The 7 Reasons Most Machine Learning Funds Fail

Key Points

- Over the past 20 years, I have seen many new faces arrive to the financial industry, only to leave shortly after.
- The rate of failure is particularly high in machine learning (ML) funds.
- In my experience, the reasons boil down to 7 common errors:

1. The Sisyphus paradigm
2. Integer differentiation
3. Inefficient sampling
4. Wrong labeling
5. Weighting of non-IID samples
6. Cross-validation leakage
7. Backtest overfitting

- **Warning**: A ML algorithm will always find a pattern, even if there is none.
- **Prediction**: The pervasive misuse of ML techniques by "quants" will continue to lead to False Positives, losses and failures. *ML does not fail, researchers fail.*

FINANCIAL ML ≠ ML ALGORITHMS + FINANCIAL DATA

Pitfall #1:

The Sisyphean Quants

The silo approach works for discretionary PMs

- Discretionary portfolio managers (PMs) make investment decisions that do not follow a particular theory or rigorous rationale.
- Because nobody fully understands the logic behind their bets, they can hardly work as a team and develop deeper insights beyond the initial intuition.
- If 50 PMs tried to work together, they would influence each other until eventually 49 would follow the lead of 1.

For this reason, investment firms ask discretionary PMs to work in silos.

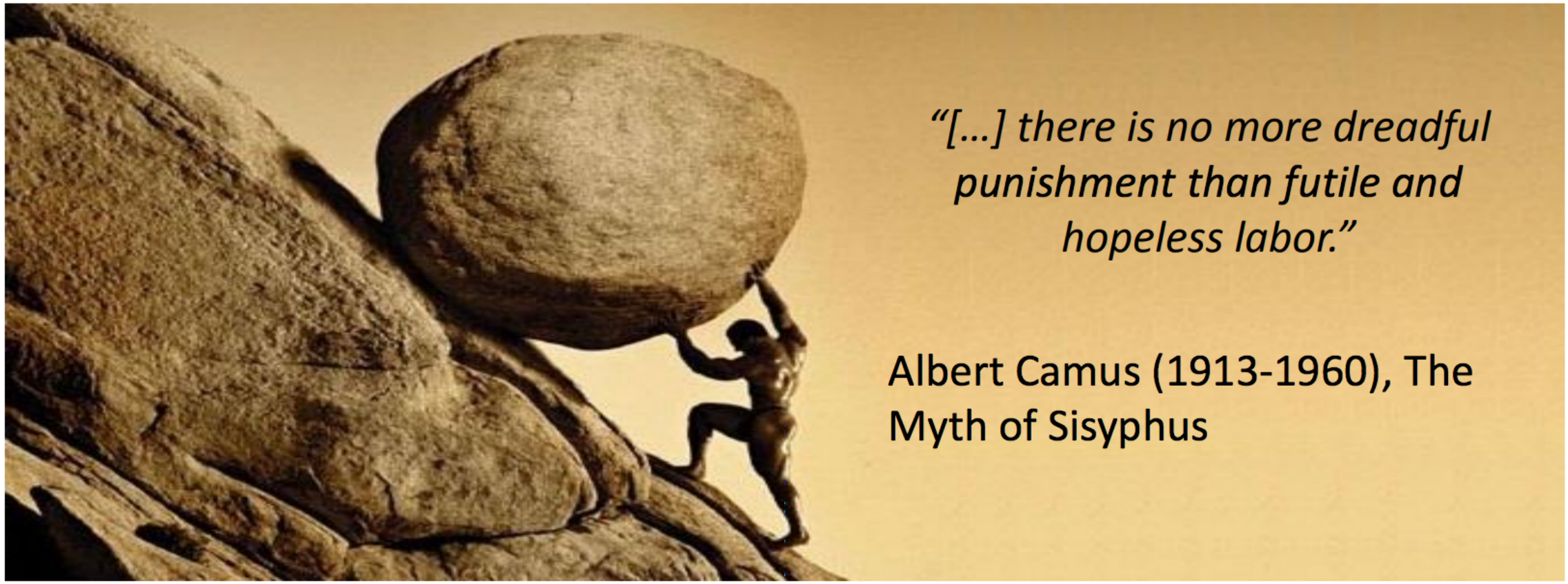
Silos prevent one PM from influencing the rest, hence protecting diversification.

