Can Machines "Learn" Finance?*

Ronen Israel
AQR Capital Management

Bryan Kelly
Yale University and
AQR Capital Management

Tobias Moskowitz
Yale University and
AQR Capital Management

This Version: January 10, 2020

1 Introduction

There are many potential places where machine learning can be used to improve our understanding of financial markets. Return prediction is the most important task underlying the portfolio construction problem that lies at the heart of the investing industry. Our discussion focuses on the unique challenges in applying machine learning to return prediction and aims to establish realistic expectations for how and where machine learning is and will be impactful in asset management.

We begin with an overview of machine learning, and why it has emerged as a topic of conversation (in asset management and other fields) that is distinct from statistics more broadly. Next, we characterize basic conditions in which machine learning thrives and provide examples of how these conditions were met in some famous machine learning success stories. Next, we argue that finance—and return prediction in particular—faces a challenging set of conditions that differ markedly from other domains where machine learning has excelled. We discuss why it is crucial to understand these differences in order to develop effective approaches and realistic expectations for machine learning in asset management. Finally, we outline some beneficial use cases for financial machine learning. We conclude with a view that machine learning is the most recent embodiment of the longstanding quantitative investing paradigm—the idea of using data-driven approaches to build more efficient

*We are grateful to Vasant Dhar for numerous comments that improved this draft, as well as from discussions with Jordan Brooks, Antti Ilmanen, Tal Kachman, Michael Katz, Michael Lock, Scott Richardson, and Dan Villalon.

portfolios—and argue that it is a natural evolution of quantitative tools in asset management, and not a revolutionary shift in the business model.

2 What Is Machine Learning?

"Machine learning is the study of computer algorithms that allow computer programs to automatically improve through experience." This definition from Mitchell (1997) provides an elegant and concise summary of machine learning as a broad research discipline. It captures how impactful machine learning has the potential to be in just about any application one can fathom.

Mitchell's definition also encompasses large swathes of traditional quantitative investment research (and much of the field of statistics more broadly). For investment practitioners, then, it is helpful to distinguish the finer aspects of cutting edge machine learning that are differentiated from traditional quant research, such as the ability to study thousands of predictors in a single model, or the ability to survey forecasts from a wide variety of models rather than relying on the quant researcher to settle on a single model at the outset.

To this end, Gu, Kelly, and Xiu (GKX, forthcoming) propose a more detailed working definition of machine learning in their study of financial markets. They note that,

> "The definition of 'machine learning' is inchoate and is often context specific. We use the term to describe (i) a diverse collection of high-dimensional models for statistical prediction, combined with (ii) so-called 'regularization' methods for model selection and mitigation of overfit, and (iii) efficient algorithms for searching among a vast number of potential model specifications."

The statement begins with a reminder of the conflicting perspectives—ranging from hype to skepticism (and sometimes confusion)—often heard in asset managers' discussions of machine learning. These conflicts originate from the fact that machine learning is a quickly evolving and technical field, which means that most people are still developed a basic understanding of it. This gives rise to different parties using highly variable terminology and standards when discussing machine learning, usually shaped to suit their own marketing purposes. But first and foremost, one should recognize that, in any of its incarnations, machine learning amounts to a set of procedures for estimating a statistical model and using that model to make decisions. So, at its core, machine learning *need*

not be differentiated from applied statistical analysis more generally. Most of the ideas underlying machine learning have lived comfortably under the umbrella of statistics for decades.

So why have we moved to using new terminology to describe old ideas? Above and beyond the marketing angle—machine learning is a sexy name that carries the connotation of bleeding edge Silicon Valley technology—there are at least three substantive reasons for this shift. First, the historical practical usage of statistics was frequently confined to "small" models—those with a handful of input predictor variables (or "features" in ML terminology) and simple, often linear, association rules between those inputs and the output (i.e., dependent variables) of interest. The term "machine learning" has come to serve as a shorthand to signal an explicit interest in "large" models, those with many input variables and/or those allowing for complex nonlinear associations between the inputs and output.

This idea is captured by part (i) of the GXK definition above. In order to learn through experience, the machine needs a representation of what it is trying to learn, which requires a research choice. Machine learning brings an open-mindedness for statistical representations that are highly parameterized and often nonlinear. Such models are of course not new to statistics, so it would be misleading to describe this as a contrast with "traditional" statistics. But it is fair to say that machine learning specializes in this sophisticated end of the model spectrum. Small models are rigid and oversimplified, but have the virtue that they can be used with small data sets. They are also "robust" in the sense that their behavior can be relatively insensitive to reasonable changes in the data. Large and sophisticated models are much more flexible, but can also suffer from poor out-of-sample performance when they overfit noise in the system. Researchers turn to models like these when they believe the benefits from more accurately describing the complexities of real world phenomena outweigh the costs of potential overfit. Part (i) of this definition also points out that the primary objective of machine learning is to generate accurate predictions. As emphasized by Breiman et al. (2001), its focus on maximizing prediction accuracy in the face of an unknown data model is the central differentiating feature of machine learning from the traditional statistical objective of estimating a known data generating model and conducting hypothesis tests.

