The Insidious Threat of Data Mining Bias

In the context of systematic trading research, data mining bias refers to the risk of attributing significance to a result that was in fact due to chance. I refer to it as an "insidious threat" because it creeps into the research process naturally and quietly and can have disastrous results if not accounted for. In the extreme, this leads to the decision to trade a losing strategy, but at the very least, neglecting data mining bias leads to inflated and unrealistic performance expectations.

Data mining bias arises due to two fundamental characteristics of trading system research and development. The first is randomness, and the second is sequential comparison, which is the search for the 'best' parameter set.

In any series of trades arising from a given system, the realised performance will have an element of randomness as well as an element due to the inherent edge of the system. If we examine only a single system variant, we have no way of knowing the ratio of the two components. The random component is equally likely to be positive or negative, and in some cases, may be extreme. In any event, this random component results in variation around the long-term expectation of the system.

Sequential comparison and selection leads to the system variant with the best performance in the historical simulation being selected. Consider this selection process in the context of the randomness associated with strategy performance: the selected strategy variant's random component is guaranteed to have been favourable. As we increase the number of variants tested, the probability of a favourable result being due to chance increases. To put this another way, the component of the selected strategy's performance due to chance increases with the number of variants tested.

With modern tools like machine learning and fast computing, it is easy to assess thousands or tens of thousands of system variants relatively quickly. This is dangerous, because given a finite historical data set, you will almost certainly stumble across something that worked on that data set due to chance alone, even if you have used an appropriate out of sample validation regime and controlled overfitting at the algorithm level. This is akin to attributing skill to the winner of a lottery. With enough entries in the lottery (several hundred million), at least one is likely to consist of the winning numbers. In this case, it is obvious that the winning entry was due to chance. In trading strategy research, whether a result was due to chance or a real edge is not so immediately apparent.

This comic from XKCD illustrates data mining bias very eloquently.

Quantify

Importantly, this sequential comparison or optimisation process is not the problem. It is of course an indispensable aspect of strategy research. However, the problem arises when the performance of the chosen variant is used as a yardstick for future performance. The statistical law of regression to the mean (not be confused with mean reversion as a characteristic of financial markets) tells us that the same amount of favourable randomness that led to the selection of the particular strategy variant is not likely to repeat. Supposing otherwise leads to inflated performance expectations and biased decision making.

Surmounting this obstacle requires a procedure for accounting for data-mining bias. White's Reality Check is one such method, but it has its own limitations, such as being prone to producing false negatives (that is, rejecting good systems). There have been improvements proposed over the years, such as the papers by Romano and Wolf (2005), Hanson (2005) and Corradi and Swanson (2011). Another useful tool is System Parameter Permutation (SPP) which was proposed by Dave Walton in his 2014 Wagner Award-winning paper. SPP accounts for data mining bias by approaching strategy evaluation not via sequential comparison, but by assessing the distribution of performance across a strategy's entire parameter space.

How to manage data mining bias

Addressing this issue properly involves developing a thorough understanding of the potential risks and rewards associated with a trading system, allowing a fund manager to make data-driven, probabilistic decisions regarding potential allocation. In comparison, traditional out-of-sample testing and cross-validation approaches necessitate binary decision making and if used in isolation cause the fund manager to forego a much deeper understanding of the system's performance characteristics.

Out-of-the-box toolkits exist to help manage this process which are suitable for use in any parameter-based trading system. The tools typically utilise all available data and as many strategy variants as possible to uncover empirical estimates of true strategy risk and reward. Contrast this with traditional out-of-sample testing and cross-validation methods which select only a single strategy variant based on a finite back-test, thus discarding the vast majority of available information. These latter approaches, while theoretically sound when used for their intended purposes, in reality tend to introduce large, sometimes extreme, amounts of data mining bias into the strategy development and selection process. If unchecked, this bias can lead to unrealistically optimistic expectations of future performance.

A quality data mining bias toolkit consists of the following:

1. Train-Test Analysis: to determine whether performance in one period correlates with performance in the next sequential period. A positive result implies that there is value in selecting strategy variants that exhibit better performance in an in-sample testing period.