The silo approach fails with quant PMs

- The boardroom's mentality is, let us do with quants what has worked with discretionary PMs.
- Let us hire 50 PhDs, and demand from each of them to produce an investment strategy within 6 months.
- This approach typically backfires, because each of these PhDs will frantically search for investment opportunities and eventually settle for:
  - A false positive that looks great in an overfit backtest; or
  - A standard factor model, which is an overcrowded strategy with low Sharpe ratio, but at least has academic support.
- Both outcomes will disappoint the investment board, and the project will be cancelled.
- Even if 5 of those 50 PhDs found something, they would quit.

Sisyphean Quants

- Firms directing quants to work in silos, or to develop individual strategies, are asking the impossible.
- Identifying new strategies requires specialized teams working together.



The Meta-Strategy Paradigm (1/3)

- The complexities involved in developing a true investment strategy are overwhelming:
  - Data collection, curation, processing, structuring,
  - HPC infrastructure,
  - software development,
  - feature analysis,
  - execution simulators,
  - backtesting, etc.
- Even if the firm provides you with shared services in those areas, you are like a worker at a BMW factory who has been asked to build the entire car alone, by using all the workshops around you.
  - One week you need to be a master welder, another week an electrician, another week a mechanical engineer, another week a painter, ... try, fail and circle back to welding. It is a futile endeavor.

The Meta-Strategy Paradigm (2/3)

- It takes almost as much effort to produce one true investment strategy as to produce a hundred.
- Every successful quantitative firm I am aware of applies the meta-strategy paradigm.
- Your firm must set up a research factory
  - where tasks of the assembly line are clearly divided into subtasks.
  - where quality is independently measured and monitored for each subtask.
  - where the role of each quant is to specialize in a particular subtask, to become the best there is at it, while having a holistic view of the entire process.
- This is how Berkeley Lab and other U.S. National laboratories routinely make scientific discoveries, such as adding 16 elements to the periodic table, or laying out the groundwork for MRIs and PET scans: https://youtu.be/G5nK3B5uuY8

The Meta-Strategy Paradigm (3/3)

Practical Application	Classic approach	Quantitative Meta-Strategy
Selection & Hiring **(Example 1)**	Interview candidates with SR (or any other performance statistic) and track record length above a given threshold. Pros: Trivial to implement. Cons: Unknown (possibly high) probability of hiring unskilled PMs.	Design an interview process that recognizes the variables that affect the probability of making the wrong hire: • False positive rate. • False negative rate. • Skill-to-unskilled odds ratio. • Number of independent trials. • Sampling mechanism. Pros: It is objective and can be improved over time, based on measurable outcomes. Cons: More laborious.
Oversight **(Example 2)**	Allocate capital as if PMs were asset classes. Pros: Trivial to implement. Cons: Correlations are unstable, meaningless. Risks are likely to be concentrated.	Recognize that PMs styles evolve over time, as they adapt to a changing environment. Pros: It provides an early signal while the style is still emerging. Allocations can be revised before it is too late. Cons: Allocation revisions may be needed on an irregular calendar frequency.
Stop-Out **(Example 3)**	Stop-out a PM once a certain loss limit has been exceeded. Pros: Trivial to implement. Cons: It allows preventable problems to grow until it is too late.	For any drawdown, large or small, determine the expected time underwater and monitor every recovery. Even if a loss is small, a failure to recover within the expected timeframe indicates a latent problem. Pros: Proactive. Address problems before they force a stop-out. Cons: PMs may feel under tighter scrutiny.