Second, machine learning seeks to choose a preferred model from a "diverse collection" of candidate models. Again, this idea has a long history in statistics under the heading of "model selection" and therefore is not a new contribution of machine learning. But the process of searching through

many models to find the best performer is characteristic of essentially all machine learning methods—it is closely connected with what machine learners call model "tuning." Of course, by looking at multiple models and selecting the top performers in-sample mechanically leads to overfit and poor out-of-sample performance. Because of this, the model search process is always accompanied by so-called "regularization" techniques and methods for identifying models that are likely to perform best out-of-sample. Regularization is a blanket term for constraining the size of a model. An optimal model is a "Goldilocks" model. It is large enough so that it can reliably identify the true and potentially complex predictive relationships in the data, but not so flexible that it overfits and suffers out-of-sample. Regularization methods encourage smaller models, and make sure that a richer model only gets selected if it is likely to give a genuine boost to out-of-sample prediction accuracy. A cornerstone method in the model selection process is cross-validation, in which the researcher simulates out-of-sample tests in historical data and picks models that would have performed best in these "as-if" out-of-sample scenarios. Element (ii) of our machine learning definition describes refinements in implementation that emphasize reliable out-of-sample performance in order to explicitly guard against overfit.

Third, and perhaps the clearest differentiator of machine learning from traditional statistics, are its innovative approaches to model optimization. In truly big data environments, the computational demands of model estimation can be exorbitant. To ease this burden, machine learning has developed a variety of approximate and computationally efficient optimization routines. For example, model estimation traditionally uses all data points in every step of an iterative optimization routine. In big data environments, it is usually overkill to use the full data set in optimization, and optimizers can be dramatically accelerated by instead using random subset of the data with little loss of accuracy. This idea is the foundation of the "stochastic gradient descent (SGD)" optimization routine that is a staple in machine learning implementation. Element (iii) describes innovative machine learning solutions such as SGD and early stopping that are designed to approximate an optimal specification with large reductions in computational cost.

Lastly, the differentiation between machine learning and statistics is based in large part on how ubiquitous machine learning has become in a wide range of commercial problems. Traditionally, statistics showed up in business processes primarily in the form of testing—such as determining whether failure rates of an engine part exceeded a given threshold or evaluating efficacy of a new

drug. Machine learning focuses on maximizing predictive accuracy, and as such it finds natural uses in every commercial application one can imagine. Much of the commercial success of machine learning has emerged from recognizing that many pre-conceived approaches to problem solving (largely based on deterministic computer algorithms) can be improved with algorithms that adapt to feedback in the form of data. That is, giving statistics and prediction a more direct role in commercial processes leads to better products and services.

3 Where Has Machine Learning Worked?

The gamut of famous machine learning success stories—problems like image and voice recognition, strategic gaming, autonomous vehicles, and robotics—occur in environments with two critical conditions in common. Each is a truly big data environment, and each is a high signal-to-noise ratio environment. Understanding these conditions, and how they differ in many financial applications, is key to developing a foundational understanding of financial machine learning.

Machine learning thrives in data rich environments. Models like neural networks are valuable for describing complex predictive associations because they have the flexibility to match complicated patterns. This flexibility comes from rich parameterizations. For example, the famous image-recognizing neural network model, "AlexNet" (Krizhevsky et al., 2012), has roughly 61 million parameters. All hope of estimating such a heavily parameterized model lies in having a truly massive amount of training data (not to mention exorbitant computing power). And many big data environments benefit from the ability of the researcher to generate new data through experiments. For example, if you haven't succeeded yet in training your autonomous vehicle, then drive the car another 100,000 miles.

The second, and perhaps more subtle, feature of most machine learning success stories can be understood in terms of their "signal-to-noise ratios." A signal-to-noise ratio describes how much predictability exists within a system. Some systems are by nature very predictable—their signal-to-ratio is high. Others are inherently dominated by randomness and thereby have low signal-to-noise ratios. Take, for example, image recognition. If handed a thousand Instagram photos, you will correctly identify those that contain cats with a success rate of almost 100%. It's generally easy for a human to distinguish signal (the cat) from noise (blur, background images, etc.).

Machine learning thrives in data rich environments with strong signals and little noise.

4 Finance Is Different

As the popular press so frequently reminds us, machine learning can accomplish the once unthinkable. It recognizes images and speech, drive cars, and beats grandmasters at complex games of strategy. This is where the excitement, hype, and extrapolation kicks in for finance. Because machine learning has done so many amazing things, it may seem a foregone conclusion that it will dominate at financial tasks like stock picking.

In order to develop realistic expectations about the benefits of machine learning for asset management, we must understand what makes finance different.