2. Uncorrelated Parameter Selection: to determine whether it is possible to construct a portfolio of strategy variants that increases our chances of achieving the strategy's long-term performance expectation.
3. Parameter Exchange Analysis: to estimate the strategy's long-term performance.
4. Build-Rebuild-Compare Analysis: to estimate the worst-case short-term performance across multiple time horizons of interest to the fund manager.
5. Focused Feature Space Investigation: to determine whether the strategy variant resulting in peak performance of the back-test period is an outlier.

Train-Test Analysis

Train-Test Analysis (TTA) consists of quantifying the correlation between a strategy's in-sample (IS) and out-of-sample (OOS) performance across a representative sample of its feature space. Ultimately, the goal of TTA is to determine whether the IS performance informs the OOS performance.

IS-OOS correlation is important because if present, it implies that parameter selection has a meaningful impact on the strategy's performance, that is, that there is value in selecting strategy variants on the basis of their recent performance. This may be best illustrated by considering the alternative: if the IS performance of a given parameter set is in no way informative of the adjacent period's performance, at least in probabilistic terms, then there is no basis for selecting one parameter set over another.

Uncorrelated Parameter Selection

Even if the TTA indicates that the performance of the strategy in one period is informative of the performance in the adjacent period, there will nearly always be a degree of uncertainty. This is evident in the spread of out-of-sample performance for a given in-sample performance.

One way to reduce this uncertainty is to select several strategy variants, each with high in-sample performance. If variants with uncorrelated returns series are chosen, it can be inferred that out-of-sample performance of the aggregated parameter sets is likely to be closer to the out-of-sample performance described by the least squares regression line between the in- and out-of-sample performance.

The process of selecting these parameter sets is Uncorrelated Parameter Selection (UPS).

Parameter Exchange Analysis

Parameter Exchange Analysis, also known as System Parameter Permutation (SPP) is used to estimate long-term performance in a manner that is not subject to data mining bias. A significant benefit of SPP is that it uses the performance data of all (or as many as feasible) sets of parameters evaluated during optimisation. Contrast this

with regular optimisation, which picks the best set of parameters, as defined in a single test run, and discards the rest.

SPP can be performed for any performance metric of interest to the trader. Here, Empirical Cumulative Distribution Functions (ECDFs) are created for each performance metric of interest and the median of each calculated. The median then serves as the best, most unbiased, estimate of true long-term system performance. The median is useful because it is not subject to data mining bias because no selection was involved in its calculation, it requires no assumptions as to the shape of the underlying distribution and is robust to outliers.

The ECDF can also be used for statistical inference and hence for making data-driven, probabilistic decisions around allocation and investment into the strategy. SPP generates complete sampling distributions, therefore estimated p-values may be observed directly from the ECDF.

