took long positions until the close. This experience illustrates the two most important lessons contained in this Element.

1.2.1 Lesson 1: You Need a Theory

Contrary to popular belief, backtesting is not a research tool. Backtests can never prove that a strategy is a true positive, and they may only provide evidence that a strategy is a false positive. Never develop a strategy solely through backtests. Strategies must be supported by theory, not historical simulations. Your theories must be general enough to explain particular cases, even if those cases are black swans. The existence of black holes was predicted by the theory of general relativity more than five decades before the first black hole was observed. In the above story, our market microstructure theory (which later on became known as the VPIN theory; see Easley et al. 2011b) helped us predict and profit from a black swan. Not only that, but our theoretical work also contributed to the market's bounce back (my colleagues used to joke that we helped put the "flash" into the "flash crash"). This Element contains some of the tools you need to discover your own theories.

1.2.2 Lesson 2: ML Helps Discover Theories

Consider the following approach to discovering new financial theories. First, you apply ML tools to uncover the hidden variables involved in a complex

phenomenon. These are the ingredients that the theory must incorporate in order to make successful forecasts. The ML tools have identified these ingredients; however, they do not directly inform you about the exact equation that binds the ingredients together. Second, we formulate a theory that connects these ingredients through a structural statement. This structural statement is essentially a system of equations that hypothesizes a particular cause–effect mechanism. Third, the theory has a wide range of testable implications that go beyond the observations predicted by the ML tools in the first step.[2] A successful theory will predict events out-of-sample. Moreover, it will explain not only positives (x causes y) but also negatives (the absence of y is due to the absence of x).

In the above discovery process, ML plays the key role of decoupling the search for variables from the search for specification. Economic theories are often criticized for being based on "facts with unknown truth value" (Romer 2016) and "generally phony" assumptions (Solow 2010). Considering the complexity of modern financial systems, it is unlikely that a researcher will be able to uncover the ingredients of a theory by visual inspection of the data or by running a few regressions. Classical statistical methods do not allow this decoupling of the two searches.

Once the theory has been tested, it stands on its own feet. In this way, the theory, not the ML algorithm, makes the predictions. In the above anecdote, the theory, not an online forecast produced by an autonomous ML algorithm, shut the position down. The forecast was theoretically sound, and it was not based on some undefined pattern. It is true that the theory could not have been discovered without the help of ML techniques, but once the theory was discovered, the ML algorithm played no role in the decision to close the positions two hours prior to the flash crash. The most insightful use of ML in finance is for discovering theories. You may use ML successfully for making financial forecasts; however, that is not necessarily the best scientific use of this technology (particularly if your goal is to develop high-capacity investment strategies).

1.3 How Scientists Use ML

An ML algorithm learns complex patterns in a high-dimensional space with little human guidance on model specification. That ML models need not be specified by the researcher has led many to, erroneously, conclude that ML must

[2] A theory can be tested with more powerful tools than backtests. For instance, we could investigate which market makers lost money during the flash crash. Did they monitor for order flow imbalance? Did market makers that monitor for order flow imbalance fare better? Can we find evidence of their earlier retreat in the FIX messages of that day? A historical simulation of a trading rule cannot give us this level of insight.

be a black box. In that view, ML is merely an "oracle,"[3] a prediction machine from which no understanding can be extracted. The black box view of ML is a misconception. It is fueled by popular industrial applications of ML, where the search for better predictions outweighs the need for theoretical understanding. A review of recent scientific breakthroughs reveals radically different uses of ML in science, including the following:

1 **Existence:** ML has been deployed to evaluate the plausibility of a theory across all scientific fields, even beyond the empirical sciences. Notably, ML algorithms have helped make mathematical discoveries. ML algorithms cannot prove a theorem, however they can point to the existence of an undiscovered theorem, which can then be conjectured and eventually proved. In other words, if something can be predicted, there is hope that a mechanism can be uncovered (Gryak et al., forthcoming).
2 **Importance:** ML algorithms can determine the relative informational content of explanatory variables (features, in ML parlance) for explanatory and/or predictive purposes (Liu 2004). For example, the mean-decrease accuracy (MDA) method follows these steps: (1) Fit a ML algorithm on a particular data set; (2) derive the out-of-sample cross-validated accuracy; (3) repeat step (2) after shuffling the time series of individual features or combinations of features; (4) compute the decay in accuracy between (2) and (3). Shuffling the time series of an important feature will cause a significant decay in accuracy. Thus, although MDA does not uncover the underlying mechanism, it discovers the variables that should be part of the theory.
3 **Causation:** ML algorithms are often utilized to evaluate causal inference following these steps: (1) Fit a ML algorithm on historical data to predict outcomes, absent of an effect. This model is nontheoretical, and it is purely driven by data (like an oracle); (2) collect observations of outcomes under the presence of the effect; (3) use the ML algorithm fit in (1) to predict the observation collected in (2). The prediction error can be largely attributed to the effect, and a theory of causation can be proposed (Varian 2014; Athey 2015).
4 **Reductionist:** ML techniques are essential for the visualization of large, high-dimensional, complex data sets. For example, manifold learning algorithms can cluster a large number of observations into a reduced subset of peer groups, whose differentiating properties can then be analyzed (Schlecht et al. 2008).

[3] Here we use a common definition of oracle in complexity theory: a black box that is able to produce a solution for any instance of a given computational problem.

5 **Retriever:** ML is used to scan through big data in search of patterns that humans failed to recognize. For instance, every night ML algorithms are fed millions of images in search of supernovae. Once they find one image with a high probability of containing a supernova, expensive telescopes can be pointed to a particular region in the universe, where humans will scrutinize the data (Lochner et al. 2016). A second example is outlier detection. Finding outliers is a prediction problem rather than an explanation problem. A ML algorithm can detect an anomalous observation, based on the complex structure it has found in the data, even if that structure is not explained to us (Hodge and Austin 2004).

Rather than replacing theories, ML plays the critical role of helping scientists form theories based on rich empirical evidence. Likewise, ML opens the opportunity for economists to apply powerful data science tools toward the development of sound theories.

1.4 Two Types of Overfitting

The dark side of ML's flexibility is that, in inexperienced hands, these algorithms can easily overfit the data. The primary symptom of overfitting is a divergence between a model's in-sample and out-of-sample performance (known as the generalization error). We can distinguish between two types of overfitting: the overfitting that occurs on the train set, and the overfitting that occurs on the test set. Figure 1.1 summarizes how ML deals with both kinds of overfitting.

1.4.1 Train Set Overfitting

Train set overfitting results from choosing a specification that is so flexible that it explains not only the signal, but also the noise. The problem with confounding signal with noise is that noise is, by definition, unpredictable. An overfit model will produce wrong predictions with an unwarranted confidence, which in turn will lead to poor performance out-of-sample (or even in a pseudo-out-of-sample, like in a backtest).

ML researchers are keenly aware of this problem, which they address in three complementary ways. The first approach to correct for train set overfitting is evaluating the generalization error, through resampling techniques (such as cross-validation) and Monte Carlo methods. Appendix A describes these techniques and methods in greater detail. The second approach to reduce train set overfitting is regularization methods, which prevent model complexity unless it can be justified in terms of greater explanatory power. Model

Figure 1.1 Solutions to two kinds of overfitting.

parsimony can be enforced by limiting the number of parameters (e.g., LASSO) or restricting the model's structure (e.g., early stopping). The third approach to address train set overfitting is ensemble techniques, which reduce the variance of the error by combining the forecasts of a collection of estimators. For example, we can control the risk of overfitting a random forest on a train set in at least three ways: (1) cross-validating the forecasts; (2) limiting the depth of each tree; and (3) adding more trees.

In summary, a backtest may hint at the occurrence of train set overfitting, which can be remedied using the above approaches. Unfortunately, backtests are powerless against the second type of overfitting, as explained next.

1.4.2 Test Set Overfitting

Imagine that a friend claims to have a technique to predict the winning ticket at the next lottery. His technique is not exact, so he must buy more than one ticket. Of course, if he buys all of the tickets, it is no surprise that he will win. How many tickets would you allow him to buy before concluding that his method is useless? To evaluate the accuracy of his technique, you should adjust for the fact that he has bought multiple tickets. Likewise, researchers running multiple statistical tests on the same data set are more likely to make a false discovery. By applying the same test on the same data set multiple times, it is guaranteed that eventually a researcher will make a false discovery. This selection bias comes from fitting the model to perform well on the test set, not the train set.