Small Data

The core task in asset management—return prediction—is a *small* data problem. This may sound surprising given the constant marketing barrage of "big data in finance." But the marketing betrays a misunderstanding of what big data and machine learning are about.

$$y_t = \sum_{i=1}^{N} \beta_i x_{i,t-1}, \quad t = 1, ..., T. \tag{1}$$

To fix ideas, consider a regression context like equation (1) where on the left hand side are the future returns that we want to predict (y_t) and on the right hand side are the predictor variables ($x_{1,t-1}, ..., x_{N,t-1}$). The right way to think about whether you have big data or small data is to ask "how rich of a model can my data accommodate?" The answer to this question is determined first and foremost by the number of independent observations you have for your left hand side variable, denoted by T. So, the key question in the asset management context is "how many independent return observations do I have."

However, most of the discussion around so-called big data in finance is about the number of predictor variables, N. Asset managers will talk about introducing big data sets like news text, satellite images, web traffic, and geolocation to the return prediction problem. But those are *right-hand-side* variables. The richness of a model is constrained not by the number of regressors one can conjure, but by the number of left-hand-side observations one can *learn* from. If I only have

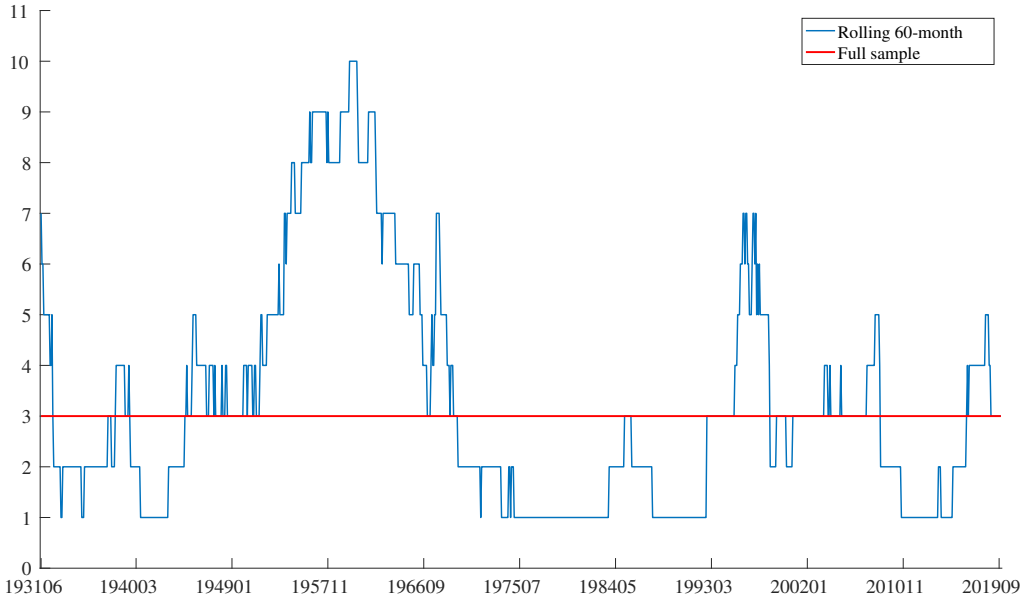
a hundred observations for y, then it doesn't matter if my predictor variables number $N = 10^3$ or $N = 10^{10}$. In either case, the number of parameters I can estimate is effectively capped by the number of y observations. Without sufficient y observations, models are constrained to be small.

When most people say big data in finance, they mean they have many variables to predict returns—they mean large N. But the number of predictors has never been a limiting factor in quant analysis. It is not the number of regressors that tells you if you have big data, it is the number of parameters that you can reliably estimate—the richness of your model—and that is determined predominantly by the number of observations, T.

So, with that in place, it is straightforward to answer the question "is return prediction a big data problem?" First, we have to decide on the frequency of returns we are counting—are they annual return, daily, tick data, etc. Clearly the appropriate frequency of returns for prediction and portfolio analysis should be based on the frequency with which the investor can rebalance with reasonably low trading costs. This will vary by investor and by asset class. In economics, it is common to imagine a so-called "representative" investor, which is the AUM-weighted average of all investors in the economy. You might imagine this as a large pension fund crossed with a high net worth individual. For large AUM investors like this, the monthly frequency is a sensible starting point to think about rebalancing because anything more frequent, even weekly let alone intra-day, quickly becomes prohibitive as trading costs rack up, leading to a large gap between what a statistical prediction model says and what can actually be implemented in practice.

