A Backtesting Protocol in the Era of Machine Learning

Rob Arnott, Campbell R. Harvey, and Harry Markowitz

Rob Arnott
is chairman and founder of Research Affiliates, LLC, in Newport Beach, CA.
arnott@rallc.com

Campbell R. Harvey
is a professor of finance at Duke University in Durham, NC, and a partner and senior advisor at Research Affiliates, LLC, in Newport Beach, CA.
cam.harvey@duke.edu

Harry Markowitz
is founder of Harry Markowitz Company in San Diego, CA.
harryhmm@aol.com

Data mining is the search for replicable patterns, typically in large sets of data, from which we can derive benefit. In empirical finance, data mining has a pejorative connotation. We prefer to view data mining as an unavoidable element of research in finance. We are all data miners, even if only by living through a particular history that shapes our beliefs. In the past, data collection was costly, and computing resources were limited. As a result, researchers had to focus their efforts on the hypotheses that made the most sense. Today, both data and computing resources are cheap, and in the era of machine learning, researchers no longer even need to specify a hypothesis—the algorithm will supposedly figure it out.

Researchers are fortunate today to have a variety of statistical tools available, among which machine learning, and the array of techniques it represents, is a prominent and valuable one. Indeed, machine learning has already advanced our knowledge in the physical and biological sciences and has also been successfully applied to the analysis of consumer behavior. All of these applications benefit from a vast amount of data. With large data, patterns will emerge purely by chance. One of the big advantages of machine learning is that it is hardwired to try to avoid overfitting by constantly cross-validating discovered patterns. Again, this advantage serves well in the presence of a large amount of data.

In investment finance, apart from tick data, the data are much more limited in scope. Indeed, most equity-based strategies that purport to provide excess returns to a passive benchmark rely on monthly and quarterly data. In this case, cross-validation does not alleviate the curse of dimensionality. As a noted researcher remarked to one of us:

> [T]uning 10 different hyperparameters using k-fold cross-validation is a terrible idea if you are trying to predict returns with 50 years of data (it might be okay if you had millions of years of data). It is always necessary to impose structure, perhaps arbitrary structure, on the problem you are trying to solve.

Machine learning and other statistical tools, which have been impractical to use in the past, hold considerable promise for the development of successful trading strategies, especially in higher-frequency trading. They might also hold great promise in other applications, such as risk management. Nevertheless, we need to be careful in applying these tools. Indeed, we argue that given the limited nature of the standard data that we use in finance, many of the challenges we face in the era of machine learning are very similar

to the issues we have long faced in quantitative finance in general. We want to avoid backtest overfitting of investment strategies, and we want a robust environment to maximize the discovery of new (true) strategies.

We believe the time is right to take a step back and to re-examine how we do our research. Many have warned about the dangers of data mining in the past (e.g., Leamer 1978; Lo and MacKinlay 1990; and Markowitz and Xu 1994), but the problem is even more acute today. The playing field has leveled in computing resources, data, and statistical expertise. As a result, new ideas run the risk of becoming very crowded, very quickly. Indeed, the mere publishing of an anomaly may well begin the process of arbitraging the opportunity away.

Our article develops a protocol for empirical research in finance. Research protocols are popular in other sciences and are designed to minimize obvious errors, which might lead to false discoveries. Our protocol applies to both traditional statistical methods and modern machine learning methods.

HOW DID WE GET HERE?

The early days of quantitative investing brought many impressive successes. Severe constraints on computing and data led research to be narrowly focused. In addition, much of the client marketplace was skeptical of quantitative methods. Consequently, given the limited capital deployed on certain strategies, the risk of crowding was minimal. Today, however, the playing field has changed. Now almost everyone deploys quantitative methods—even discretionary managers—and clients are far less averse to quantitative techniques.

The pace of transformation is striking. Consider the Cray 2, the fastest supercomputer in the world in the late 1980s and early 1990s (Bookman 2017). It weighed 5,500 pounds and, adjusted for inflation, cost over US$30 million in 2019 dollars. The Cray 2 performed an extraordinary (at the time) 1.9 billion operations per second (Anthony 2012). Today's iPhone Xs is capable of 5 trillion operations per second and weighs just six ounces. Whereas a gigabyte of storage cost $10,000 in 1990, it costs only about a penny today. Furthermore, a surprising array of data and application software is available for free, or very nearly free. The barriers to entry in the data-mining business, once lofty, are now negligible.

Sheer computing power and vast data are only part of the story. We have witnessed many advances in statistics, mathematics, and computer science, notably in the fields of machine learning and artificial intelligence. In addition, the availability of open-source software has also changed the game: It is no longer necessary to invest in (or create) costly software. Essentially, anyone can download software and data and potentially access massive cloud computing to join the data-mining game.