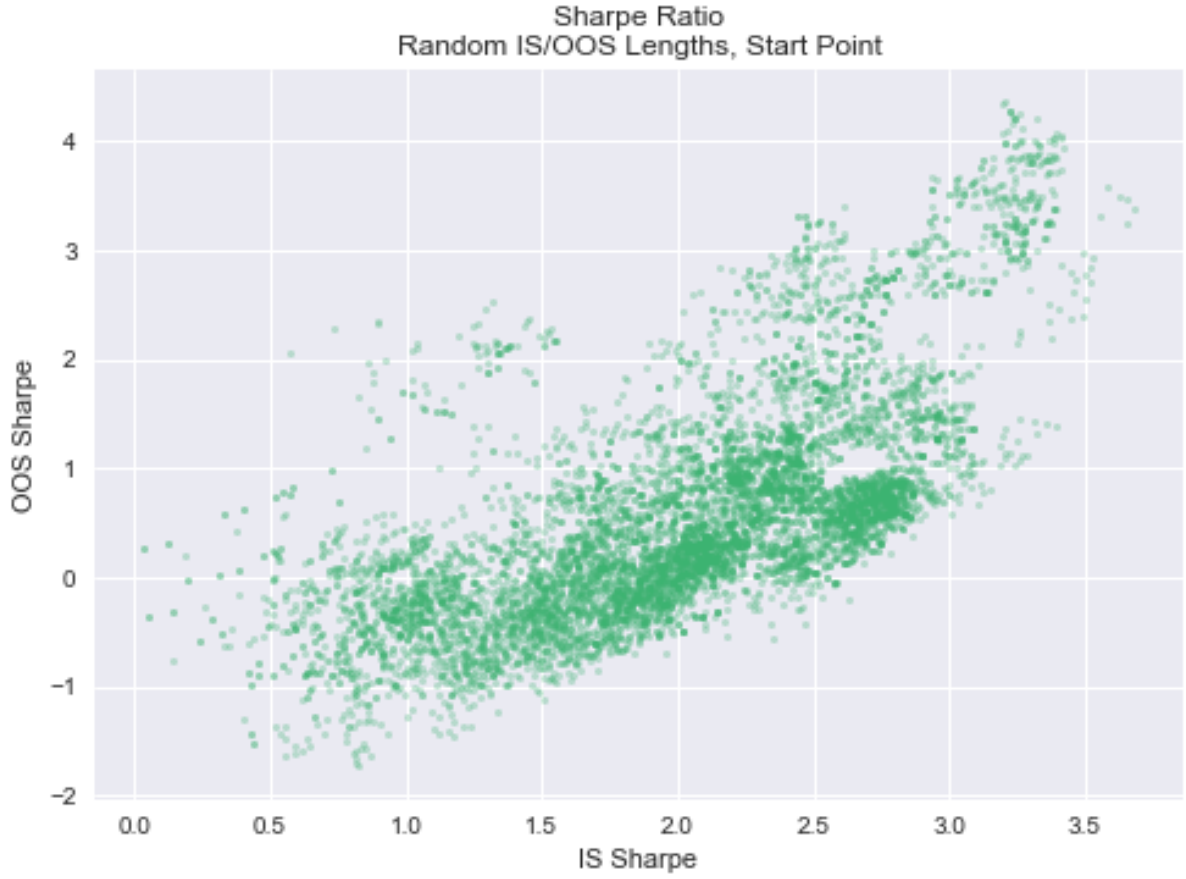
Worst-Case Short-Term Performance Estimate

In order to achieve its long-term performance, a trading system typically must endure significant short-term variability in performance. Therefore, to harvest a system's long-term performance, it is critical to understand the worst-case contingencies that must be endured. Understanding this enables the trader to again make data-driven, probabilistic decisions about allocation into the strategy.