Another example of test set overfitting occurs when a researcher backtests a strategy and she tweaks it until the output achieves a target performance. That backtest–tweak–backtest cycle is a futile exercise that will inevitably end with an overfit strategy (a false positive). Instead, the researcher should have spent her time investigating how the research process misled her into backtesting a false strategy. In other words, a poorly performing backtest is an opportunity to fix the research process, not an opportunity to fix a particular investment strategy.

Most published discoveries in finance are likely false, due to test set overfitting. ML did not cause the current crisis in financial research (Harvey et al. 2016). That crisis was caused by the widespread misuse of classical statistical methods in finance, and *p-hacking* in particular. ML can help deal with the problem of test set overfitting, in three ways. First, we can keep track of how many independent tests a researcher has run, to evaluate the probability that at least one of the outcomes is a false discovery (known as familywise error rate, or FWER). The deflated Sharpe ratio (Bailey and López de Prado 2014) follows

a similar approach in the context of backtesting, as explained in Section 8. It is the equivalent to controlling for the number of lottery tickets that your friend bought. Second, while it is easy to overfit a model to one test set, it is hard to overfit a model to thousands of test sets for each security. Those thousands of test sets can be generated by resampling combinatorial splits of train and test sets. This is the approached followed by the combinatorial purged cross-validation method, or CPCV (AFML, chapter 12). Third, we can use historical series to estimate the underlying data-generating process, and sample synthetic data sets that match the statistical properties observed in history. Monte Carlo methods are particularly powerful at producing synthetic data sets that match the statistical properties of a historical series. The conclusions from these tests are conditional to the representativeness of the estimated data-generating process (AFML, chapter 13). The main advantage of this approach is that those conclusions are not connected to a particular (observed) realization of the data-generating process but to an entire distribution of random realizations. Following with our example, this is equivalent to replicating the lottery game and repeating it many times, so that we can rule luck out.

In summary, there are multiple practical solutions to the problem of train set and test set overfitting. These solutions are neither infallible nor incompatible, and my advice is that you apply all of them. At the same time, I must insist that no backtest can replace a theory, for at least two reasons: (1) backtests cannot simulate black swans – only theories have the breadth and depth needed to consider the never-before-seen occurrences; (2) backtests may insinuate that a strategy is profitable, but they do not tell us why. They are not a controlled experiment. Only a theory can state the cause–effect mechanism, and formulate a wide range of predictions and implications that can be independently tested for facts and counterfacts. Some of these implications may even be testable outside the realm of investing. For example, the VPIN theory predicted that market makers would suffer stop-outs under persistent order flow imbalance. Beyond testing whether order flow imbalance causes a reduction in liquidity, researchers can also test whether market makers suffered losses during the flash crash (hint: they did). This latter test can be conducted by reviewing financial statements, independently from the evidence contained in exchange records of prices and quotes.

1.5 Outline

This Element offers asset managers a step-by-step guide to building financial theories with the help of ML methods. To that objective, each section uses what we have learned in the previous ones. Each section (except for this introduction)

contains an empirical analysis, where the methods explained are put to the test in Monte Carlo experiments.

The first step in building a theory is to collect data that illustrate how some variables relate to each other. In financial settings, those data often take the form of a covariance matrix. We use covariance matrices to run regressions, optimize portfolios, manage risks, search for linkages, etc. However, financial covariance matrices are notoriously noisy. A relatively small percentage of the information they contain is signal, which is systematically suppressed by arbitrage forces. Section 2 explains how to denoise a covariance matrix without giving up the little signal it contains. Most of the discussion centers on random matrix theory, but at the core of the solution sits an ML technique: the kernel density estimator.

Many research questions involve the notion of similarity or distance. For example, we may be interested in understanding how *closely* related two variables are. Denoised covariance matrices can be very useful for deriving distance metrics from linear relationships. Modeling nonlinear relationships requires more advanced concepts. Section 3 provides an information-theoretic framework for extracting complex signals from noisy data. In particular, it allows us to define distance metrics with minimal assumptions regarding the underlying variables that characterize the metric space. These distance metrics can be thought of as a nonlinear generalization of the notion of correlation.

One of the applications of distance matrices is to study whether some variables are more closely related among themselves than to the rest, hence forming clusters. Clustering has a wide range of applications across finance, like in asset class taxonomy, portfolio construction, dimensionality reduction, or modeling networks of agents. A general problem in clustering is finding the optimal number of clusters. Section 4 introduces the ONC algorithm, which provides a general solution to this problem. Various use cases for this algorithm are presented throughout this Element.

Clustering is an unsupervised learning problem. Before we can delve into supervised learning problems, we need to assess ways of labeling financial data. The effectiveness of a supervised ML algorithm greatly depends on the kind of problem we attempt to solve. For example, it may be harder to forecast tomorrow's S&P 500 return than the sign of its next 5% move. Different features are appropriate for different types of labels. Researchers should consider carefully what labeling method they apply on their data. Section 5 discusses the merits of various alternatives.

AFML warned readers that backtesting is not a research tool. Feature importance is. A backtest cannot help us develop an economic or financial theory.

In order to do that, we need a deeper understanding of what variables are involved in a phenomenon. Section 6 studies ML tools for evaluating the importance of explanatory variables, and explains how these tools defeat many of the caveats of classical methods, such as the *p*-value. A particular concern is how to overcome *p*-value's lack of robustness under multicollinearity. To tackle this problem, we must apply what we learned in all prior sections, including denoising (Section 2), distance metrics (Section 3), clustering (Section 4), and labeling (Section 5).

Once you have a financial theory, you can use your discovery to develop an investment strategy. Designing that strategy will require making some investment decisions under uncertainty. To that purpose, mean-variance portfolio optimization methods are universally known and used, even though they are notorious for their instability. Historically, this instability has been addressed in a number of ways, such as introducing strong constraints, adding priors, shrinking the covariance matrix, and other robust optimization techniques. Many asset managers are familiar with instability caused by noise in the covariance matrix. Fewer asset managers realize that certain data structures (types of signal) are also a source of instability for mean-variance solutions. Section 7 explains why signal can be a source of instability, and how ML methods can help correct it.

Finally, a financial ML book would not be complete without a detailed treatment of how to evaluate the probability that your discovery is false, as a result of test set overfitting. Section 8 explains the dangers of backtest overfitting, and provides several practical solutions to the problem of selection bias under multiple testing.

1.6 Audience

If, like most asset managers, you routinely compute covariance matrices, use correlations, search for low-dimensional representations of high-dimensional spaces, build predictive models, compute *p*-values, solve mean-variance optimizations, or apply the same test multiple times on a given data set, you need to read this Element. In it, you will learn that financial covariance matrices are noisy and that they need to be cleaned before running regressions or computing optimal portfolios (Section 2). You will learn that correlations measure a very narrow definition of codependency and that various information-theoretic metrics are more insightful (Section 3). You will learn intuitive ways of reducing the dimensionality of a space, which do not involve a change of basis. Unlike PCA, ML-based dimensionality reduction methods provide intuitive results (Section 4). Rather than aiming for implausible fixed-horizon predictions, you will learn alternative ways of posing financial prediction problems that can be solved with higher accuracy (Section 5). You will learn modern

alternatives to the classical *p*-values (Section 6). You will learn how to address the instability problem that plagues mean-variance investment portfolios (Section 7). And you will learn how to evaluate the probability that your discovery is false as a result of multiple testing (Section 8). If you work in the asset management industry or in academic finance, this Element is for you.

1.7 Five Popular Misconceptions about Financial ML

Financial ML is a new technology. As it is often the case with new technologies, its introduction has inspired a number of misconceptions. Below is a selection of the most popular.

1.7.1 ML Is the Holy Grail versus ML Is Useless

The amount of hype and counterhype surrounding ML defies logic. Hype creates a set of expectations that may not be fulfilled for the foreseeable future. Counterhype attempts to convince audiences that there is nothing special about ML and that classical statistical methods already produce the results that ML-enthusiasts claim.

ML critics sometimes argue that "caveat X in linear regression is no big deal," where X can either mean model misspecification, multicollinearity, missing regressors, nonlinear interaction effects, etc. In reality, any of these violations of classical assumptions will lead to accepting uninformed variables (a false positive) and/or rejecting informative variables (a false negative). For an example, see Section 6.