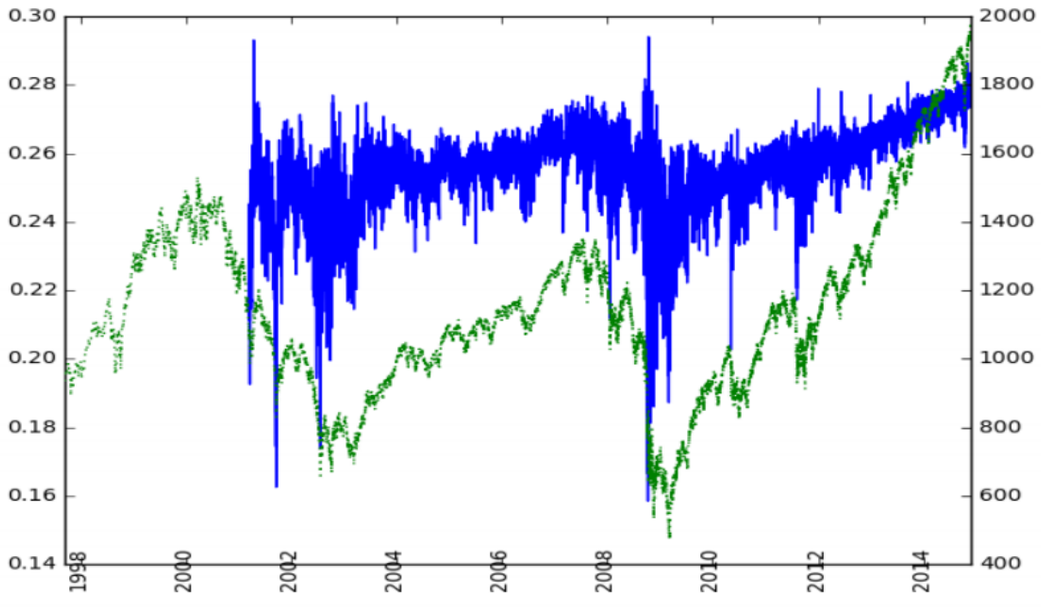
Pitfall #2:
Integer Differentiation

The Stationarity vs. Memory Dilemma

- In order to perform inferential analyses, researchers need to work with invariant processes, such as returns on prices (or changes in log-prices), changes in yield, changes in volatility, …
- These operations make the series stationary, at the expense of removing all memory from the original series.
- Memory is the basis for the model's predictive power.
  - For example, equilibrium (stationary) models need some memory to assess how far the price process has drifted away from the long-term expected value in order to generate a forecast.
- The dilemma is
  - returns are stationary however memory-less; and
  - prices have memory however they are non-stationary.