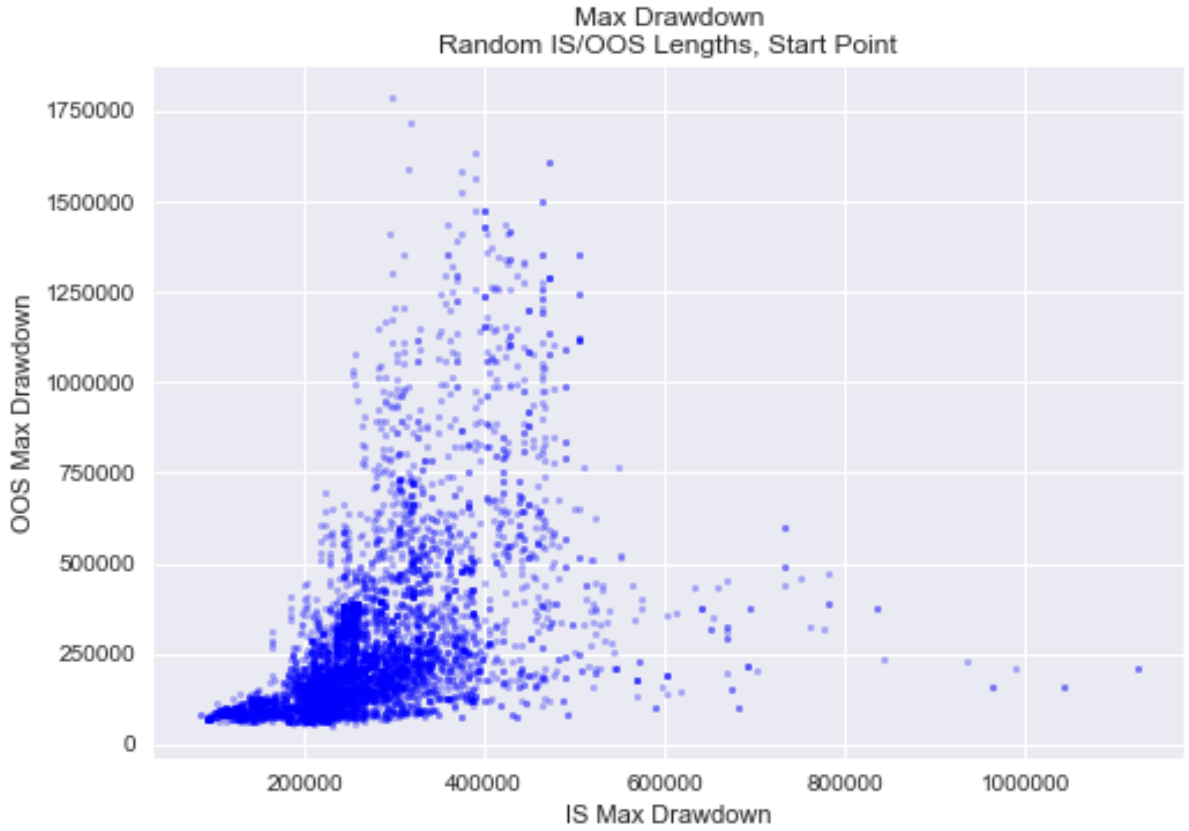
A useful by-product of this analysis is that it also provides insight into when it is wise to cease trading a strategy by providing quantitative bounds on the expected short-term performance and hence providing insight into whether the underlying returns distribution has changed significantly.

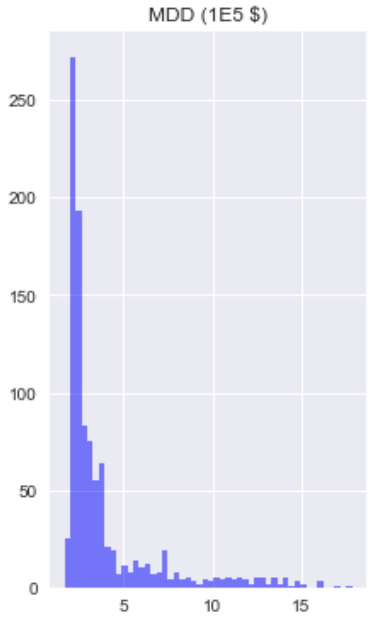
Examples

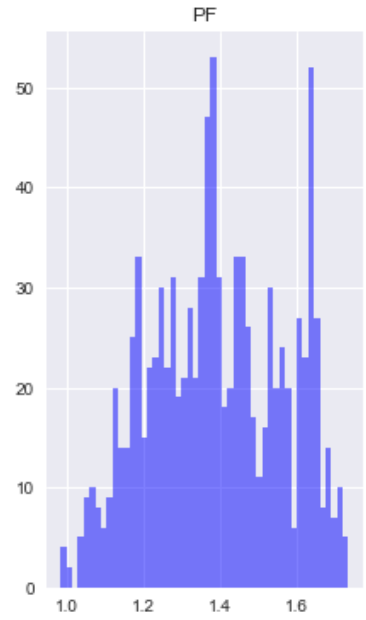
Quantify

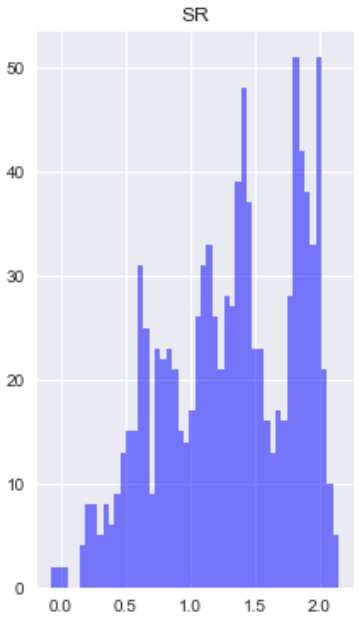

Sharpe Ratio
Random IS/OOS Lengths, Start Point
OOS Sharpe
IS Sharpe
4
3
2
1
0
−1
−2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5