Another common error is to believe that the central limit theorem somehow justifies the use of linear regression models everywhere. The argument goes like this: with enough observations, Normality prevails, and linear models provide a good fit to the asymptotic correlation structure. This "CLT Hail Mary pass" is an undergrad fantasy: yes, the *sample mean* converges in distribution to a Gaussian, but not the sample itself! And that converge only occurs if the observations are independent and identically distributed. It takes a few lines of code to demonstrate that a misspecified regression will perform poorly, whether we feed it thousands or billions of observations.

Both extremes (hype and counterhype) prevent investors from recognizing the real and differentiated value that ML delivers today. ML is modern statistics, and it helps overcome many of the caveats of classical techniques that have preoccupied asset managers for decades. See López de Prado (2019c) for multiple examples of current applications of ML in finance.

1.7.2 ML Is a Black Box

This is perhaps the most widespread myth surrounding ML. Every research laboratory in the world uses ML to some extent, so clearly ML is compatible with the scientific method. Not only is ML not a black box, but as Section 6 explains, ML-based research tools can be more insightful than traditional statistical methods (including econometrics). ML models can be interpreted through a number of procedures, such as PDP, ICE, ALE, Friedman's H-stat, MDI, MDA, global surrogate, LIME, and Shapley values, among others. See Molnar (2019) for a detailed treatment of ML interpretability.

Whether someone applies ML as a black box or as a white-box is a matter of personal choice. The same is true of many other technical subjects. I personally do not care much about how my car works, and I must confess that I have never lifted the hood to take a peek at the engine (my thing is math, not mechanics). So, my car remains a black box to me. I do not blame the engineers who designed my car for my lack of curiosity, and I am aware that the mechanics who work at my garage see my car as a white box. Likewise, the assertion that ML is a black box reveals how some people have chosen to apply ML, and it is not a universal truth.

1.7.3 Finance Has Insufficient Data for ML

It is true that a few ML algorithms, particularly in the context of price prediction, require a lot of data. That is why a researcher must choose the right algorithm for a particular job. On the other hand, ML critics who wield this argument seem to ignore that many ML applications in finance do not require any historical data at all. Examples include risk analysis, portfolio construction, outlier detection, feature importance, and bet-sizing methods. Each section in this Element demonstrates the mathematical properties of ML without relying on any historical series. For instance, Section 7 evaluates the accuracy of an ML-based portfolio construction algorithm via Monte Carlo experiments. Conclusions drawn from millions of Monte Carlo simulations teach us something about the general mathematical properties of a particular approach. The anecdotal evidence derived from a handful of historical simulations is no match to evaluating a wide range of scenarios.

Other financial ML applications, like sentiment analysis, deep hedging, credit ratings, execution, and private commercial data sets, enjoy an abundance of data. Finally, in some settings, researchers can conduct randomized controlled experiments, where they can generate their own data and establish precise cause–effect mechanisms. For example, we may reword a news article and compare ML's sentiment extraction with a human's conclusion, controlling for

various changes. Likewise, we may experiment with the market's reaction to alternative implementations of an execution algorithm under comparable conditions.

1.7.4 The Signal-to-Noise Ratio Is Too Low in Finance

There is no question that financial data sets exhibit lower signal-to-noise ratio than those used by other ML applications (a point that we will demonstrate in Section 2). Because the signal-to-noise ratio is so low in finance, data alone are not good enough for relying on black box predictions. That does not mean that ML cannot be used in finance. It means that we must use ML differently, hence the notion of financial ML as a distinct subject of study. Financial ML is not the mere application of standard ML to financial data sets. Financial ML comprises ML techniques specially designed to tackle the specific challenges faced by financial researchers, just as econometrics is not merely the application of standard statistical techniques to economic data sets.

The goal of financial ML ought to be to assist researchers in the discovery of new economic theories. The theories so discovered, and not the ML algorithms, will produce forecasts. This is no different than the way scientists utilize ML across all fields of research.

1.7.5 The Risk of Overfitting Is Too High in Finance

Section 1.4 debunked this myth. In knowledgeable hands, ML algorithms overfit less than classical methods. I concede, however, that in nonexpert hands ML algorithms can cause more harm than good.

1.8 The Future of Financial Research

The International Data Corporation has estimated that 80% of all available data are unstructured (IDC 2014). Many of the new data sets available to researchers are high-dimensional, sparse, or nonnumeric. As a result of the complexities of these new data sets, there is a limit to how much researchers can learn using regression models and other linear algebraic or geometric approaches. Even with older data sets, traditional quantitative techniques may fail to capture potentially complex (e.g., nonlinear and interactive) associations among variables, and these techniques are extremely sensitive to the multicollinearity problem that pervades financial data sets (López de Prado 2019b).

Economics and finance have much to benefit from the adoption of ML methods. As of November 26, 2018, the Web of Science[4] lists 13,772 journal

[4] www.webofknowledge.com.

articles on subjects in the intersection of "Economics" and "Statistics & Probability." Among those publications, only eighty-nine articles (0.65%) contain any of the following terms: classifier, clustering, neural network, or machine learning. To put it in perspective, out of the 40,283 articles in the intersection of "Biology" and "Statistics & Probability," a total of 4,049 (10.05%) contained any of those terms, and out of the 4,994 articles in the intersection of "Chemistry, Analytical" and "Statistics & Probability," a total of 766 (15.34%) contained any of those terms.

The econometric canon predates the dawn of digital computing. Most econometric models were devised for estimation by hand and are a product of their time. In the words of Robert Tibshirani, "people use certain methods because that is how it all started and that's what they are used to. It's hard to change it."[5] Students in the twenty-first century should not be overexposed to legacy technologies. Moreover, the most successful quantitative investment firms in history rely primarily on ML, not econometrics, and the current predominance of econometrics in graduate studies prepares students for academic careers, not for jobs in the industry.

This does not mean that econometrics has outlived its usability. Researchers asked to decide between econometrics and ML are presented with a false choice. ML and econometrics complement each other, because they have different strengths. For example, ML can be particularly helpful at suggesting to researchers the ingredients of a theory (see Section 6), and econometrics can be useful at testing a theory that is well grounded on empirical observation. In fact, sometimes we may want to apply both paradigms at the same time, like in semiparametric methods. For example, a regression could combine observable explanatory variables with control variables that are contributed by an ML algorithm (Mullainathan and Spiess 2017). Such approach would address the bias associated with omitted regressors (Clarke 2005).

1.9 Frequently Asked Questions

Over the past few years, attendees at seminars have asked me all sorts of interesting questions. In this section I have tried to provide a short answer to some of the most common questions. I have also added a couple of questions that I am still hoping that someone will ask one day.

In Simple Terms, What Is ML?

Broadly speaking, ML refers to the set of algorithms that learn complex patterns in a high-dimensional space without being specifically directed. Let us break

[5] https://qz.com/1206229/this-is-the-best-book-for-learning-modern-statistics-its-free/.

that definition into its three components. First, ML learns without being specifically directed, because researchers impose very little structure on the data. Instead, the algorithm derives that structure from the data. Second, ML learns complex patterns, because the structure identified by the algorithm may not be representable as a finite set of equations. Third, ML learns in a high-dimensional space, because solutions often involve a large number of variables, and the interactions between them.

For example, we can train an ML algorithm to recognize human faces by showing it examples. We do not define what a face is, hence the algorithm learns without our direction. The problem is never posed in terms of equations, and in fact the problem may not be expressible in terms of equations. And the algorithm uses an extremely large number of variables to perform this task, including the individual pixels and the interaction between the pixels.

In recent years, ML has become an increasingly useful research tool throughout all fields of scientific research. Examples include drug development, genome research, new materials, and high-energy physics. Consumer products and industrial services have quickly incorporated these technologies, and some of the most valuable companies in the world produce ML-based products and services.

How Is ML Different from Econometric Regressions?

Researchers use traditional regressions to fit a predefined functional form to a set of variables. Regressions are extremely useful when we have a high degree of conviction regarding that functional form and all the interaction effects that bind the variables together. Going back to the eighteenth century, mathematicians developed tools that fit those functional forms using estimators with desirable properties, subject to certain assumptions on the data.

Starting in the 1950s, researchers realized that there was a different way to conduct empirical analyses, with the help of computers. Rather than imposing a functional form, particularly when that form is unknown ex ante, they would allow algorithms to figure out variable dependencies from the data. And rather than making strong assumptions on the data, the algorithms would conduct experiments that evaluate the mathematical properties of out-of-sample predictions. This relaxation in terms of functional form and data assumptions, combined with the use of powerful computers, opened the door to analyzing complex data sets, including highly nonlinear, hierarchical, and noncontinuous interaction effects.