Given the low cost of entering the data-mining business, investors need to be wary. Consider the long–short equity strategy whose results are illustrated in Exhibit 1. This is not a fake exhibit.[1] It represents a market-neutral strategy developed on NYSE stocks from 1963 to 1988, then validated out of sample with even stronger results over the years 1989 through 2015. The Sharpe ratio is impressive—over a 50-year span, far longer than most backtests—and the performance is both economically meaningful, generating nearly 6% alpha a year, and statistically significant.

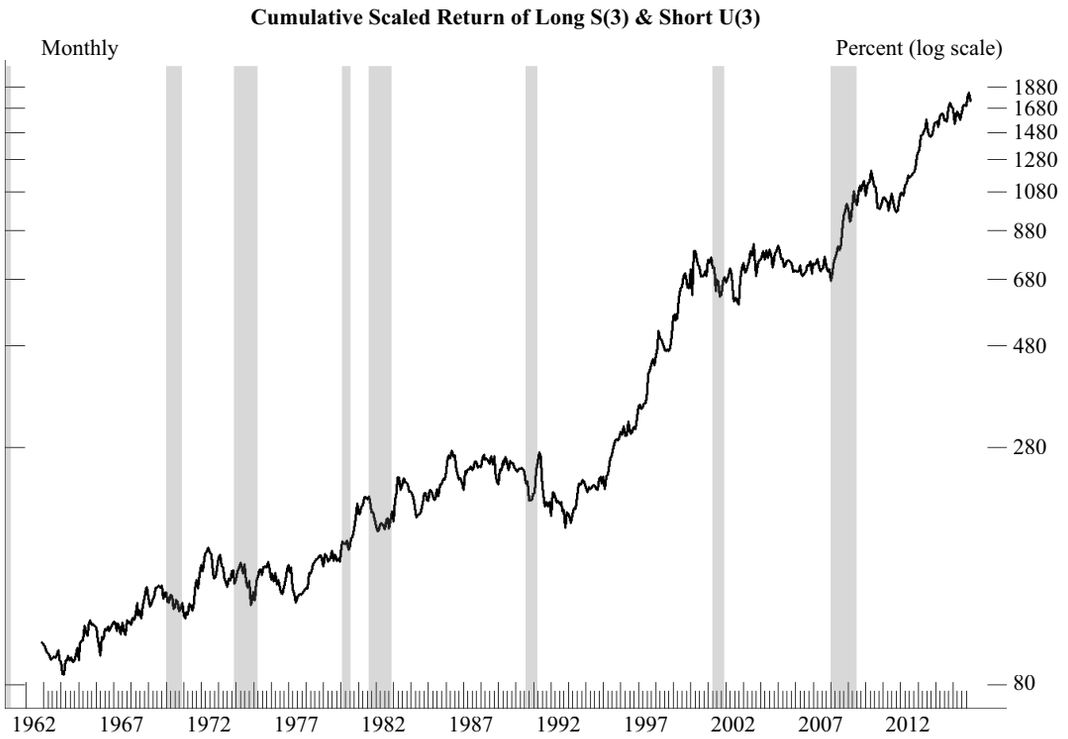
Better still, the strategy has five very attractive practical features. First, it relies on a consistent methodology through time. Second, performance in the most recent period does not trail off, indicating that the strategy is not crowded. Third, the strategy does well during the financial crisis, gaining nearly 50%. Fourth, the strategy has no statistically significant correlations with any of the well-known factors, such as value, size, and momentum, or with the market as a whole. Fifth, the turnover of the strategy is extremely low, less than 10% a year, so the trading costs should be negligible.

This strategy might seem too good to be true. And it is. This data-mined strategy forms portfolios based on letters in a company's ticker symbol. For example, A(1)–B(1) goes long all stocks with "A" as the first letter of their ticker symbol and short all stocks with "B" as the first letter, equally weighting in both portfolios. The strategy in Exhibit 1 considers all combinations of the first three letters of the ticker symbol, denoted as S(3)–U(3). With 26 letters in the alphabet and with two pairings on three possible letters in the ticker symbol, thousands of combinations are possible. In searching

[1] Harvey and Liu (2014) presented a similar exhibit with purely simulated (fake) strategies.

EXHIBIT 1

Long–Short Market-Neutral Strategy Based on NYSE Stocks, January 1963 to December 2015



Notes: Gray areas denote NBER recessions. Strategy returns scaled to match S&P 500 T-bill volatility during this period.
Source: Campbell Harvey, using data from CRSP.

all potential combinations,[2] the chances of finding a strategy that looks pretty good are pretty high.