Quantify

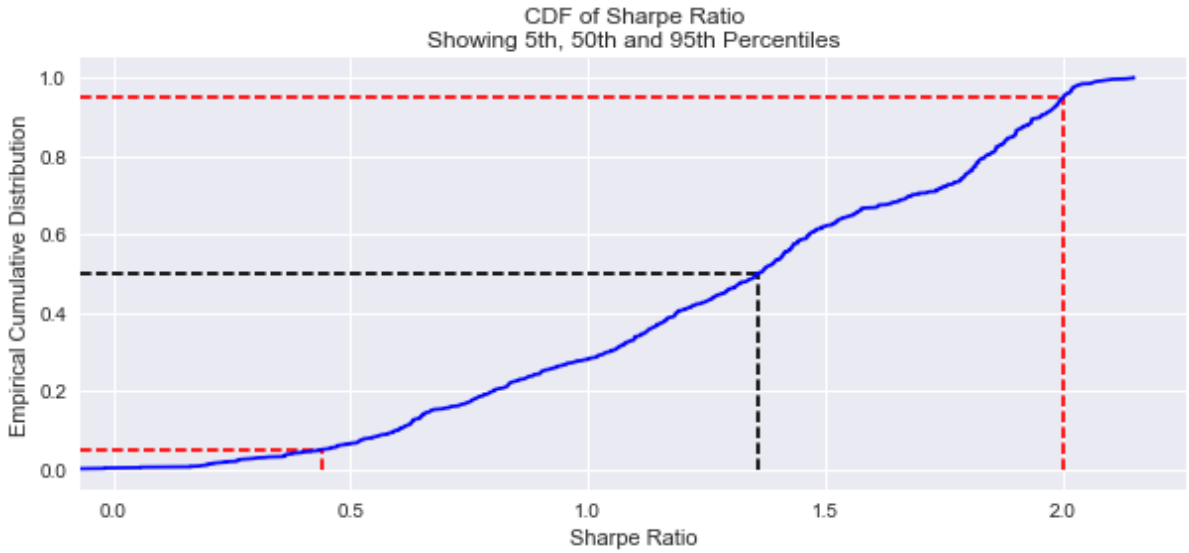

Max Drawdown
Random IS/OOS Lengths, Start Point
OOS Max Drawdown
IS Max Drawdown
1750000
1500000
1250000
1000000
750000
500000
250000
0
200000
400000
600000
800000
1000000