Consider the following example: a researcher wishes to estimate the survival probability of a passenger on the *Titanic*, based on a number of variables, such

as gender, ticket class, and age. A typical regression approach would be to fit a logit model to a binary variable, where 1 means survivor and 0 means deceased, using gender, ticket class, and age as regressors. It turns out that, even though these regressors are correct, a logit (or probit) model fails to make good predictions. The reason is that logit models do not recognize that this data set embeds a hierarchical (treelike) structure, with complex interactions. For example, adult males in second class died at a much higher rate than each of these attributes taken independently. In contrast, a simple "classification tree" algorithm performs substantially better, because we allow the algorithm to find that hierarchical structure (and associated complex interactions) for us.

As it turns out, hierarchical structures are omnipresent in economics and finance (Simon 1962). Think of sector classifications, credit ratings, asset classes, economic linkages, trade networks, clusters of regional economies, etc. When confronted with these kinds of problems, ML tools can complement and overcome the limitations of econometrics or similar traditional statistical methods.

How Is ML Different from Big Data?

The term *big data* refers to data sets that are so large and/or complex that traditional statistical techniques fail to extract and model the information contained in them. It is estimated that 90% of all recorded data have been created over the past two years, and 80% of the data is unstructured (i.e., not directly amenable to traditional statistical techniques).

In recent years, the quantity and granularity of economic data have improved dramatically. The good news is that the sudden explosion of administrative, private sector, and micro-level data sets offers an unparalleled insight into the inner workings of the economy. The bad news is that these data sets pose multiple challenges to the study of economics. (1) Some of the most interesting data sets are unstructured. They can also be nonnumerical and noncategorical, like news articles, voice recordings, or satellite images. (2) These data sets are high-dimensional (e.g., credit card transactions.) The number of variables involved often greatly exceeds the number of observations, making it very difficult to apply linear algebra solutions. (3) Many of these data sets are extremely sparse. For instance, samples may contain a large proportion of zeros, where standard notions such as correlation do not work well. (4) Embedded within these data sets is critical information regarding networks of agents, incentives, and aggregate behavior of groups of people. ML techniques are designed for analyzing big data, which is why they are often cited together.

How Is the Asset Management Industry Using ML?

Perhaps the most popular application of ML in asset management is price prediction. But there are plenty of equally important applications, like hedging, portfolio construction, detection of outliers and structural breaks, credit ratings, sentiment analysis, market making, bet sizing, securities taxonomy, and many others. These are real-life applications that transcend the hype often associated with expectations of price prediction.

For example, factor investing firms use ML to redefine value. A few years ago, price-to-earnings ratios may have provided a good ranking for value, but that is not the case nowadays. Today, the notion of value is much more nuanced. Modern asset managers use ML to identify the traits of value, and how those traits interact with momentum, quality, size, etc. Meta-labeling (Section 5.5) is another hot topic that can help asset managers size and time their factor bets.

High-frequency trading firms have utilized ML for years to analyze real-time exchange feeds, in search for footprints left by informed traders. They can utilize this information to make short-term price predictions or to make decisions on the aggressiveness or passiveness in order execution. Credit rating agencies are also strong adopters of ML, as these algorithms have demonstrated their ability to replicate the ratings generated by credit analysts. Outlier detection is another important application, since financial models can be very sensitive to the presence of even a small number of outliers. ML models can help improve investment performance by finding the proper size of a position, leaving the buy-or-sell decision to traditional or fundamental models.

And Quantitative Investors Specifically?

All of the above applications, and many more, are relevant to quantitative investors. It is a great time to be a quant. Data are more abundant than ever, and computers are finally delivering the power needed to make effective use of ML. I am particularly excited about real-time prediction of macroeconomic statistics, following the example of MIT's Billion Prices Project (Cavallo and Rigobon 2016). ML can be specially helpful at uncovering relationships that until now remained hidden, even in traditional data sets. For instance, the economic relationships between companies may not be effectively described by traditional sector-group-industry classifications, such as GICS.[6] A network approach, where companies are related according to a variety of factors, is likely

[6] www.msci.com/gics.

to offer a richer and more accurate representation of the dynamics, strengths, and vulnerabilities of specific segments of the stock or credit markets (Cohen and Frazzini 2008).

What Are Some of the Ways That ML Can Be Applied to Investor Portfolios?

Portfolio construction is an extremely promising area for ML (Section 7). For many decades, the asset management industry has relied on variations and refinements of Markowitz's efficient frontier to build investment portfolios. It is known that many of these solutions are optimal in-sample, however, they can perform poorly out-of-sample due to the computational instabilities involved in convex optimization. Numerous classical approaches have attempted, with mixed success, to address these computational instabilities. ML algorithms have shown the potential to produce robust portfolios that perform well out-of-sample, thanks to their ability to recognize sparse hierarchical relationships that traditional methods miss (López de Prado 2016).

What Are the Risks? Is There Anything That Investors Should Be Aware of or Look Out For?

Finance is not a plug-and-play subject as it relates to ML. Modeling financial series is harder than driving cars or recognizing faces. The reason is, the signal-to-noise ratio in financial data is extremely low, as a result of arbitrage forces and nonstationary systems. The computational power and functional flexibility of ML ensures that it will always find a pattern in the data, even if that pattern is a fluke rather than the result of a persistent phenomenon. An "oracle" approach to financial ML, where algorithms are developed to form predictions divorced from all economic theory, is likely to yield false discoveries. I have never heard a scientist say "Forget about theory, I have this oracle that can answer anything, so let's all stop thinking, and let's just believe blindly whatever comes out."

It is important for investors to recognize that ML is not a substitute for economic theory, but rather a powerful tool for building modern economic theories. We need ML to develop better financial theories, and we need financial theories to restrict ML's propensity to overfit. Without this theory–ML interplay, investors are placing their trust on high-tech horoscopes.

How Do You Expect ML to Impact the Asset Management Industry in the Next Decade?

Today, the amount of ML used by farmers is staggering: self-driving tractors, drones scanning for irregular patches of land, sensors feeding cattle and

administering nutrients as needed, genetically engineered crops, satellite images for estimating yields, etc. Similarly, I think in ten years we will look back, and ML will be an important aspect of asset management. And just like in the farming industry, although this transformation may not happen overnight, it is clear that there is only one direction forward.

Economic data sets will only get bigger, and computers will only get more powerful. Most asset managers will fail either by not evolving or by rushing into the unknown without fully recognizing the dangers involved in the "oracle" approach. Only a few asset managers will succeed by evolving in a thoughtful and responsible manner.

How Do You Expect ML to Impact Financial Academia in the Next Decade?

Imagine if physicists had to produce theories in a universe where the fundamental laws of nature are in a constant flux; where publications have an impact on the very phenomenon under study; where experimentation is virtually impossible; where data are costly, the signal is dim, and the system under study is incredibly complex . . . I feel utmost admiration for how much financial academics have achieved in the face of paramount adversity.

ML has a lot to offer to the academic profession. First, ML provides the power and flexibility needed to find dim signals in the sea of noise caused by arbitrage forces. Second, ML allows academics to decouple the research process into two stages: (1) search for important variables irrespective of functional form, and (2) search for a functional form that binds those variables. López de Prado (2019b) demonstrates how even small specification errors mislead researchers into rejecting important variables. It is hard to overstate the relevance of decoupling the specification search from the variables search. Third, ML offers the possibility of conducting simulations on synthetic data. This is as close as finance will ever get to experimentation, in the absence of laboratories. We live an exciting time to do academic research on financial systems, and I expect tremendous breakthroughs as more financial researchers embrace ML.

Isn't Financial ML All about Price Prediction?

One of the greatest misunderstandings I perceive from reading the press is the notion that ML's main (if not only) objective is price prediction. Asset pricing is undoubtedly a very worthy endeavor, however its importance is often overstated. Having an edge at price prediction is just one necessary, however entirely

insufficient, condition to be successful in today's highly competitive market. Other areas that are equally important are data processing, portfolio construction, risk management, monitoring for structural breaks, bet sizing, and detection of false investment strategies, just to cite a few.

Consider the players at the World Series of Poker. The cards are shuffled and distributed randomly. These players obviously cannot predict what cards will be handed to players with any meaningful accuracy. And yet, the same handful of players ends up in top positions year after year. One reason is, bet sizing is more important than card prediction. When a player receives a good hand, he evaluates the probability that another player may hold a strong hand too, and bets strategically. Likewise, investors may not be able to predict prices, however they may recognize when an out-of-the-normal price has printed, and bet accordingly. I am not saying that bet sizing is the key to successful investing. I am merely stating that bet sizing is at least as important as price prediction, and that portfolio construction is arguably even more important.