A data-mined strategy that has a nonsensical basis is, of course, unlikely to fool investors. We do not see exchange-traded funds popping up that offer "alphabets," each specializing in a letter of the alphabet. Although a strategy with no economic foundation might have worked in the past by luck, any future success would be the result of equally random luck.

The strategy detailed in Exhibit 1, as preposterous as it seems, holds important lessons in both data mining and machine learning. First, the S(3)−U(3) strategy was discovered by brute force, not machine learning. Machine learning implementations would carefully cross-validate the data by training the algorithm on part of the data and then validating on another part of the data. As Exhibit 1 shows, however, in a simple implementation when the S(3)−U(3) strategy was identified in the first quarter-century of the sample, it would be "validated" in the second quarter-century. In other words, it is possible that a false strategy can work in the cross-validated sample. In this case, the cross-validation is not randomized; as a result, a single historical path can be found.

The second lesson is that the data are very limited. Today, we have about 55 years of high-quality equity data (or less than 700 monthly observations) for many of the metrics in each of the stocks we may wish to consider. This tiny sample is far too small for most machine learning applications and impossibly small for advanced approaches such as deep learning. Third, we have a strong prior that the strategy is false: If it works, it is only because of luck. Machine learning, and particularly unsupervised machine learning, does not impose

[2] Online tools, such as those available at http://datagrid.lbl.gov/backtest/index.php, generate fake strategies that are as impressive as the one illustrated in Exhibit 1.

economic principles. If it works, it works in retrospect but not necessarily in the future.

When data are limited, economic foundations become more important. Chordia, Goyal, and Saretto (2017) examined 2.1 million equity-based trading strategies that use different combinations of indicators based on data from Compustat. They carefully took data mining into account by penalizing each discovery (i.e., by increasing the hurdle for significance). They identified 17 strategies that "survive the statistical and economic thresholds."

One of the strategies is labeled (dltis-pstkr)/mrc4. This strategy sorts stocks as follows: The numerator is long-term debt issuance minus preferred/preference stock redeemable. The denominator is minimum rental commitments four years into the future. The statistical significance is impressive, nearly matching the high hurdle established by researchers at CERN when combing through quintillions of observations to discover the elusive Higgs boson (ATLAS Collaboration 2012; CMS Collaboration 2012). All 17 of the best strategies Chordia, Goyal, and Saretto identified have a similarly peculiar construction, which—in our view and in the view of the authors of the paper—leaves them with little or no economic foundation, even though they are based on financial metrics.

Our message on the use of machine learning in backtests is one of caution and is consistent with the admonitions of López de Prado (2018). Machine learning techniques have been widely deployed for uses ranging from detection of consumer preferences to autonomous vehicles, all situations that involve big data. The large amount of data allows for multiple layers of cross-validation, which minimizes the risk of overfitting. We are not so lucky in finance. Our data are limited. We cannot flip a 4TeV switch at a particle accelerator and create trillions of fresh (not simulated) out-of-sample data. But we are lucky in that finance theory can help us filter out ideas that lack an ex ante economic basis.[3]

We also do well to remember that we are not investing in signals or data; we are investing in financial assets that represent partial ownership of a business, or of debt, or of real properties, or of commodities. The quantitative community is sometimes so focused on its models that we seem to forget that these models are crude approximations of the real world and cannot possibly reflect all nuances of the assets that actually comprise our portfolios. The amount of noise usually dwarfs the signal. Finance is a world of human beings, with emotions, herding behavior, and short memories, and market anomalies—opportunities that are the main source of intended profit for the quantitative community and their clients—are hardly static. They change with time and are often easily arbitraged away. We ignore the gaping chasm between our models and the real world at our peril.

THE WINNER'S CURSE

Most in the quantitative community will acknowledge the many pitfalls in model development. Considerable incentives exist to beat the market and to outdo the competition. Countless thousands of models are tried. In contrast to our example with ticker symbols, most of this research explores variables that most would consider reasonable. An overwhelming number of these models do not work and are routinely discarded. Some, however, do appear to work. Of the models that appear to work, how many really do, and how many are just the product of overfitting?

Many opportunities exist for quantitative investment managers to make mistakes. The most common mistake is being seduced by the data into thinking a model is better than it is. This mistake has a behavioral underpinning. Researchers want their model to work. They seek evidence to support their hypothesis—and all of the rewards that come with it. They believe if they work hard enough, they will find the golden ticket. This induces a type of selection problem in which the models that make it through are likely to be the result of a biased selection process.