The Optimal Stationarity-Memory Trade Off

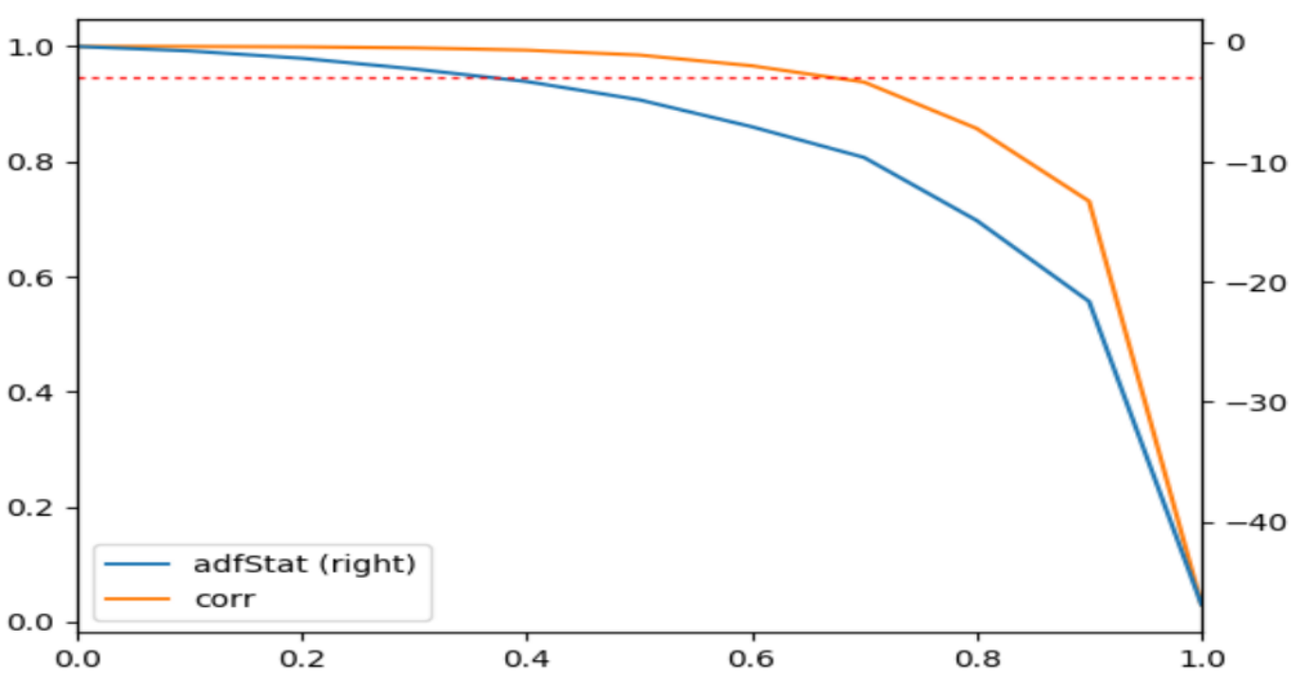
- Question: What is the minimum amount of differentiation that makes a price series stationary while preserving as much memory as possible?
- Answer: We would like to generalize the notion of returns to consider stationary series where not all memory is erased.
- Under this framework, returns are just one kind of (and in most cases suboptimal) price transformation among many other possible.



- Green line: E-mini S&P 500 futures trade bars of size 1E4
- Blue line: Fractionally differentiated ($d = .4$)
- Over a short time span, it resembles returns
- Over a longer time span, it resembles price levels

Eg.: E-mini S&P 500 Futures

- On the x-axis, the *d* value used to generate the series on which the ADF stat was computed.
- On the left y-axis, the correlation between the original series ($d = 0$) and the differentiated series at various *d* values.
- On the right y-axis, ADF stats computed on log prices.



The original series ($d = 0$) has an ADF stat of -0.3387, while the returns series ($d = 1$) has an ADF stat of -46.9114.

At a 95% confidence level, the test's critical value is -2.8623.