Now, focusing on monthly returns, how much data do we have? If you are trading a macro strategy, using instruments things like currencies, government bonds, and commodity futures, you have a few decades of data, or a few hundred observations per asset. This is "tiny" data as it can only support a model with a handful of parameters if one hopes to achieve a degree of model stability. What about cross-sectional asset classes, like single-name equities or corporate bonds. Here the situation is brighter, as we can have at any time up to a few thousand assets trading, each with anywhere from a few years to a few decades of data, totaling to a few hundred thousand observations. While this is more than in the macro asset case, it is still small by any machine learning standard, especially once one considers that the *effective* number observations is smaller due to significant cross-sectional correlation in returns. For example, there is a 71% average pairwise correlation among the

Exhibit 1: Number of PC's Explaining 80% of Equity Return Variation



Note. The figure shows the number of PCs necessary to explain at least 80% of the covariance among monthly returns of the Fama-French 100 size and value portfolios over the 1927–2019 sample. The red line shows the result for the full sample and the blue line shows the result in rolling 60-month subsamples.

widely studied Fama-French 100 size and value portfolios.[1] Exhibit 2 shows that over the 1927–2019 sample, just three principal components (PCs) are necessary to explain at least 80% of the covariance among monthly returns of the 100 portfolios. Depending on the sub-sample, it is often the case that a single PC captures 80% of the common variation in returns, and it never take more than 10 PCs to achieve this threshold. Evidently, the number of effective cross section observations is vastly overstated by looking at asset count alone without regard to asset correlation.

In return modeling, there is only one way to expand the size of your data set—wait for time to pass. That is, a key distinguishing feature of finance versus many other machine learning domains is that we cannot generate data through experimentation. In 100 years, and regardless of the amount of technological progress in that time, return prediction will *still* be a small data problem. In this way, asset allocation significantly contrasts with the environments where machine learning has famously succeeded.

It is worth noting that investors are heterogeneous, and some can rebalance more frequently than others. High-frequency trading (HFT) firms, for example, routinely trade at millisecond frequencies.

[1] Data from https://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

The concept of the representative investor is valuable because it leads us to focus on what *most* investors can accomplish, and brings to the fore the idea of trading strategy *capacity*. HFT has nearly instantaneous turnover and which translates into vastly more return data than in the monthly problem; as a result, HFT can use far more highly parameterized models than low frequency traders. But the scale of these trades is forced to be small, and this limits who can participate. If the representative investor wanted in on HFT, the turnover of her huge AUM at high frequencies would magnify the scale of trades, leading to large price impacts that compress any HFT alpha, cannibalizing the profits that might be achievable by a small trader.

Dhar (2015) further discusses the critical role of data size in determining the impact of machine learning in finance. He provides a thoughtful discussion of machine versus human strengths in decision making in the asset management context, pointing out that "As a guiding principle, however, a robot should be considered seriously in situations where there is sufficient data from which it can learn."

Low Signal-to-noise Ratios

The second major difference between return prediction and many other machine learning problems is that the signal-to-noise ratio in returns is weak (and is constantly being pulled toward zero). One of the most important reasons for this is that economic growth, and hence financial market behavior, is difficult to predict. The best stock or investment portfolio in the world will, on any given day, quarter, or year, experience drastic swings in performance due to unanticipated news.[2] Second, the predictable signal in financial markets—often called the risk premium or the expected return in excess of cash—is small.

The low signal-to-noise is not some unfortunate coincidence of markets. On the contrary, it is a feature ensured, and constantly reinforced, by simple economic forces of profit maximization and competition. Traders with information that reliably predicts, say, a future rise in prices, don't sit passively on that information. Instead they start trading. The very act of exploiting their predictive information pushes up prices, and thereby sucks some of the predictability out of the market. And they don't stop after prices have risen just a little. They continue buying until they have exhausted their information—until prices adjust to the level that their information predicts. By leveraging

[2]To give a sense of the volatility in markets, a single stock on average will have an expected return around 5% per year above cash and a volatility of returns of nearly 40% per year. That is, its volatility is eight times larger than the expected excess return. Even at the market portfolio level, volatility is nearly four times as large as the expected return.

Exhibit 2: The Tradeoff of Signal Strength and Observation Count

Horizon	Coeff.	S.E.	R^2	N
1	0.02	0.00	0.5%	1115
12	0.25	0.05	7.7%	93
36	0.75	0.13	28.3%	31
60	0.96	0.17	29.6%	19

Note. In-sample CRSP value-weighted log market return predictive regressions using CAPE (scaled by 100) over forecast horizons of one month to five years from 1926–2018. CAPE data from http://www.econ.yale.edu/~shiller/data.htm.

information for profit-oriented trading, informed investors leave minimal predictability on the table. With the predictability already priced in, the only thing that moves markets is unanticipated news—shocks to the system—i.e., noise. This idea, that competition in markets wipes out return predictability, is not new. It is the very idea underpinning the Nobel prize-winning work on the efficient markets hypothesis (Fama, 1970).