Models with strong results will be tested, modified, and retested, whereas models with poor results will be quickly expunged. This creates two problems. One is that some good models will fail in the test period, perhaps for reasons unique to the dataset, and will be forgotten. The other problem is that researchers seek a narrative to justify a bad model that works well in the test period, again perhaps for reasons irrelevant to the future efficacy of the model. These outcomes are false negatives and false positives, respectively. Even more common

[3] Economists have an advantage over physicists in that societies are human constructs. Economists research what humans have created, and as humans, we know how we created it. Physicists are not so lucky.

than a false positive is an *exaggerated* positive, an outcome that seems stronger, perhaps much stronger, than it is likely to be in the future.

In other areas of science, this phenomenon is sometimes called the *winner's curse*. This is not the same winner's curse as in auction theory. The researcher who is first to publish the results of a clinical trial is likely to face the following situation: Once the trial is replicated, one of three different outcomes can occur.[4] First (sadly the least common outcome), the trial stands up to many replication tests, even with a different sample, different time horizons, and other out-of-sample tests, and continues to work after its original publication roughly as well as in the backtests. Second, after replication, the effect is far smaller than in the original finding (e.g., if microcap stocks are excluded or if the replication is out of sample). The third outcome is the worst: There is no effect, and the research is eventually discredited. Once published, models rarely work as well as in the backtest.[5]

Can we avoid the winner's curse? Not entirely, but with a strong research culture, it is possible to mitigate the damage.

AVOIDING FALSE POSITIVES: A PROTOCOL

The goal of investment management is to present strategies to clients that perform, as promised, in live trading. Researchers want to minimize false positives but to do so in a way that does not miss too many good strategies. Protocols are widely used both in scientific experiments and in practical applications. For example, every pilot is now required to go through a protocol (sometimes called a checklist) before takeoff, and airline safety has greatly improved in recent years. More generally, the use of protocols has been shown to increase performance standards and prevent failure, as tasks become increasingly complex (e.g., Gawande 2009). We believe that the use of protocols for quantitative research in finance should become de rigueur, especially for machine learning–based techniques, as computing power and process complexity grow. Our goal is to improve investor outcomes in the context of backtesting.

Many items in the protocol we suggest are not new (e.g., Harvey 2017, Fabozzi and López de Prado 2018, and López de Prado 2018), but in this modern era of data science and machine learning, we believe it worthwhile to specify best research practices in quantitative finance.

CATEGORY #1: RESEARCH MOTIVATION

Establish an Ex Ante Economic Foundation

Empirical research often provides the basis for the development of a theory. Consider the relation between experimental and theoretical physics. Researchers in experimental physics measure (generate data) and test the existing theories. Theoretical physicists often use the results of experimental physics to develop better models. This process is consistent with the concept of the scientific method: A hypothesis is developed, and the empirical tests attempt to find evidence inconsistent with the hypothesis—so-called falsifiability.[6]

The hypothesis provides a discipline that reduces the chance of overfitting. Importantly, the hypothesis needs to have a logical foundation. For example, the "alpha-bet" long–short trading strategy in Exhibit 1 has no theoretical foundation, let alone a prior hypothesis. Bem (2011) published a study in a top academic journal that "supported" the existence of extrasensory perception using over 1,000 subjects in 10 years of experiments. The odds of the results being a fluke were 74 billion to 1. They were a fluke: The tests were not successfully replicated.

The researcher invites future problems by starting an empirical investigation without an ex ante economic hypothesis. First, it is inefficient even to consider models or variables without an ex ante economic hypothesis (such as scaling a predictor by rental payments due in the fourth year, as in Exhibit 1). Second, no matter the outcome, without an economic foundation for the

[4] In investing, two of these three outcomes pose a twist to the winner's curse: private gain and social loss. The investment manager pockets the fees until the flaw of the strategy becomes evident, and the investor bears the losses until the great reveal that it was a bad strategy all along.

[5] See McLean and Pontiff (2016). Arnott, Beck, and Kalesnik (2016) examined eight of the most popular factors and showed an average return of 5.8% a year in the span before the factors' publication and a return of only 2.4% after publication. This loss of nearly 60% of the alpha on a long–short portfolio before any fees or trading costs is far more slippage than most observers realize.

[6] One of the most damning critiques of theories in physics is to be deemed unfalsifiable. Should we hold finance theories to a lesser standard?

model, the researcher maximizes the chance that the model will not work when taken into live trading. This is one of the drawbacks of machine learning.

One of our recommendations is to carefully structure the machine learning problem so that the inputs are guided by a reasonable hypothesis. Here is a simple example: Suppose the researcher sets a goal of finding a long–short portfolio of stocks that outperforms on a risk-adjusted basis, using the full spectrum of independent variables available in Compustat and I/B/E/S. This is asking for trouble. With no particular hypothesis, and even with the extensive cross-validation done in many machine learning applications, the probability of a false positive is high.