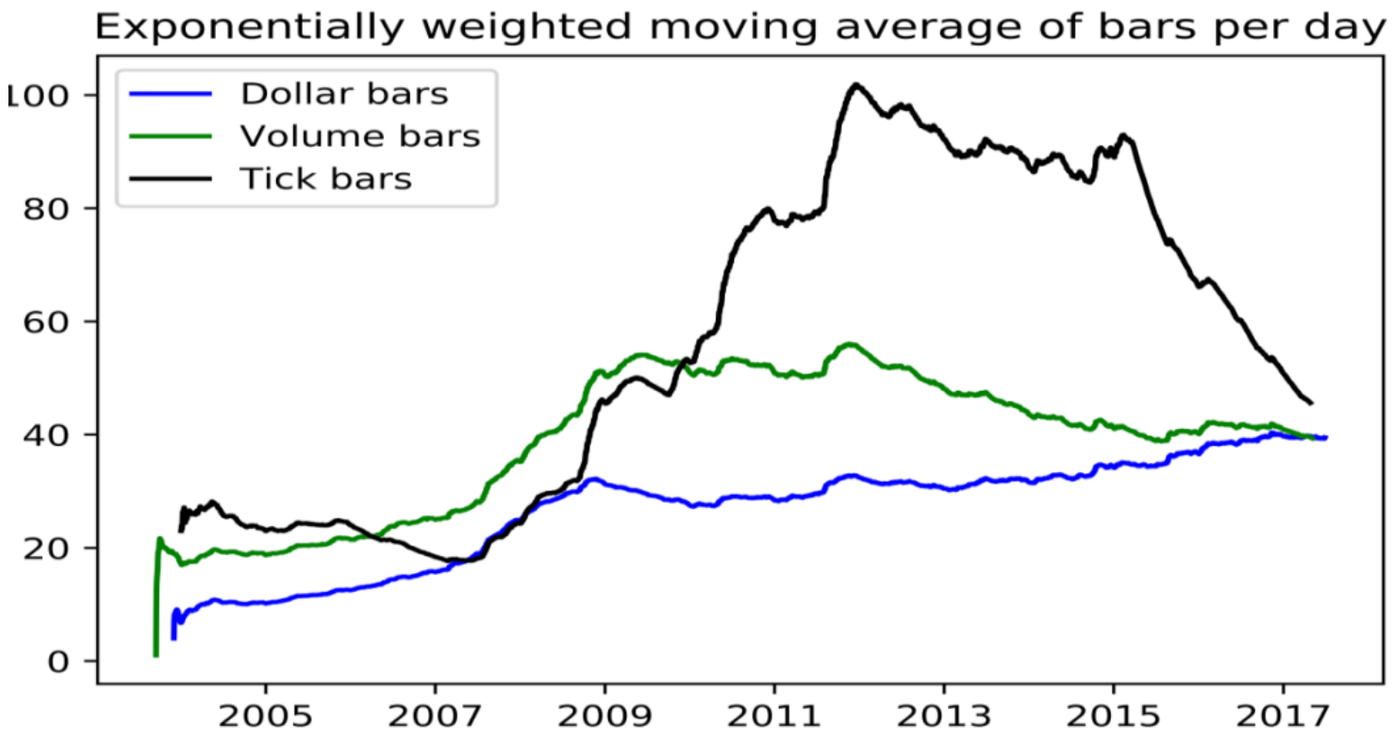
The ADF stat crosses that threshold in the vicinity of $d = 0.35$, where correlation is still very high (0.995).

Pitfall #3:
Inefficient Sampling

Chronological Sampling

- Information does not arrive to the market at a constant entropy rate.
- Sampling data in chronological intervals means that the informational content of the individual observations is far from constant.
- A better approach is to sample observations as a subordinated process of the amount of information exchanged:
  - Trade bars.
  - Volume bars.
  - Dollar bars.
  - Volatility or runs bars.
  - Order imbalance bars.
  - Entropy bars.

Sampling Frequencies



Three bar types computed on E-mini S&P 500 futures.

Tick bars tend to exhibit a wide range of sampling frequencies, for multiple microstructural reasons.

Sampling frequencies for **volume bars** are often inversely proportional to price levels.

In general, **dollar bars** tend to exhibit more stable sampling frequencies.

Pitfall #4:
Wrong Labeling

Labeling in Finance

- Virtually all ML papers in finance label observations using the fixed-time horizon method.
- Consider a set of features $\{X_i\}_{i=1,\ldots,I}$, drawn from some bars with index $t = 1, \ldots, T$, where $I \leq T$. An observation X_i is assigned a label $y_i \in \{-1,0,1\}$,

$$y_i = \begin{cases} -1 & \text{if } r_{t_{i,0},t_{i,0}+h} < -\tau \\ 0 & \text{if } \left|r_{t_{i,0},t_{i,0}+h}\right| \leq \tau \\ 1 & \text{if } r_{t_{i,0},t_{i,0}+h} > \tau \end{cases}$$

where τ is a pre-defined constant threshold, $t_{i,0}$ is the index of the bar immediately after X_i takes place, $t_{i,0} + h$ is the index of *h* bars after $t_{i,0}$, and $r_{t_{i,0},t_{i,0}+h}$ is the price return over a bar horizon *h*.