If well-functioning markets lack predictability, why bother looking for signal at all? In an efficient market, returns need not be *entirely* devoid of predictability. Investors may stop short of using their full information if, for example, it requires taking on too much risk, if they possess behavioral biases, if they face transaction costs, or if they are subject to legal or regulatory restrictions (such as insider trading rules). The remaining predictability should be small and difficult to capture, as any easy profits will be quickly seized by competitive traders, particularly over short horizons. An interesting feature of risk-based return predictability is that it tends to become stronger with forecast horizon, a point emphasized by Cochrane (2009). Exhibit 2 shows how stock predictability varies across horizons from one month to five years ahead. For simplicity, we illustrate this point in a linear regression with the Shiller (2015) cyclically-adjusted price-earnings ratios (CAPE). On one hand, machine learning can benefit from the higher signal-to-noise ratios at longer horizons (the R^2 reaches 29.6% at five years). But then we are faced with the fact that long-horizon forecasts take a fixed time series sample and cut it into fewer and fewer observations, driving up standard errors and prediction error variance along with it. A five year forecasting model only has 19 distinct return observations to learn from! A central task facing financial machine learners researchers is to identify the most profitable point along this tantalizing signal strength and observation count tradeoff.

In addition to small data and low signal-to-noise ratios, asset management possess a variety of other challenges that pose a problem to adaptation of machine learning methods. We discuss three

of these below.

Evolving Markets

The machine learning challenges posed by low signal-to-noise ratios are further confounded by the adaptive nature and dynamic character of markets. If a researcher identifies a new signal that captures a particular form of asset mispricing useful for predicting prices, then as the signal becomes more widely known, more traders act on it, and prices correct more quickly. The market eventually absorbs that information and the data generating process changes due to the very actions of agents in the market. Likewise, technological innovations can alter the structure of the economy and reshape the way humans interact with markets. While the frontiers of machine learning have developed some tools that may help with such adaptive phenomena (such as the online learning algorithms in Arora et al., 2012; Li and Hoi, 2014), it highlights the fact that finance is more complex than many other domains of ML research (cats don't begin morphing into dogs once the algorithm becomes good at cat recognition).

Unstructured Data

The traditional inputs to quantitative asset management are the kinds of well-structured data sets that reside in an Excel spreadsheet. Columns are predictor variables, rows are repeat observations—these are the types of data that lend themselves easily to statistical analysis. In contrast, many interesting new data sources are best characterized as "unstructured" data. They include text data such as news articles and Tweets, image data such as Instagram posts or Youtube videos, and even some forms of market data such as detailed limit-order books. The finance industry often refers to such unstructured data as alternative, or "alt," data.

For most alt data sets, the data history is short. For example, social media outlets you may have less than a decade of data to work with, while some sources of web traffic data or geolocation data are available at most for a few years. The limited time series presents a challenge for reliable backtesting. With a short history it's hard to form a precise estimate of strategy performance which ultimately means that even very strong signals might prudently receive only small weights in a portfolio.

Need for Interpretability

Some machine learning models are proverbial black boxes, and it can be extremely challenging to draw meaningful interpretations of underlying mechanisms from machine learning models (see, e.g., Ghorbani et al., 2019). Yet the ability to understand the inner workings of one's model is a basic requirement in most asset management applications. While any asset manager prefers a more predictive model, all else equal, they can be averse to using historically reliable models that they cannot interpret. This fact—that all else is not always equal—is another version of the risk-return tradeoff that any investor solves. So while asset managers prefer a model with more predictability to less, their fiduciary duty of understanding and communicating the risks in their clients' portfolios leads them to also prefer more interpretable models. In the end, choosing a point on the predictability/interpretability frontier is a ultimately a business decision of the asset manager.

Finance is not alone in its need for interpretable models. Doctors seek to understand the drivers of machine learning medical diagnoses to avoid adverse unintended consequences of relying on algorithms (Cabitza et al., 2017) and governments and regulators remain vigilant against implicit or explicit biases in policy (such as lending decisions of financial institutions Hardt et al., 2016). This broad demand has made interpretability a priority in machine learning research (Doshi-Velez and Kim, 2017; Vellido et al., 2012). Machine learning need not be an opaque black box. First, researchers are making progress in improving how humans interpret machine learning models (Zhang et al., 2018; Horel and Giesecke, 2019). Second, and perhaps more interestingly, structural modeling approaches can embed machine learning techniques, which makes efficient use of the data and enhances discovery, within an overarching theoretical model that provides interpretation and intuition (an idea we discuss in more detail below). There are many interesting potential research avenues for drawing more meaningful and intuitive conclusions from financial machine learning models.

5 The Research Frontier

Because of these critical differences between finance and other fields where machine learning thrives, the answer to the question "Can machines learn finance?" is by no means obvious. As an industry, the understanding of how impactful machine learning will be for asset management is only just emerging. This is exactly why new research in this area is so valuable. With so much at stake, and

the best path forward is digging in and conducting diligent research.

Analysis, Not Anecdotes

When people discuss machine learning in finance, the conversation is predominantly anecdotal—"I heard a story about how manager XYZ does it." Methodical research into the benefits of machine learning for asset management is in its infancy. But early research, such as Dhar et al. (2000) and Dhar (2011), tells a hopeful story.