Beware an Ex Post Economic Foundation

It is also almost always a mistake to create an economic story—a rationale to justify the findings—after the data mining has occurred. The story is often flimsy, and if the data mining had delivered the opposite result, the after-the-fact story might easily have been the opposite. An economic foundation should exist first, and a number of empirical tests should be designed to test how resilient that foundation is. Any suspicion that the hypothesis was developed *after* looking at the data is an obvious red flag.

Another subtle point: In other disciplines such as medicine, researchers often do not have a prespecified theory, and data exploration is crucial in shaping future clinical trials. These trials provide the researcher with truly out-of-sample data. In finance and economics, we do not have the luxury of creating a large out-of-sample test. It is therefore dangerous to appropriate this exploratory approach into our field. We may not jeopardize customer health, but we will jeopardize their wealth. This is particularly relevant when it comes to machine learning methods, which were developed for more data-rich disciplines.

CATEGORY #2: MULTIPLE TESTING AND STATISTICAL METHODS

Keep Track of What Is Tried

Given 20 randomly selected strategies, one strategy will likely exceed the two-sigma threshold (t-statistic of 2.0 or above) purely by chance. As a result, the t-statistic of 2.0 is not a meaningful benchmark if more than one strategy is tested. Keeping track of the number of strategies tried is crucial, as is measuring their correlations (Harvey 2017; López de Prado 2018). A bigger penalty in terms of threshold is applied to strategies that are relatively uncorrelated. For example, if the 20 strategies tested had a near 1.0 correlation, then the process is equivalent to trying only one strategy.

Keep Track of Combinations of Variables

Suppose the researcher starts with 20 variables and experiments with some interactions, say (variable 1 × variable 2) and (variable 1 × variable 3). This single interaction does not translate into only 22 tests (the original 20, plus two additional interactions) but into 190 possible interactions. Any declared significance should take the full range of interactions into account.[7]

Beware the Parallel Universe Problem

Suppose a researcher develops an economic hypothesis and tests the model once; that is, the researcher decides on the data, variables, scaling, and type of test—all in advance. Given the single test, the researcher believes the two-sigma rule is appropriate, but perhaps it is not. Think of being in 20 different parallel universes. In each, the researcher chooses a different model informed on the identical history. In each, the researcher performs a single test. One of them works. Is it significant at two sigma? Probably not.

Another way to think about this is to suppose that (in a single universe) the researcher compiles a list of 20 variables to test for predictive ability. The first one "works." The researcher stops and claims to have done a single test. True, but the outcome may be lucky. Think of another researcher with the same 20 variables who tests in a different order, and only the last variable "works." In this case, a discovery at two sigma would be discarded because a two-sigma threshold is too low for 20 different tests.

CATEGORY #3: SAMPLE CHOICE AND DATA

Define the Test Sample Ex Ante

The training sample needs to be justified in advance. The sample should never change after the research begins. For example, suppose the model

[7] There are *20 choose 2* interactions, which is 20!/(18!2!).

"works" if the sample begins in 1970 but does not work if the sample begins in 1960—in such a case, the model does not work. A more egregious example would be to delete the global financial crisis data, the tech bubble, or the 1987 market crash because they hurt the predictive ability of the model. The researcher must not massage the data to make the model work.

Ensure Data Quality

Flawed data can lead researchers astray. Any statistical analysis of the data is only as good as the quality of the data that are input, especially in the case of certain machine learning applications that try to capture nonlinearities. A nonlinearity might simply be a bad data point.

The idea of garbage in/garbage out is hardly new. Provenance of the data needs to be taken into account. For example, data from CRSP, Compustat, or some other "neutral" provider should have a far higher level of trust than raw data supplied by some broker. In the past, researchers would literally eyeball smaller datasets and look for anomalous observations. Given the size of today's datasets, the human eyeball is insufficient. Cleaning the data before employing machine learning techniques in the development of investment models is crucial. Interestingly, some valuable data science tools have been developed to check data integrity. These need to be applied as a first step.

Document Choices in Data Transformations

Manipulation of the input data (e.g., volatility scaling or standardization) is a choice and is analogous to trying extra variables. The choices need to be documented and ideally decided in advance. Furthermore, results need to be robust to minor changes in the transformation. For example, given 10 different volatility-scaling choices, if the one the researcher chose is the one that performed the best, this is a red flag.

Do Not Arbitrarily Exclude Outliers

By definition, outliers are influential observations for the model. Inclusion or exclusion of influential observations can make or break the model. Ideally, a solid economic case should be made for exclusion—*before* the model is estimated. In general, no influential observations should be deleted. Assuming the observation is based on valid data, the model should explain all data, not just a select number of observations.