Why Don't You Discuss a Wide Range of ML Algorithms?

The purpose of this Element is not to introduce the reader to the vast population of ML algorithms used today in finance. There are two reasons for that. First, there are lengthy textbooks dedicated to the systematic exposition of those algorithms, and another one is hardly needed. Excellent references include James et al. (2013), Hastie et al. (2016), and Efron and Hastie (2016). Second, financial data sets have specific nuisances, and the success or failure of a project rests on understanding them. Once we have engineered the features and posed the problem correctly, choosing an algorithm plays a relatively secondary role.

Allow me to illustrate the second point with an example. Compare an algorithm that forecasted a change of 1, but received a realized change of 3, with another algorithm that forecasted a change of −1, but received a realized change of 1. In both cases, the forecast error is 2. In many industrial applications, we would be indifferent between both errors. That is not the case in finance. In the first instance, an investor makes one-third of the predicted profit, whereas in the second instance the investor suffers a loss equal to the predicted profit. Failing to predict the size is an opportunity loss, but failing to predict the sign is an actual loss. Investors penalize actual losses much more than opportunity losses. Predicting the sign of an outcome is often more important than predicting its size, and a reason for favoring classifiers over regression methods in finance. In addition, it is common in finance to find that the sign and size of an outcome depend on different features, so jointly

forecasting the sign and size of an outcome with a unique set of features can lead to subpar results.[7] ML experts who transition into finance from other fields often make fundamental mistakes, like posing problems incorrectly, as explained in López de Prado (2018b). Financial ML is a subject in its own right, and the discussion of generic ML algorithms is not the heart of the matter.

Why Don't You Discuss a Specific Investment Strategy, Like Many Other Books Do?

There are plenty of books in the market that provide recipes for implementing someone else's investment strategy. Those cookbooks show us how to prepare someone else's cake. This Element is different. I want to show you how you can use ML to discover new economic and financial theories that are relevant to you, on which you can base your proprietary investment strategies. Your investment strategies are just the particular implementation of the theories that first you must discover independently. You cannot bake someone else's cake and expect to retain it for yourself.

1.10 Conclusions

The purpose of this Element is to introduce ML tools that are useful for discovering economic and financial theories. Successful investment strategies are specific implementations of general theories. An investment strategy that lacks a theoretical justification is likely to be false. Hence, a researcher should concentrate her efforts on developing a theory, rather than of backtesting potential strategies.

ML is not a black box, and it does not necessarily overfit. ML tools complement rather than replace the classical statistical methods. Some of ML's strengths include (1) Focus on out-of-sample predictability over variance adjudication; (2) usage of computational methods to avoid relying on (potentially unrealistic) assumptions; (3) ability to "learn" complex specifications, including nonlinear, hierarchical, and noncontinuous interaction effects in a high-dimensional space; and (4) ability to disentangle the variable search from the specification search, in a manner robust to multicollinearity and other substitution effects.

[7] See López de Prado (2018a) for a discussion of meta-labeling algorithms, where the sign and size decision is made by independent algorithms.

1.11 Exercises

1 Can quantitative methods be used to predict events that never happened before? How could quantitative methods predict a black swan?
2 Why is theory particularly important in finance and economics? What is the best use of ML in finance?
3 What are popular misconceptions about financial ML? Are financial data sets large enough for ML applications?
4 How does ML control for overfitting? Is the signal-to-noise ratio too low in finance for allowing the use of ML?
5 Describe a quantitative approach in finance that combines classical and ML methods. How is ML different from a large regression? Describe five applications of financial ML.

Bibliography

Aggarwal, C., and C. Reddy (2014): *Data Clustering – Algorithms and Applications*. 1st ed. CRC Press.

Ahmed, N., A. Atiya, N. Gayar, and H. El-Shishiny (2010): "An Empirical Comparison of Machine Learning Models for Time Series Forecasting." *Econometric Reviews*, Vol. 29, No. 5–6, pp. 594–621.

Anderson, G., A. Guionnet, and O. Zeitouni (2009): *An Introduction to Random Matrix Theory.* 1st ed. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Ballings, M., D. van den Poel, N. Hespeels, and R. Gryp (2015): "Evaluating Multiple Classifiers for Stock Price Direction Prediction." *Expert Systems with Applications*, Vol. 42, No. 20, pp. 7046–56.

Bansal, N., A. Blum, and S. Chawla (2004): "Correlation Clustering." *Machine Learning*, Vol. 56, No. 1, pp. 89–113.

Benjamini, Y., and D. Yekutieli (2001): "The Control of the False Discovery Rate in Multiple Testing under Dependency." *Annals of Statistics*, Vol. 29, pp. 1165–88.

Benjamini, Y., and W. Liu (1999): "A Step-Down Multiple Hypotheses Testing Procedure that Controls the False Discovery Rate under Independence." *Journal of Statistical Planning and Inference*, Vol. 82, pp. 163–70.

Benjamini, Y., and Y. Hochberg (1995): "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society*, Series B, Vol. 57, pp. 289–300.

Bontempi, G., S. Taieb, and Y. Le Borgne (2012): "Machine Learning Strategies for Time Series Forecasting." *Lecture Notes in Business Information Processing*, Vol. 138, No. 1, pp. 62–77.

Booth, A., E. Gerding, and F. McGroarty (2014): "Automated Trading with Performance Weighted Random Forests and Seasonality." *Expert Systems with Applications*, Vol. 41, No. 8, pp. 3651–61.

Cao, L., and F. Tay (2001): "Financial Forecasting Using Support Vector Machines." *Neural Computing and Applications*, Vol. 10, No. 2, pp. 184–92.

Cao, L., F. Tay, and F. Hock (2003): "Support Vector Machine with Adaptive Parameters in Financial Time Series Forecasting." *IEEE Transactions on Neural Networks*, Vol. 14, No. 6, pp. 1506–18.

Cervello-Royo, R., F. Guijarro, and K. Michniuk (2015): "Stockmarket Trading Rule Based on Pattern Recognition and Technical Analysis: Forecasting the

DJIA Index with Intraday Data." *Expert Systems with Applications*, Vol. 42, No. 14, pp. 5963–75.

Chang, P., C. Fan, and J. Lin (2011): "Trend Discovery in Financial Time Series Data Using a Case-Based Fuzzy Decision Tree." *Expert Systems with Applications*, Vol. 38, No. 5, pp. 6070–80.

Chen, B., and J. Pearl (2013): "Regression and Causation: A Critical Examination of Six Econometrics Textbooks." *Real-World Economics Review*, Vol. 65, pp. 2–20.

Creamer, G., and Y. Freund (2007): "A Boosting Approach for Automated Trading." *Journal of Trading*, Vol. 2, No. 3, pp. 84–96.

Creamer, G., and Y. Freund (2010): "Automated Trading with Boosting and Expert Weighting." *Quantitative Finance*, Vol. 10, No. 4, pp. 401–20.

Creamer, G., Y. Ren, Y. Sakamoto, and J. Nickerson (2016): "A Textual Analysis Algorithm for the Equity Market: The European Case." *Journal of Investing*, Vol. 25, No. 3, pp. 105–16.

Dixon, M., D. Klabjan, and J. Bang (2017): "Classification-Based Financial Markets Prediction Using Deep Neural Networks." *Algorithmic Finance*, Vol. 6, No. 3, pp. 67–77.

Dunis, C., and M. Williams (2002): "Modelling and Trading the Euro/US Dollar Exchange Rate: Do Neural Network Models Perform Better?" *Journal of Derivatives and Hedge Funds*, Vol. 8, No. 3, pp. 211–39.

Easley, D., and J. Kleinberg (2010): *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. 1st ed. Cambridge University Press.

Easley, D., M. López de Prado, M. O'Hara, and Z. Zhang (2011): "Microstructure in the Machine Age." Working paper.

Efroymson, M. (1960): "Multiple Regression Analysis." In A. Ralston and H. Wilf (eds.), *Mathematical Methods for Digital Computers*. 1st ed. Wiley.

Einav, L., and J. Levin (2014): "Economics in the Age of Big Data." *Science*, Vol. 346, No. 6210. Available at http://science.sciencemag.org/content/346/6210/1243089

Feuerriegel, S., and H. Prendinger (2016): "News-Based Trading Strategies." *Decision Support Systems*, Vol. 90, pp. 65–74.