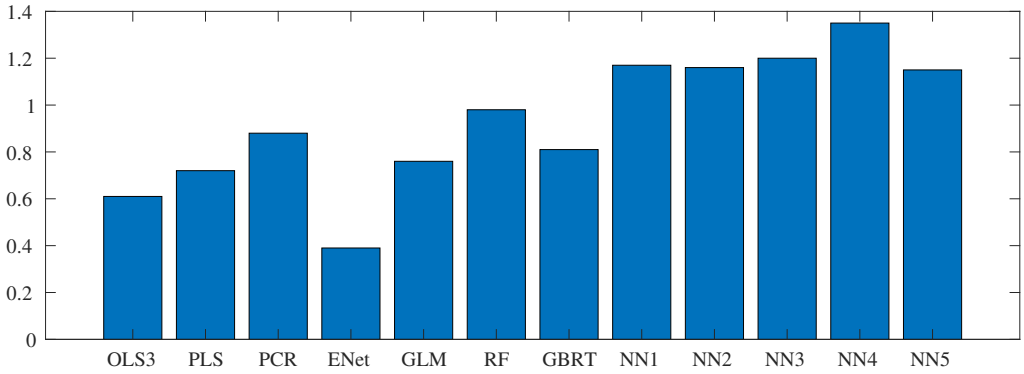
Gu, Kelly, and Xiu (forthcoming) suggest that machine learning methods can deliver significant out-of-sample improvement in the performance of stock selection strategies (see Exhibit 3). They also offer some new insights into the nature of its outperformance. For example, improvements arise most prominently among the more sophisticated models (trees and neural networks), and are due in large part to allowance of non-linear predictor interactions that are missed by simpler methods.

On balance, Gu, Kelly, and Xiu (forthcoming) find portfolio gains from using machine learning methods versus a simple benchmark. They are economically and statistically significant, but the gains are by no means revolutionary. First, note that their benchmark is an OLS regression with only three predictors: size, value, and momentum (OLS3). This is likely an overly simplistic benchmark, as even unsophisticated asset managers tend to use more advanced models than this and are even likely to use some basic trappings of machine learning. Second, while performance of all strategies is measured gross of trading costs, it does not mean the comparison is exactly apples-to-apples. The more sophisticated a model, the more likely it is to detect incremental predictability in regions of the data that are the most expensive to trade, which is exactly why the predictability was left on the table in the first place. Both of the considerations further narrow the gap between performance of machine learning-based strategies and the benchmark.

Combining Economic Theory and Machine Learning

A basic principle of statistical analysis is that theory and model parameters are substitutes. The more structure you can impose in your model, the fewer parameters you need to estimate and the more efficiently your model can use observations at its disposal to cut through noise. That is, models are helpful because they filter out noise (a map of New York is helpful for navigating the city in part because of all of the detail it *omits*). But an over-simplified model can also filter out some signal too,

Exhibit 3: Sharpe Ratio Comparison of Machine Learning Strategies



Note. Out-of-sample annualized Sharpe ratios for value-weighted decile spread long-short portfolios based on predictions from 12 machine learning models including a three-predictor OLS model (OLS3), partial least squares (PLS), principal components regression (PCR), elastic net (ENet), generalized linear model (GLM), random forest (RF), gradient-boosted regression trees (GBRT), and feedforward neural networks with one to five hidden layers (NN1–NN5). Source: Gu, Kelly, and Xiu (forthcoming).

so in a data-rich and high signal-to-noise environment, you would not want to use an unnecessarily small model. Simplicity, however, can be a virtue when signal-to-noise is low, where the benefit of filtering out noise might outweigh the cost of missing some signal.

In asset management, one can begin to tackle the low signal-to-noise problem by bringing economic theory to describe some aspects of the data, complemented by machine learning tools to capture aspects of the data for which theory is silent. A revealing example of this is the method of instrumented principal components analysis (IPCA) proposed by Kelly et al. (2017), Kelly et al. (2018), and its extension to instrumented autoencoder models by Gu, Kelly, and Xiu (2019). They begin from a simple economic structure—a low dimensional factor model for returns. In contrast, a fully unstructured machine learning analysis would be agnostic of the factor structure and would instead rely on the statistical logic that, with enough data, non-parametric methods will eventually detect the factor structure if it's in the data.

In the context of return modeling, an unstructured approach would be a severely inefficient use of data because, as a profession, we have high conviction that returns follow a factor structure. A common feature of essentially every economic theory is that returns obey a factor structure—there are a few sources of common risk that drive returns of essentially all assets and the rest of the variation in individual asset prices is idiosyncratic noise. Furthermore, a long history of empirical

research in finance has demonstrated that a few large eigenvectors dominate the covariance matrix of all returns in the economy. Using non-parametric methods to learn that there is a factor structure would waste data; you don't want to "spend" your observations re-learning something you already knew.