Select Winsorization Level before Constructing the Model

Winsorization is related to data exclusion. Winsorized data are truncated at a certain threshold (e.g., truncating outliers to the 1% or 2% tails) rather than deleted. Winsorization is a useful tool because outliers can have an outsize influence on any model, but the choice to winsorize, and at which level, should be decided before constructing the model. An obvious sign of a faulty research process is a model that "works" at a winsorization level of 5% but fails at 1%, and the 5% level is then chosen.

CATEGORY #4: CROSS-VALIDATION

Acknowledge Out of Sample Is Not Really Out of Sample

Researchers have lived through the hold-out sample and thus understand the history, are knowledgeable about when markets rose and fell, and associate leading variables with past experience. As such, no true out-of-sample data exist; the only true out of sample is the live trading experience.

A better out-of-sample application is on freshly uncovered historical data; for example, some researchers have tried to backfill the historical database of US fundamental data to the 1920s. It is reasonable to assume these data have not been data mined because the data were not previously available in machine readable form. But beware: Although these data were not previously available, well-informed researchers are aware of how history unfolded and how macroeconomic events were correlated with market movements. For those well versed on the history of markets, these data are in sample in their own experience and in shaping their own prior hypotheses. Even for those less knowledgeable, today's conventional wisdom is informed by past events.

As with deep historical data, applying the model in different settings is a good idea but should be done with caution because correlations exist across countries. For example, a data-mined (and potentially fake) anomaly that works in the US market over a certain sample may also work in Canada or the United Kingdom over the same time span, given the correlation between these markets.

Recognize That Iterated Out of Sample Is Not Out of Sample

Suppose a model is successful in the in-sample period but fails out of sample. The researcher observes that the model fails for a particular reason. The researcher modifies the initial model so it then works both in sample and out of sample. This is no longer an out-of-sample test. It is overfitting.

Do Not Ignore Trading Costs and Fees

Almost all of the investment research published in academic finance ignores transactions costs.[8] Even with modest transactions costs, the statistical significance of many published anomalies essentially vanishes. Any research on historical data needs to take transactions costs and, more generally, implementation shortfall into account in both the in-sample and out-of-sample analysis (Arnott 2006).

CATEGORY #5: MODEL DYNAMICS

Be Aware of Structural Changes

Certain machine applications have the ability to adapt through time. In economic applications, structural changes—or nonstationarities—exist. This concern is largely irrelevant in the physical and biological sciences. In finance, we are not dealing with physical constants; we are dealing with human beings and with changing preferences and norms. Once again, the amount of available data is limiting, and the risk of overfitting the dynamics of a relation through time is high.

Acknowledge the Heisenberg Uncertainty Principle and Overcrowding

In physics, the Heisenberg uncertainty principle states that we cannot know a particle's position and momentum simultaneously with precision. The more accurately we know one characteristic, the less accurately we can know the other. A similar principle can apply in finance. As we move from the study of past data into the live application of research, market inefficiencies are hardly static. The cross-validated relations of the past may seem powerful for reasons that no longer apply or may dissipate merely because we are now aware of them and are trading based on them.

Indeed, the mere act of studying and refining a model serves to increase the mismatch between our expectations of a model's efficacy and the true underlying efficacy of the model—and that is before we invest live assets, moving asset prices and shrinking the efficacy of the models through our own collective trading.

Refrain from Tweaking the Model

Suppose the model is running but not doing as well as expected. Such a case should not be a surprise because the backtest of the model is likely overfit to some degree. It may be tempting to tweak the model, especially as a means to improve its fit in recent, now in-sample, data. Although these modifications are a natural response to failure, we should be fully aware that they will generally lead to further overfitting of the model and may lead to even worse live-trading performance.

CATEGORY #6: MODEL COMPLEXITY

Beware the Curse of Dimensionality

Multidimensionality works against the viability of machine learning applications; the reason is related to the limitations of data. Every new piece of information increases dimensionality and requires more data. Recall the research of Chordia, Goyal, and Saretto (2017), who examined 2.1 million equity models based on Compustat data. There are orders of magnitude more models than assets. With so many models, some will work very well in sample.

Consider a model to predict the cross section of stock prices. One reasonable variable to explore is past stock prices (momentum), but many other variables, such as volume, trailing volatility, bid–ask spread, and option skew, could be considered. As each possible predictor variable is added, more data are required, but history is limited and new data cannot be created or simulated.[9]

[8] See Asness and Frazzini (2013). Hou, Xue, and Zhang (2017) showed that most anomaly excess returns disappear once microcaps are excluded.

[9] Monte Carlo simulations are part of the toolkit, perhaps less used today than in the past. Of course, simulations will produce results entirely consonant with the assumptions that drive the simulations.