Greene, W. (2012): *Econometric Analysis*. 7th ed. Pearson Education.

Harvey, C., and Y. Liu (2015): "Backtesting." *The Journal of Portfolio Management*, Vol. 42, No. 1, pp. 13–28.

Harvey, C., and Y. Liu (2018): "False (and Missed) Discoveries in Financial Economics." Working paper. Available at https://ssrn.com/abstract=3073799

Harvey, C., and Y. Liu (2018): "Lucky Factors." Working paper. Available at https://ssrn.com/abstract=2528780

Hastie, T., R. Tibshirani, and J. Friedman (2016): *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd ed. Springer.

Hayashi, F. (2000): *Econometrics*. 1st ed. Princeton University Press.

Holm, S. (1979): "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics*, Vol. 6, pp. 65–70.

Hsu, S., J. Hsieh, T. Chih, and K. Hsu (2009): "A Two-Stage Architecture for Stock Price Forecasting by Integrating Self-Organizing Map and Support Vector Regression." *Expert Systems with Applications*, Vol. 36, No. 4, pp. 7947–51.

Huang, W., Y. Nakamori, and S. Wang (2005): "Forecasting Stock Market Movement Direction with Support Vector Machine." *Computers and Operations Research*, Vol. 32, No. 10, pp. 2513–22.

Ioannidis, J. (2005): "Why Most Published Research Findings Are False." *PLoS Medicine*, Vol. 2, No. 8. Available at https://doi.org/10.1371/journal.pmed.0020124

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013): *An Introduction to Statistical Learning*. 1st ed. Springer.

Kahn, R. (2018): *The Future of Investment Management*. 1st ed. CFA Institute Research Foundation.

Kara, Y., M. Boyacioglu, and O. Baykan (2011): "Predicting Direction of Stock Price Index Movement Using Artificial Neural Networks and Support Vector Machines: The Sample of the Istanbul Stock Exchange." *Expert Systems with Applications*, Vol. 38, No. 5, pp. 5311–19.

Kim, K. (2003): "Financial Time Series Forecasting Using Support Vector Machines." *Neurocomputing*, Vol. 55, No. 1, pp. 307–19.

Kolanovic, M., and R. Krishnamachari (2017): "Big Data and AI Strategies: Machine Learning and Alternative Data Approach to Investing." *J.P. Morgan Quantitative and Derivative Strategy*, May.

Kolm, P., R. Tutuncu, and F. Fabozzi (2010): "60 Years of Portfolio Optimization." *European Journal of Operational Research*, Vol. 234, No. 2, pp. 356–71.

Krauss, C., X. Do, and N. Huck (2017): "Deep Neural Networks, Gradient-Boosted Trees, Random Forests: Statistical Arbitrage on the S&P 500." *European Journal of Operational Research*, Vol. 259, No. 2, pp. 689–702.

Kuan, C., and L. Tung (1995): "Forecasting Exchange Rates Using Feedforward and Recurrent Neural Networks." *Journal of Applied Econometrics*, Vol. 10, No. 4, pp. 347–64.

Kuhn, H. W., and A. W. Tucker (1952): "Nonlinear Programming." In *Proceedings of 2nd Berkeley Symposium*. University of California Press, pp. 481–92.

Laborda, R., and J. Laborda (2017): "Can Tree-Structured Classifiers Add Value to the Investor?" *Finance Research Letters*, Vol. 22, pp. 211–26.

López de Prado, M. (2018): "A Practical Solution to the Multiple-Testing Crisis in Financial Research." *Journal of Financial Data Science*, Vol. 1, No. 1. Available at https://ssrn.com/abstract=3177057

López de Prado, M., and M. Lewis (2018): "Confidence and Power of the Sharpe Ratio under Multiple Testing." Working paper. Available at https://ssrn.com/abstract=3193697

MacKay, D. (2003): *Information Theory, Inference, and Learning Algorithms*. 1st ed. Cambridge University Press.

Marcenko, V., and L. Pastur (1967): "Distribution of Eigenvalues for Some Sets of Random Matrices." *Matematicheskii Sbornik*, Vol. 72, No. 4, pp. 507–36.

Michaud, R. (1998): *Efficient Asset Allocation: A Practical Guide to Stock Portfolio Optimization and Asset Allocation*. Boston: Harvard Business School Press.

Nakamura, E. (2005): "Inflation Forecasting Using a Neural Network." *Economics Letters*, Vol. 86, No. 3, pp. 373–78.

Olson, D., and C. Mossman (2003): "Neural Network Forecasts of Canadian Stock Returns Using Accounting Ratios." *International Journal of Forecasting*, Vol. 19, No. 3, pp. 453–65.

Otto, M. (2016): *Chemometrics: Statistics and Computer Application in Analytical Chemistry*. 3rd ed. Wiley.

Patel, J., S. Sha, P. Thakkar, and K. Kotecha (2015): "Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques." *Expert Systems with Applications*, Vol. 42, No. 1, pp. 259–68.

Pearl, J. (2009): "Causal Inference in Statistics: An Overview." *Statistics Surveys*, Vol. 3, pp. 96–146.

Plerou, V., P. Gopikrishnan, B. Rosenow, L. Nunes Amaral, and H. Stanley (1999): "Universal and Nonuniversal Properties of Cross Correlations in Financial Time Series." *Physical Review Letters*, Vol. 83, No. 7, pp. 1471–74.

Porter, K. (2017): "Estimating Statistical Power When Using Multiple Testing Procedures." Available at www.mdrc.org/sites/default/files/PowerMultiplicity-IssueFocus.pdf

Potter, M., J. P. Bouchaud, and L. Laloux (2005): "Financial Applications of Random Matrix Theory: Old Laces and New Pieces." *Acta Physica Polonica B*, Vol. 36, No. 9, pp. 2767–84.

Qin, Q., Q. Wang, J. Li, and S. Shuzhi (2013): "Linear and Nonlinear Trading Models with Gradient Boosted Random Forests and Application to Singapore Stock Market." *Journal of Intelligent Learning Systems and Applications*, Vol. 5, No. 1, pp. 1–10.

Robert, C. (2014): "On the Jeffreys–Lindley Paradox." *Philosophy of Science*, Vol. 81, No. 2, pp. 216–32.

Shafer, G. (1982): "Lindley's Paradox." *Journal of the American Statistical Association*, Vol. 77, No. 378, pp. 325–34.

Simon, H. (1962): "The Architecture of Complexity." *Proceedings of the American Philosophical Society*, Vol. 106, No. 6, pp. 467–82.

SINTEF (2013): "Big Data, for Better or Worse: 90% of World's Data Generated over Last Two Years." *Science Daily*, May 22. Available at www.sciencedaily.com/releases/2013/05/130522085217.htm

Sorensen, E., K. Miller, and C. Ooi (2000): "The Decision Tree Approach to Stock Selection." *Journal of Portfolio Management*, Vol. 27, No. 1, pp. 42–52.

Theofilatos, K., S. Likothanassis, and A. Karathanasopoulos (2012): "Modeling and Trading the EUR/USD Exchange Rate Using Machine Learning Techniques." *Engineering, Technology and Applied Science Research*, Vol. 2, No. 5, pp. 269–72.

Trafalis, T., and H. Ince (2000): "Support Vector Machine for Regression and Applications to Financial Forecasting." *Neural Networks*, Vol. 6, No. 1, pp. 348–53.

Trippi, R., and D. DeSieno (1992): "Trading Equity Index Futures with a Neural Network." *Journal of Portfolio Management*, Vol. 19, No. 1, pp. 27–33.

Tsai, C., and S. Wang (2009): "Stock Price Forecasting by Hybrid Machine Learning Techniques." *Proceedings of the International Multi-Conference of Engineers and Computer Scientists*, Vol. 1, No. 1, pp. 755–60.

Tsai, C., Y. Lin, D. Yen, and Y. Chen (2011): "Predicting Stock Returns by Classifier Ensembles." *Applied Soft Computing*, Vol. 11, No. 2, pp. 2452–59.

Tsay, R. (2013): *Multivariate Time Series Analysis: With R and Financial Applications*. 1st ed. Wiley.

Wang, J., and S. Chan (2006): "Stock Market Trading Rule Discovery Using Two-Layer Bias Decision Tree." *Expert Systems with Applications*, Vol. 30, No. 4, pp. 605–11.

Wang, Q., J. Li, Q. Qin, and S. Ge (2011): "Linear, Adaptive and Nonlinear Trading Models for Singapore Stock Market with Random Forests." In *Proceedings of the 9th IEEE International Conference on Control and Automation*, pp. 726–31.