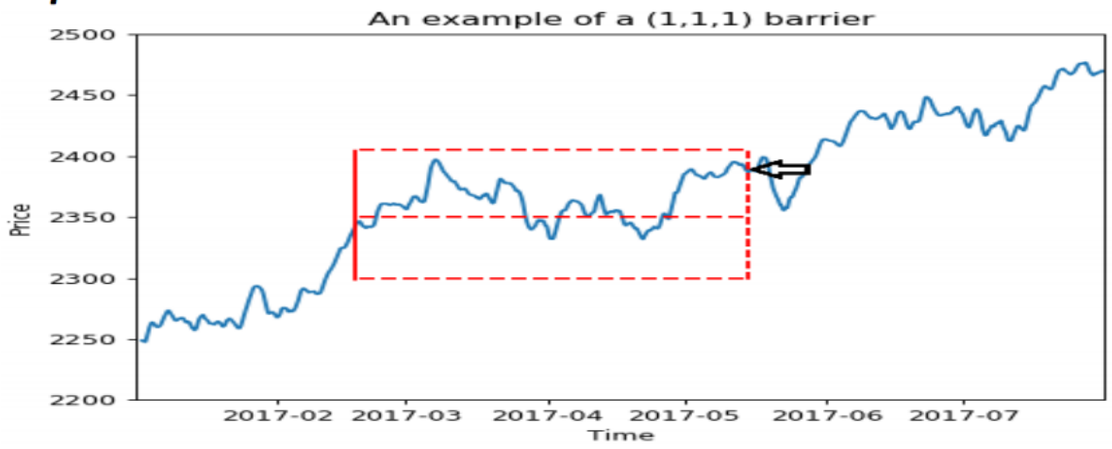
- Because the literature almost always works with time bars, *h* implies a fixed-time horizon.

Caveats of the Fixed Horizon Method

- There are several reasons to avoid such labeling approach:
  - Time bars do not exhibit good statistical properties.
  - The same threshold τ is applied regardless of the observed volatility.
    - Suppose that $\tau = 1E-2$, where sometimes we label an observation as $y_i = 1$ subject to a realized bar volatility of $\sigma_{t_{i,0}} = 1E-4$ (e.g., during the night session), and sometimes $\sigma_{t_{i,0}} = 1E-2$ (e.g., around the open). The large majority of labels will be 0, even if return $r_{t_{i,0},t_{i,0}+h}$ was predictable and statistically significant.
- A couple of better alternatives would be:
  - Label per a varying threshold $\sigma_{t_{i,0}}$, estimated using a rolling exponentially-weighted standard deviation of returns.
  - Use volume or dollar bars, as their volatilities are much closer to constant (homoscedasticity).
- A key flaw of the fixed-time horizon method: It ignores the *path* followed by prices. We will address this with the Triple Barrier Method.

The Triple Barrier Method

- It is simply unrealistic to build a strategy that profits from positions that would have been stopped-out by the fund, exchange (margin call) or investor.
- The Triple Barrier Method labels an observation according to the first barrier touched out of three barriers.
  - Two horizontal barriers are defined by profit-taking and stop-loss limits, which are a dynamic function of estimated volatility (whether realized or implied).
  - A third, vertical barrier, is defined in terms of number of bars elapsed since the position was taken (an expiration limit).
- The barrier that is touched first by the *price path* determines the label:
  - Upper horizontal barrier: Label 1.
  - Lower horizontal barrier: Label -1.
  - Vertical barrier: Label 0.