Macroeconomic analysis provides another example. Although most believe that certain economic state variables are important drivers of market behavior and expected returns, macroeconomic data, generally available on a monthly or quarterly basis, are largely offside for most machine learning applications. Over the post-1960 period,[10] just over 200 quarterly observations and fewer than 700 monthly observations exist.

Although the number of historical observations is limited for each time series, a plethora of macroeconomic variables is available. If we select one or two to be analyzed, we create an implicit data-mining problem, especially given that we have lived through the chosen out-of-sample period.

Pursue Simplicity and Regularization

Given data limitations, regularizing by imposing structure on the data is important. Regularization is a key component of machine learning. It might be the case that a machine learning model decides that a linear regression is the best model. If, however, a more elaborate machine learning model beats the linear regression model, it had better win by an economically significant amount before the switch to a more complex model is justified.

A simple analogy is a linear regression model of Y on X. The in-sample fit can almost always be improved by adding higher powers of X to the model. In out-of-sample testing, the model with the higher powers of X will often perform poorly.

Current machine learning tools are designed to minimize the in-sample overfitting by extensive use of cross-validation. Nevertheless, these tools may add complexity (which is potentially nonintuitive) that leads to disappointing performance in true out-of-sample live trading. The greater the complexity and the reliance on nonintuitive relationships, the greater the likely slippage between backtest simulations and live results.

Seek Interpretable Machine Learning

It is important to look under the hood of any machine learning application. It cannot be a black box. Investment managers should know what to expect with any machine learning–based trading system. Indeed, an interesting new subfield in computer science focuses on interpretable classification and interpretable policy design (e.g., Wang et al. 2017).

CATEGORY #7: RESEARCH CULTURE

Establish a Research Culture That Rewards Quality

The investment industry rewards research that produces backtests with winning results. If we do this in actual asset management, we create a toxic culture that institutionalizes incentives to hack the data to produce a seemingly good strategy. Researchers should be rewarded for good science, not good results. A healthy culture will also set the expectation that most experiments will fail to uncover a positive result. Both management and researchers must have this common expectation.

Be Careful with Delegated Research

No one can perform every test that could potentially render an interesting result, so researchers will often delegate. Delegated research needs to be carefully monitored. Research assistants have an incentive to please their supervisor by presenting results that support the supervisor's hypothesis. This incentive can lead to a free-for-all data-mining exercise that is likely to lead to failure when applied to live data.

Exhibit 2 condenses the foregoing discussion into a seven-point protocol for research in quantitative finance.

CONCLUSIONS

The nexus of unprecedented computing power, free software, widely available data, and advances in scientific methods provide us with unprecedented opportunities for quantitative research in finance. Given these unprecedented capabilities, we believe it is useful to take a step back and reflect on the investment industry's research process. It is naïve to think we no longer need economic models in the era of machine learning. Given that the quantity (and quality) of data is relatively limited in finance, machine learning applications face many of the same issues quantitative finance researchers have struggled with for decades.

[10] Monthly macroeconomic data generally became available in 1959.

EXHIBIT 2
Seven-Point Protocol for Research in Quantitative Finance

1. Research Motivation
- a. Does the model have a solid economic foundation?
- b. Did the economic foundation or hypothesis exist before the research was conducted?

2. Multiple Testing and Statistical Methods
- a. Did the researcher keep track of all models and variables that were tried (both successful and unsuccessful), and are the researchers aware of the multiple-testing issue?
- b. Is there a full accounting of all possible interaction variables if interaction variables are used?
- c. Did the researchers investigate all variables set out in the research agenda, or did they cut the research as soon as they found a good model?

3. Data and Sample Choice
- a. Do the data chosen for examination make sense? And, if other data are available, is it reasonable to exclude these data?
- b. Did the researchers take steps to ensure the integrity of the data?
- c. Do the data transformations, such as scaling, make sense? Were they selected in advance? And are the results robust to minor changes in these transformations?
- d. If outliers are excluded, are the exclusion rules reasonable?
- e. If the data are winsorized, was there a good reason to do it? Was the winsorization rule chosen before the research was started? Was only one winsorization rule tried (as opposed to many)?

4. Cross-Validation
- a. Are the researchers aware that true out-of-sample tests are only possible in live trading?
- b. Are steps in place to eliminate the risk of out-of-sample iterations (i.e., an in-sample model that is later modified to fit out-of-sample data)?
- c. Is the out-of-sample analysis representative of live trading? For example, are trading costs and data revisions taken into account?

5. Model Dynamics
- a. Is the model resilient to structural change, and have the researchers taken steps to minimize the overfitting of the model dynamics?
- b. Does the analysis take into account the risk/likelihood of overcrowding in live trading?
- c. Do researchers take steps to minimize the tweaking of a live model?