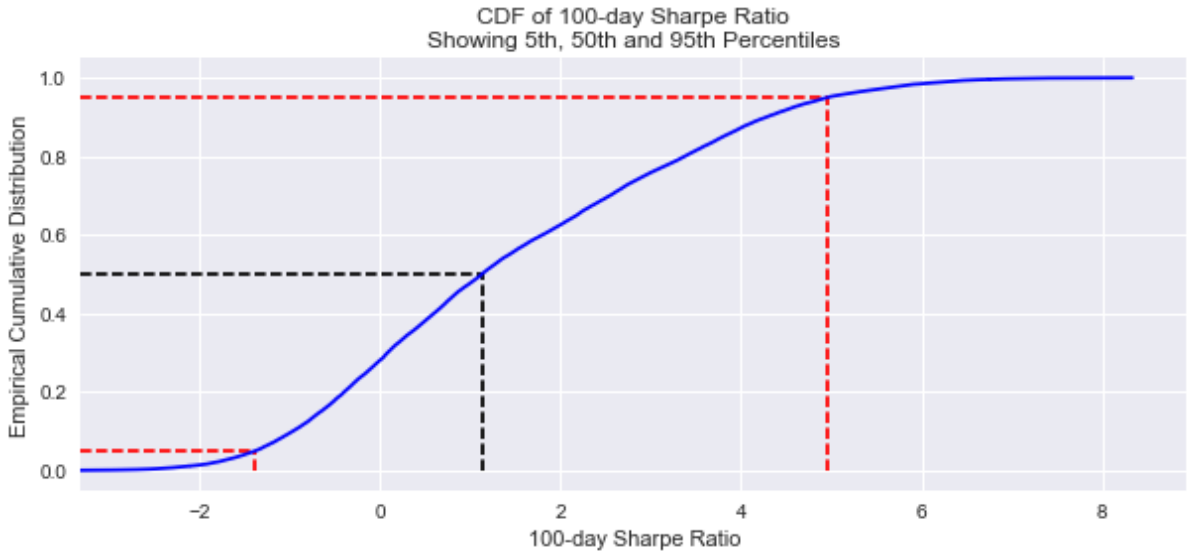

MDD (1E5 $)
250
200
150
100
50
0
5
10
15


PF
50
40
30
20
10
0
1.0
1.2
1.4
1.6

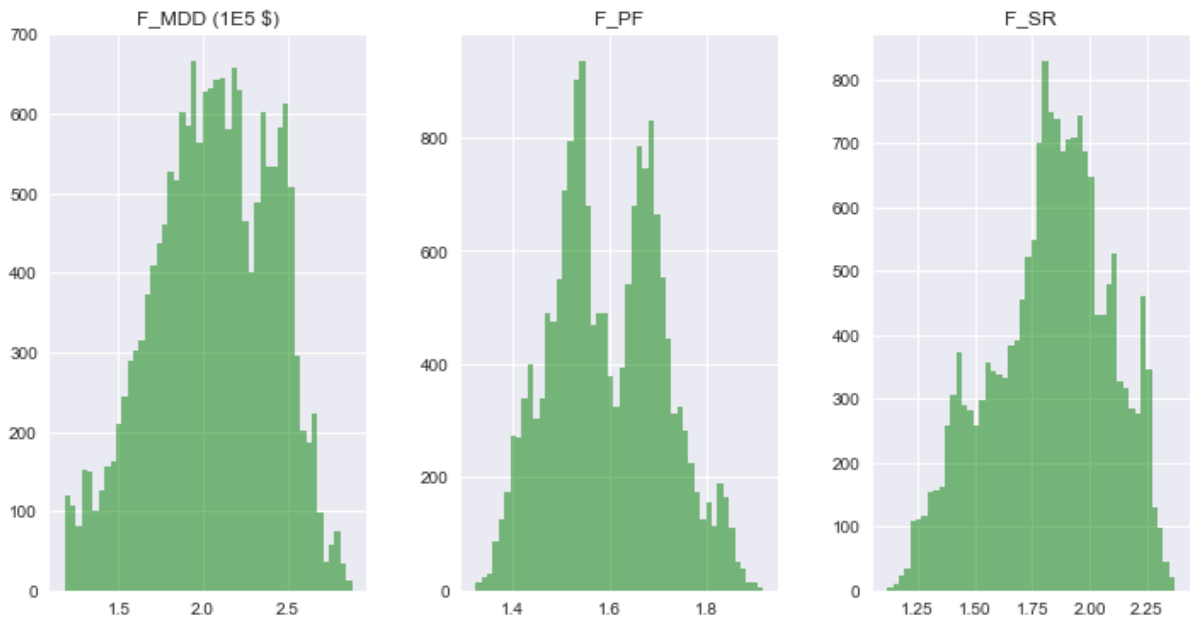

SR
50
40
30
20
10
0
0.0
0.5
1.0
1.5
2.0

Quantify


CDF of Sharpe Ratio
Showing 5th, 50th and 95th Percentiles
Empirical Cumulative Distribution
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
Sharpe Ratio


CDF of 100-day Sharpe Ratio
Showing 5th, 50th and 95th Percentiles
Empirical Cumulative Distribution
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
8
100-day Sharpe Ratio

Quantify



Kris Longmore is Founder and Head of Quantitative Research at [Quantify.]