Wei, P., and N. Wang (2016): "Wikipedia and Stock Return: Wikipedia Usage Pattern Helps to Predict the Individual Stock Movement." In *Proceedings of the 25th International Conference Companion on World Wide Web*, Vol. 1, pp. 591–94.

Wooldridge, J. (2010): *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. MIT Press.

Wright, S. (1921): "Correlation and Causation." *Journal of Agricultural Research*, Vol. 20, pp. 557–85.

Żbikowski, K. (2015): "Using Volume Weighted Support Vector Machines with Walk Forward Testing and Feature Selection for the Purpose of Creating Stock Trading Strategy." *Expert Systems with Applications*, Vol. 42, No. 4, pp. 1797–1805.

Zhang, G., B. Patuwo, and M. Hu (1998): "Forecasting with Artificial Neural Networks: The State of the Art." *International Journal of Forecasting*, Vol. 14, No. 1, pp. 35–62.

Zhu, M., D. Philpotts, and M. Stevenson (2012): "The Benefits of Tree-Based Models for Stock Selection." *Journal of Asset Management*, Vol. 13, No. 6, pp. 437–48.

Zhu, M., D. Philpotts, R. Sparks, and J. Stevenson (2011): "A Hybrid Approach to Combining CART and Logistic Regression for Stock Ranking." *Journal of Portfolio Management*, Vol. 38, No. 1, pp. 100–109.

References

American Statistical Association (2016): "Statement on Statistical Significance and P-Values." Available at www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

Apley, D. (2016): "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models." Available at https://arxiv.org/abs/1612.08468

Athey, Susan (2015): "Machine Learning and Causal Inference for Policy Evaluation." In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 5–6. ACM.

Bailey, D., and M. López de Prado (2012): "The Sharpe Ratio Efficient Frontier." *Journal of Risk*, Vol. 15, No. 2, pp. 3–44.

Bailey, D., and M. López de Prado (2013): "An Open-Source Implementation of the Critical-Line Algorithm for Portfolio Optimization." *Algorithms*, Vol. 6, No. 1, pp. 169–96. Available at http://ssrn.com/abstract=2197616

Bailey, D., and M. López de Prado (2014): "The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting and Non-Normality." *Journal of Portfolio Management*, Vol. 40, No. 5, pp. 94–107.

Bailey, D., J. Borwein, M. López de Prado, and J. Zhu (2014): "Pseudo-mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance." *Notices of the American Mathematical Society*, Vol. 61, No. 5, pp. 458–71. Available at http://ssrn.com/abstract=2308659

Black, F., and R. Litterman (1991): "Asset Allocation Combining Investor Views with Market Equilibrium." *Journal of Fixed Income*, Vol. 1, No. 2, pp. 7–18.

Black, F., and R. Litterman (1992): "Global Portfolio Optimization." *Financial Analysts Journal*, Vol. 48, No. 5, pp. 28–43.

Breiman, L. (2001): "Random Forests." *Machine Learning*, Vol. 45, No. 1, pp. 5–32.

Brian, E., and M. Jaisson (2007): "Physico-theology and Mathematics (1710–1794)." In *The Descent of Human Sex Ratio at Birth*. Springer Science & Business Media, pp. 1–25.

Brooks, C., and H. Kat (2002): "The Statistical Properties of Hedge Fund Index Returns and Their Implications for Investors." *Journal of Alternative Investments*, Vol. 5, No. 2, pp. 26–44.

Cavallo, A., and R. Rigobon (2016): "The Billion Prices Project: Using Online Prices for Measurement and Research." NBER Working Paper 22111, March.

CFTC (2010): "Findings Regarding the Market Events of May 6, 2010." *Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*, September 30.

Christie, S. (2005): "Is the Sharpe Ratio Useful in Asset Allocation?" MAFC Research Paper 31. Applied Finance Centre, Macquarie University.

Clarke, Kevin A. (2005): "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science*, Vol. 22, No. 1, pp. 341–52.

Clarke, R., H. De Silva, and S. Thorley (2002): "Portfolio Constraints and the Fundamental Law of Active Management." *Financial Analysts Journal*, Vol. 58, pp. 48–66.

Cohen, L., and A. Frazzini (2008): "Economic Links and Predictable Returns." *Journal of Finance*, Vol. 63, No. 4, pp. 1977–2011.

De Miguel, V., L. Garlappi, and R. Uppal (2009): "Optimal versus Naive Diversification: How Inefficient Is the 1/N Portfolio Strategy?" *Review of Financial Studies*, Vol. 22, pp. 1915–53.

Ding, C., and X. He (2004): "K-Means Clustering via Principal Component Analysis." In *Proceedings of the 21st International Conference on Machine Learning*. Available at http://ranger.uta.edu/~chqding/papers/KmeansPCA1.pdf

Easley, D., M. López de Prado, and M. O'Hara (2011a): "Flow Toxicity and Liquidity in a High-Frequency World." *Review of Financial Studies*, Vol. 25, No. 5, pp. 1457–93.

Easley, D., M. López de Prado, and M. O'Hara (2011b): "The Microstructure of the 'Flash Crash': Flow Toxicity, Liquidity Crashes and the Probability of Informed Trading." *Journal of Portfolio Management*, Vol. 37, No. 2, pp. 118–28.

Efron, B., and T. Hastie (2016): *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. 1st ed. Cambridge University Press.

Embrechts, P., C. Klueppelberg, and T. Mikosch (2003): *Modelling Extremal Events*. 1st ed. Springer.

Goutte, C., P. Toft, E. Rostrup, F. Nielsen, and L. Hansen (1999): "On Clustering fMRI Time Series." *NeuroImage*, Vol. 9, No. 3, pp. 298–310.

Grinold, R., and R. Kahn (1999): *Active Portfolio Management*. 2nd ed. McGraw-Hill.

Gryak, J., R. Haralick, and D. Kahrobaei (Forthcoming): "Solving the Conjugacy Decision Problem via Machine Learning." *Experimental Mathematics*. Available at https://doi.org/10.1080/10586458.2018.1434704

Hacine-Gharbi, A., and P. Ravier (2018): "A Binning Formula of Bi-histogram for Joint Entropy Estimation Using Mean Square Error Minimization." *Pattern Recognition Letters*, Vol. 101, pp. 21–28.

Hacine-Gharbi, A., P. Ravier, R. Harba, and T. Mohamadi (2012): "Low Bias Histogram-Based Estimation of Mutual Information for Feature Selection." *Pattern Recognition Letters*, Vol. 33, pp. 1302–8.

Hamilton, J. (1994): *Time Series Analysis*. 1st ed. Princeton University Press.

Harvey, C., Y. Liu, and C. Zhu (2016): ". . . and the Cross-Section of Expected Returns." *Review of Financial Studies*, Vol. 29, No. 1, pp. 5–68. Available at https://ssrn.com/abstract=2249314

Hodge, V., and J. Austin (2004): "A Survey of Outlier Detection Methodologies." *Artificial Intelligence Review*, Vol. 22, No. 2, pp. 85–126.

IDC (2014): "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things." *EMC Digital Universe with Research and Analysis*. April. Available at www.emc.com/leadership/digital-universe/2014iview/index.htm

Ingersoll, J., M. Spiegel, W. Goetzmann, and I. Welch (2007): "Portfolio Performance Manipulation and Manipulation-Proof Performance Measures." *The Review of Financial Studies*, Vol. 20, No. 5, pp. 1504–46.

Jaynes, E. (2003): *Probability Theory: The Logic of Science*. 1st ed. Cambridge University Press.

Jolliffe, I. (2002): *Principal Component Analysis*. 2nd ed. Springer.

Kraskov, A., H. Stoegbauer, and P. Grassberger (2008): "Estimating Mutual Information." Working paper. Available at https://arxiv.org/abs/cond-mat/0305641v1

Laloux, L., P. Cizeau, J. P. Bouchaud, and M. Potters (2000): "Random Matrix Theory and Financial Correlations." *International Journal of Theoretical and Applied Finance*, Vol. 3, No. 3, pp. 391–97.

Ledoit, O., and M. Wolf (2004): "A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices." *Journal of Multivariate Analysis*, Vol. 88, No. 2, pp. 365–411.

Lewandowski, D., D. Kurowicka, and H. Joe (2009): "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." *Journal of Multivariate Analysis*, Vol. 100, pp. 1989–2001.