6. Complexity
- a. Is the model designed to minimize the curse of dimensionality?
- b. Have the researchers taken steps to produce the simplest practicable model specification?
- c. Has an attempt been made to interpret the predictions of the machine learning model rather than using it as a black box?

7. Research Culture
- a. Does the research culture reward the quality of the science rather than the finding of a winning strategy?
- b. Do the researchers and management understand that most tests will fail?
- c. Are expectations clear (that researchers should seek the truth, not just something that works) when research is delegated?

In this article, we have developed a research protocol for investment strategy backtesting. The list is applicable to most research tools used in investment strategy research—from portfolio sorts to machine learning. Our list of prescriptions and proscriptions is long, but hardly exhaustive.

Importantly, the goal is not to eliminate all false positives. Indeed, that is easy—just reject every single strategy. One of the important challenges we face is satisfying the dual objectives of minimizing false strategies but not missing too many good strategies at the same time. The optimization of this trade-off is the subject of ongoing research (see Harvey and Liu 2018).

At first reading, our observations may seem trivial and obvious. Importantly, our goal is not to criticize quantitative investing. Our goal is to encourage humility, to recognize that we can easily deceive ourselves into thinking we have found the Holy Grail. Hubris is our enemy. A protocol is a simple step. Protocols can improve outcomes, whether in a machine shop, an airplane cockpit, a hospital, or for an investment manager. For the investment manager, the presumptive goal is an investment process that creates the best possible opportunity to match or exceed expectations when applied in live trading. Adopting this process is good for the client and good for the reputation of the investment manager.

ACKNOWLEDGMENTS

We have benefited from the comments of Frank Fabozzi, Marcos López de Prado, and Joseph Simonian.

REFERENCES

Anthony, S. "The History of Supercomputers." ExtremeTech .com, April 10, 2012.

Arnott, R. 2006. "Implementation Shortfall." *Financial Analysts Journal* 62 (3) (May/June): 6–8.

Arnott, R., N. Beck, and V. Kalesnik. "Timing 'Smart Beta' Strategies? Of Course! Buy Low, Sell High!" Research Affiliates Publications, September 2016.

Asness, C., and A. Frazzini. 2013. "The Devil in HML's Details." *The Journal of Portfolio Management* 39 (4): 49–68.

ATLAS Collaboration. 2012. "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC." *Physics Letters B* 716 (1): 1–29.

Bem, D. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *Journal of Personality and Social Psychology* 100 (3): 407–425.

Bookman, S. "15 Huge Supercomputers That Were Less Powerful Than Your Smartphone." TheClever.com, April 18, 2017.

Chordia, T., A. Goyal, and A. Saretto. 2017. "*p*-Hacking: Evidence from Two Million Trading Strategies." Swiss Finance Institute Research Paper No. 17–37, SSRN.

CMS Collaboration. 2012. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC." *Physics Letters B* 716 (1): 30–61.

Fabozzi, F., and M. López de Prado. 2018. "Being Honest in Backtest Reporting: A Template for Disclosing Multiple Tests." *The Journal of Portfolio Management* 45 (1): 141–147.

Gawande, A. *The Checklist Manifesto: How to Get Things Right*. New York: Henry Holt and Sons, 2009.

Harvey, C. R. 2017. "Presidential Address: The Scientific Outlook in Financial Economics." *The Journal of Finance* 72: 1399–1440.

Harvey, C. R., and Y. Liu. 2014. "Evaluating Trading Strategies." *The Journal of Portfolio Management* 40 (5): 108–118.

———. 2018. "False (and Missed) Discoveries in Financial Economics." SSRN, https://ssrn.com/abstract=3073799.

Hou, K., C. Xue, and L. Zhang. 2017. "Replicating Anomalies." SSRN, https://www.ssrn.com/abstract=2961979.

Leamer, E. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: John Wiley & Sons, 1978.

Lo, A., and A. C. MacKinlay. 1990. "Data-Snooping Biases in Tests of Financial Asset Pricing Models." *Review of Financial Studies* 3 (3): 431–467.

López de Prado, M. 2018. "The 10 Reasons Most Machine Learning Funds Fail." *The Journal of Portfolio Management* 44 (6): 120–133.

Markowitz, H., and B. L. Xu. 1994. "Data Mining Corrections." *The Journal of Portfolio Management* 21 (1): 60–69.

McLean, R. D., and J. Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *The Journal of Finance* 71 (1): 5–32.

Wang, T., C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille. 2017. "A Bayesian Framework for Learning Rule Sets for Interpretable Classification." *Journal of Machine Learning Research* 18: 1–37.