Pitfall #5:

Weighting of non-IID samples

The “spilled samples” problem (1/2)

- Most non-financial ML researchers can assume that observations are drawn from IID processes. For example, you can obtain blood samples from a large number of patients, and measure their cholesterol.
- Of course, various underlying common factors will shift the mean and standard deviation of the cholesterol distribution, but the samples are still independent: There is one observation per subject.
- Suppose you take those blood samples, and someone in your laboratory spills blood from each tube to the following 9 tubes to their right.
  - That is, tube 10 contains blood for patient 10, but also blood from patients 1 to 9. Tube 11 contains blood from patient 11, but also blood from patients 2 to 10, and so on.

The “spilled samples” problem (2/2)

- Now you need to determine the features predictive of high cholesterol (diet, exercise, age, etc.), without knowing for sure the cholesterol level of each patient.
- That is the equivalent challenge that we face in financial ML.
  - Labels are decided by outcomes.
  - Outcomes are decided over multiple observations.
  - Because labels overlap in time, we cannot be certain about what observed features caused an effect.

Pitfall #6:
Cross-Validation (CV) Leakage

Why standard CV fails in Finance

- One reason k-fold CV fails in finance is because observations cannot be assumed to be drawn from an IID process.
- *Leakage* takes place when the training set contains information that also appears in the testing set.
- Consider a serially correlated feature X that is associated with labels Y that are formed on overlapping data:
  - Because of the serial correlation, $X_t \approx X_{t+1}$.
  - Because labels are derived from overlapping data points, $Y_t \approx Y_{t+1}$.
- Then, placing *t* and *t+1* in different sets leaks information.
  - When a classifier is first trained on (X_t, Y_t), and then it is asked to predict $\mathrm{E}[Y_{t+1}]$ based on an observed X_{t+1}, this classifier is more likely to achieve $Y_{t+1} = \mathrm{E}[Y_{t+1}]$ even if X is an irrelevant feature.
- In the presence of irrelevant features, leakage leads to false discoveries.