Where economic theories differ is in the nature or definition of the common factors. They range from theories in which the factors are macroeconomic aggregates related to consumption growth and real investment to those based on volatility shocks or behavioral biases. The idea proposed in Kelly et al. (2017), Kelly et al. (2018), and Gu, Kelly, and Xiu (2019) is to impose the general factor model mold, while embedding machine learning in specific parts of that mold in order to learn what the true factors are. They use a large collection of stock-specific information to learn the latent common factors and each stock's loadings on those factors—exactly those parts of the factor model that are not pinned down by theory. They also do this while imposing a no-arbitrage restriction, which is another theory-based disciplining device that reduces model parameterization to more efficiently use available data.

These papers also highlight that machine learning is not all about alpha. This is important because most discussions, and certainly most anecdotes, of machine learning applied to finance focus on the creation of alpha. Using new data and machine learning to build alpha (i.e., to find new, unique sources of return predictability) heads straight into the most competitive aspect of financial markets. As more investors enter the market with similar data and similar tools, the mispricing corrects and that alpha compresses.

Kelly et al. (2017) and Gu, Kelly, and Xiu (2019) emphasize that a promising area of asset management research uses machine learning to improve factor investing—that is, more efficiently establishing risk exposures and earning compensation for bearing that risk. A central tenet of asset pricing theory is that risk and return are inextricably linked in equilibrium. By taking this economic restriction seriously, these papers build superior return forecasts by more accurately forecasting *risk*—in particular, covariances between stocks and common factors. Exhibit 4 presents evidence that return forecasts based on the IPCA model of Kelly et al. (2017) and the conditional autoencoder (CA_2) of Gu, Kelly, and Xiu (2019) translate into substantial improvements in out-of-sample portfolio Sharpe ratio compared to benchmark Fama-French models.

Their methods estimate statistically optimal factors which improve factor portfolio performance

Exhibit 4: Long-short Portfolio Sharpe Ratios Based on Factor Model Return Forecasts

	Number of Factors					
Model	1	2	3	4	5	6
Fama-French	-0.82	-1.13	-0.69	-0.60	0.18	-0.53
IPCA	-0.15	-0.07	0.59	0.81	1.05	0.96
CA_2	-0.03	0.08	0.92	1.39	1.45	1.53

Note: Annualized out-of-sample Sharpe ratios for portfolios that are long/short the top/bottom decile of stocks based on return forecasts from the Fama-French model with one to six factors (including momentum), the IPCA model of Kelly et al. (2017), and the two-layer conditional autoencoder model (CA_2) of Gu, Kelly, and Xiu (2019). Stocks are value-weighted within each decile. Source: Gu, Kelly, and Xiu (2019).

by reducing tracking error relative to the true, unobservable risk factor—ultimately providing a cleaner means of harvesting the risk premium. This contrasts with more traditional research factors which tend to have ad hoc and inefficient constructions.[3] Like finding alpha, optimizing factors with machine learning can significantly boost investment portfolio performance. But unlike alpha, true factor premia are underpinned by risk and do not tend to decay as more investors enter.

Beyond Return Prediction

While we emphasize that return prediction (because of its small data and low signal-to-noise ratios) poses a particularly difficult challenge for machine learning, it is important to recognize that other critical finance problems are less subject to these limitations and in turn can benefit more from machine learning. Important examples include portfolio implementation problems such as risk management and transaction cost management.

For example, the data underlying trading cost models are based on transactions, which are available is abundance. For example, the trading cost analysis of (Frazzini et al., 2018) studies a "live execution database contains 11,044,700 parent orders, 4,368,100,000 child orders, and 691,600,000 executions across 9,543 stocks globally between August 1998 and June 2016, totaling US$1,701,390,000,000 in trades." A data set of this size can support model richness that wouldn't be dreamed of in a model of monthly returns.

As another example, a long literature exemplified by the Nobel prize-winning work of Engle (1982) demonstrates that financial market risks possess a high degree of predictability. That is, risk prediction benefits from a comparatively high signal-to-noise ratio. When the signal-to-noise ratio

[3]Take for example the Fama-French value factor, HML. This factor buys 30% of stocks with the highest B/M ratios, sells 30% with low B/M, and ignores the rest.

is high, it takes fewer observations to zero in on the ground truth predictive relationship. Indeed, risk modeling has been among the finance applications with the fastest uptake of machine learning methods (early examples include Donaldson and Kamstra, 1997; Aït-Sahalia and Lo, 1998)

6 Conclusion: An Evolution, Not a Revolution

Financial machine learning has the potential to be an important step forward in quantitative investing. Two key points are crucial for understanding the current state of machine learning in the practice of asset management. The first is that research is just taking off and many important questions are yet to be answered. The second is that early research evidence indicates economically and statistically significant improvements in the performance of portfolios that leverage machine learning tools. However, the gains are evolutionary, not revolutionary.