Liu, Y. (2004): "A Comparative Study on Feature Selection Methods for Drug Discovery." *Journal of Chemical Information and Modeling*, Vol. 44, No. 5, pp. 1823–28. Available at https://pubs.acs.org/doi/abs/10.1021/ci049875d

Lo, A. (2002): "The Statistics of Sharpe Ratios." *Financial Analysts Journal*, July, pp. 36–52.

Lochner, M., J. McEwen, H. Peiris, O. Lahav, and M. Winter (2016): "Photometric Supernova Classification with Machine Learning." *The Astrophysical Journal*, Vol. 225, No. 2. Available at http://iopscience.iop.org/article/10.3847/0067-0049/225/2/31/meta

López de Prado, M. (2016): "Building Diversified Portfolios that Outperform Out-of-Sample." *Journal of Portfolio Management*, Vol. 42, No. 4, pp. 59–69.

López de Prado, M. (2018a): *Advances in Financial Machine Learning*. 1st ed. Wiley.

López de Prado, M. (2018b): "The 10 Reasons Most Machine Learning Funds Fail." *The Journal of Portfolio Management*, Vol. 44, No. 6, pp. 120–33.

López de Prado, M. (2019a): "A Data Science Solution to the Multiple-Testing Crisis in Financial Research." *Journal of Financial Data Science*, Vol. 1, No. 1, pp. 99–110.

López de Prado, M. (2019b): "Beyond Econometrics: A Roadmap towards Financial Machine Learning." Working paper. Available at https://ssrn.com/abstract=3365282

López de Prado, M. (2019c): "Ten Applications of Financial Machine Learning." Working paper. Available at https://ssrn.com/abstract=3365271

López de Prado, M., and M. Lewis (2018): "Detection of False Investment Strategies Using Unsupervised Learning Methods." Working paper. Available at https://ssrn.com/abstract=3167017

Louppe, G., L. Wehenkel, A. Sutera, and P. Geurts (2013): "Understanding Variable Importances in Forests of Randomized Trees." In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pp. 431–39.

Markowitz, H. (1952): "Portfolio Selection." *Journal of Finance*, Vol. 7, pp. 77–91.

Meila, M. (2007): "Comparing Clusterings – an Information Based Distance." *Journal of Multivariate Analysis*, Vol. 98, pp. 873–95.

Mertens, E. (2002): "Variance of the IID estimator in Lo (2002)." Working paper, University of Basel.

Molnar, C. (2019): "Interpretable Machine Learning: A Guide for Making Black-Box Models Explainable." Available at https://christophm.github.io/interpretable-ml-book/

Mullainathan, S., and J. Spiess (2017): "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives*, Vol. 31, No. 2, pp. 87–106.

Neyman, J., and E. Pearson (1933): "IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society, Series A*, Vol. 231, No. 694–706, pp. 289–337.

Opdyke, J. (2007): "Comparing Sharpe Ratios: So Where Are the p-Values?" *Journal of Asset Management*, Vol. 8, No. 5, pp. 308–36.

Parzen, E. (1962): "On Estimation of a Probability Density Function and Mode." *The Annals of Mathematical Statistics*, Vol. 33, No. 3, pp. 1065–76.

Resnick, S. (1987): *Extreme Values, Regular Variation and Point Processes*. 1st ed. Springer.

Romer, P. (2016): "The Trouble with Macroeconomics." *The American Economist*, September 14.

Rosenblatt, M. (1956): "Remarks on Some Nonparametric Estimates of a Density Function." *The Annals of Mathematical Statistics*, Vol. 27, No. 3, pp. 832–37.

Rousseeuw, P. (1987): "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis." *Computational and Applied Mathematics*, Vol. 20, pp. 53–65.

Schlecht, J., M. Kaplan, K. Barnard, T. Karafet, M. Hammer, and N. Merchant (2008): "Machine-Learning Approaches for Classifying Haplogroup from Y Chromosome STR Data." *PLOS Computational Biology*, Vol. 4, No. 6. Available at https://doi.org/10.1371/journal.pcbi.1000093

Sharpe, W. (1966): "Mutual Fund Performance." *Journal of Business*, Vol. 39, No. 1, pp. 119–38.

Sharpe, W. (1975): "Adjusting for Risk in Portfolio Performance Measurement." *Journal of Portfolio Management*, Vol. 1, No. 2, pp. 29–34.

Sharpe, W. (1994): "The Sharpe Ratio." *Journal of Portfolio Management*, Vol. 21, No. 1, pp. 49–58.

Šidàk, Z. (1967): "Rectangular Confidence Regions for the Means of Multivariate Normal Distributions." *Journal of the American Statistical Association*, Vol. 62, No. 318, pp. 626–33.

Solow, R. (2010): "Building a Science of Economics for the Real World." Prepared statement of Robert Solow, Professor Emeritus, MIT, to the House Committee on Science and Technology, Subcommittee on Investigations and Oversight, July 20.

Steinbach, M., E. Levent, and V. Kumar (2004): "The Challenges of Clustering High Dimensional Data." In L. Wille (ed.), *New Directions in Statistical Physics*. 1st ed. Springer, pp. 273–309.

Štrumbelj, E., and I. Kononenko (2014): "Explaining Prediction Models and Individual Predictions with Feature Contributions." *Knowledge and Information Systems*, Vol. 41, No. 3, pp. 647–65.

Varian, H. (2014): "Big Data: New Tricks for Econometrics." *Journal of Economic Perspectives*, Vol. 28, No. 2, pp. 3–28.

Wasserstein, R., A. Schirm, and N. Lazar (2019): "Moving to a World beyond p<0.05." *The American Statistician*, Vol. 73, No. 1, pp. 1–19.

Wasserstein, R., and N. Lazar (2016): "The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician*, Vol. 70, pp. 129–33.

Witten, D., A. Shojaie, and F. Zhang (2013): "The Cluster Elastic Net for High-Dimensional Regression with Unknown Variable Grouping." *Technometrics*, Vol. 56, No. 1, pp. 112–22.

Acknowledgments

Professor Riccardo Rebonato kindly invited me to publish this Element in the series he edits for *Cambridge Elements in Quantitative Finance*. Professor Frank Fabozzi made substantive suggestions regarding the Element's content and scope. Many of the techniques introduced in this Element were tested in the course of my work at Lawrence Berkeley National Laboratory, for which I am particularly grateful to Professor Horst Simon and Dr. Kesheng Wu. Finally, I wish to recognize my approximately thirty coauthors for the past twenty years, for their enduring support and inspiration.

About the Author

Marcos M. López de Prado is a professor of practice at Cornell University's School of Engineering, and the CIO of True Positive Technologies (TPT). Dr. López de Prado has over 20 years of experience developing investment strategies with the help of machine learning algorithms and supercomputers. In 2019, he received the 'Quant of the Year Award' from *The Journal of Portfolio Management*. For more information, visit www.QuantResearch.org

Cambridge Elements

Quantitative Finance

Elements in the Series

Machine Learning for Asset Managers
Marcos M. López de Prado

A full series listing is available at: www.cambridge.org/EQF

Successful investment strategies are specific implementations of general theories. An investment strategy that lacks a theoretical justification is likely to be false. Hence, an asset manager should concentrate her efforts on developing a theory rather than on backtesting potential trading rules. The purpose of this Element is to introduce machine learning (ML) tools that can help asset managers discover economic and financial theories. ML is not a black box, and it does not necessarily overfit. ML tools complement rather than replace the classical statistical methods. Some of ML's strengths include: (1) a focus on out-of-sample predictability instead of in-sample variance adjudication; (2) the use of computational methods to avoid relying on (potentially unrealistic) assumptions; (3) the ability to "learn" complex specifications, including nonlinear, hierarchical, and noncontinuous interaction effects in a high-dimensional space; and (4) the ability to disentangle the variable search from the specification search, in a manner that is robust to multicollinearity and other substitution effects.

About the Series

Cambridge Elements in Quantitative Finance aims for broad coverage of all major topics within the field. Written at a level appropriate for advanced undergraduate or graduate students and practitioners, *Quantitative Finance* combines reports on original research covering an author's personal area of expertise, tutorials and masterclasses on emerging methodologies, and reviews of the most important literature.

Series Editor

Riccardo Rebonato
EDHEC Business School

CAMBRIDGE
UNIVERSITY PRESS
www.cambridge.org

ISBN 978-1-108-79289-9

Cover image: NASA, ESA, M. Robberto (Space Telescope Science Institute/ESA) and the Hubble Space Telescope Orion Treasury Project Team.