The ideas behind machine learning—leveraging new data sets to identify robust additive portfolio performance and using quantitative methods to extract information systematically—are the modus operandi of quantitative investment processes. For decades, asset managers have used human-intensive, decentralized statistical learning. Machine learning offers a systematic approach to investing that mechanizes that process, allows managers to metabolize information from more new sources faster, including unstructured data previously untapped, and provides tools to search through increasingly flexible economic models that better capture complex realities of financial markets.

The evolution of asset management by incorporating machine learning is already underway, but the industry's collective machine learning marketing hype must be tempered. Asset management, and return prediction in particular, is a small data science with low signal-to-noise ratios, making it very different from disciplines where machine learning has thrived. As a result, adapting machine learning in finance is a more difficult proposition than many commentators appreciate. Attempts to model returns with the extremely rich and flexible statistical models employed in other domains are typically doomed from the outset because of the lack of data to support such models. The benefits of machine learning for return prediction are most likely to come from moving to modestly flexible nonlinear models that are buttressed by structural assumptions from economic theory and human expertise. Other beneficial uses in asset management includes machine learning approaches for better "crafting" portfolios once return predictions have been made, for example by improving models of

risk and trading cost management.

References

Aït-Sahalia, Yacine, and Andrew W Lo, 1998, Nonparametric estimation of state-price densities implicit in financial asset prices, *The Journal of Finance* 53, 499–547.

Arora, Sanjeev, Elad Hazan, and Satyen Kale, 2012, The multiplicative weights update method: a meta-algorithm and applications, *Theory of Computing* 8, 121–164.

Breiman, Leo, et al., 2001, Statistical modeling: The two cultures (with comments and a rejoinder by the author), *Statistical science* 16, 199–231.

Cabitza, Federico, Raffaele Rasoini, and Gian Franco Gensini, 2017, Unintended consequences of machine learning in medicine, *Jama* 318, 517–518.

Cochrane, John H, 2009, *Asset pricing: Revised edition* (Princeton university press).

Dhar, V, 2015, Should you trust your money to a robot?, *Big data* 3, 55.

Dhar, Vasant, 2011, Prediction in financial markets: The case for small disjuncts, *ACM Transactions on Intelligent Systems and Technology (TIST)* 2, 19.

Dhar, Vasant, Dashin Chou, and Foster Provost, 2000, Discovering interesting patterns for investment decision making with glower: A genetic learner overlaid with entropy reduction, *Data Mining and Knowledge Discovery* 4, 251–280.

Donaldson, R Glen, and Mark Kamstra, 1997, An artificial neural network-garch model for international stock return volatility, *Journal of Empirical Finance* 4, 17–46.

Doshi-Velez, Finale, and Been Kim, 2017, Towards a rigorous science of interpretable machine learning, *arXiv preprint arXiv:1702.08608* .

Engle, Robert F, 1982, Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation, *Econometrica: Journal of the Econometric Society* 987–1007.

Fama, Eugene F, 1970, Efficient capital markets: A review of theory and empirical work, *The Journal of Finance* 25, 383–417.

Frazzini, Andrea, Ronen Israel, and Tobias J Moskowitz, 2018, Trading costs, *Yale University Working Paper* .

Ghorbani, Amirata, Abubakar Abid, and James Zou, 2019, Interpretation of neural networks is fragile, in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3681–3688.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu, forthcoming, Empirical asset pricing via machine learning, Technical report.

Gu, Shihao, Bryan T Kelly, and Dacheng Xiu, 2019, Autoencoder asset pricing models, *Available at SSRN* .

Hardt, Moritz, Eric Price, Nati Srebro, et al., 2016, Equality of opportunity in supervised learning, in *Advances in neural information processing systems*, 3315–3323.

Horel, Enguerrand, and Kay Giesecke, 2019, Towards explainable ai: Significance tests for neural networks, *arXiv preprint arXiv:1902.06021* .

Kelly, Bryan, Seth Pruitt, and Yinan Su, 2018, Characteristics are covariances: A unified model of risk and return, Technical report, National Bureau of Economic Research.

Kelly, Bryan T, Seth Pruitt, and Yinan Su, 2017, Instrumented principal component analysis, *Yale University Working Paper* .

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton, 2012, Imagenet classification with deep convolutional neural networks, in *Advances in neural information processing systems*, 1097–1105.

Li, Bin, and Steven CH Hoi, 2014, Online portfolio selection: A survey 46, 35.

Mitchell, T.M., 1997, *Machine Learning*, McGraw-Hill International Editions (McGraw-Hill).

Shiller, Robert J, 2015, *Irrational exuberance: Revised and expanded third edition* (Princeton university press).

Vellido, Alfredo, José David Martín-Guerrero, and Paulo JG Lisboa, 2012, Making machine learning models interpretable., in *ESANN*, volume 12, 163–172, Citeseer.

Zhang, Quanshi, Ying Nian Wu, and Song-Chun Zhu, 2018, Interpretable convolutional neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8827–8836.