Purged K-Fold CV

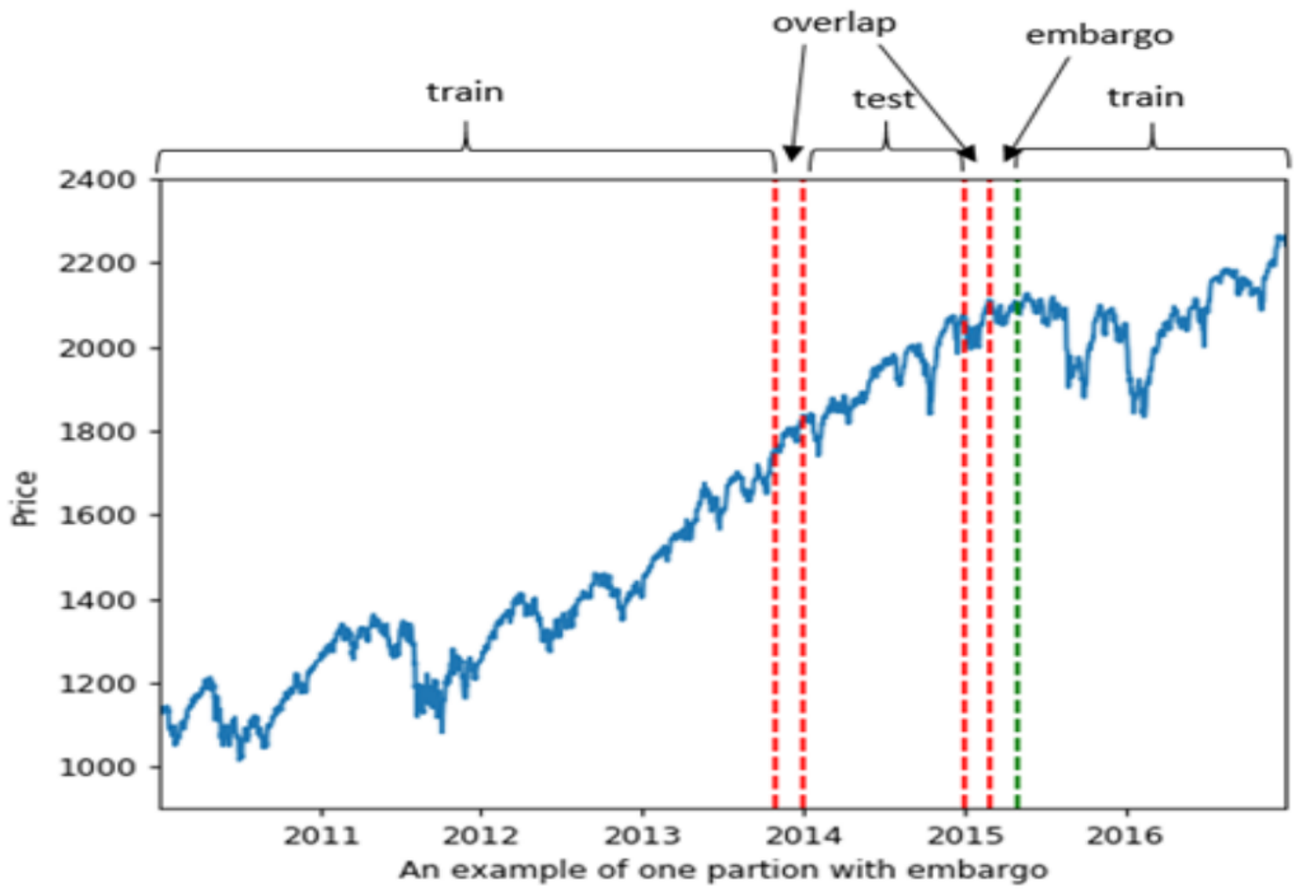
- One way to reduce leakage is to purge from the training set all observations whose labels overlapped in time with those labels included in the testing set. I call this process *purging*.
- Consider a label Y_j that is a function of observations in the closed range $t \in [t_{j,0}, t_{j,1}]$, $Y_j = f\left[[t_{j,0}, t_{j,1}]\right]$.
  - For example, in the context of the triple barrier labeling method, it means that the label is the sign of the return spanning between price bars with indices $t_{j,0}$ and $t_{j,i}$, that is $\text{sgn}\left[r_{t_{j,0},t_{j,1}}\right]$.
- A label $Y_i = f\left[[t_{j,0}, t_{j,1}]\right]$ overlaps with Y_j if any of the three sufficient conditions is met:

$$t_{j,0} \leq t_{i,0} \leq t_{j,1}; t_{j,0} \leq t_{i,1} \leq t_{j,1};\ t_{i,0} \leq t_{j,0} \leq t_{j,1} \leq t_{i,1}$$

Embargoed K-Fold CV

- Since financial features often incorporate series that exhibit serial correlation (like ARMA processes), we should eliminate from the training set observations that immediately follow an observation in the testing set. I call this process *embargo*.
  - The embargo does not need to affect training observations prior to a test, because training labels $Y_i = f\left[\left[t_{i,0}, t_{i,1}\right]\right]$, where $t_{i,1} < t_{j,0}$ (training ends before testing begins), contain information that was available at the testing time $t_{j,0}$.
  - We are only concerned with training labels $Y_i = f\left[\left[t_{i,0}, t_{i,1}\right]\right]$ that take place immediately after the test, $t_{j,1} \leq t_{i,0} \leq t_{j,1} + h$.
- We can implement this embargo period *h* by setting $Y_j = f\left[\left[t_{j,0}, t_{j,1} + h\right]\right]$ before purging. A small value $h \approx .01T$, where T is the number of bars, often suffices to prevent all leakage.

Eg.: Purging and Embargoing

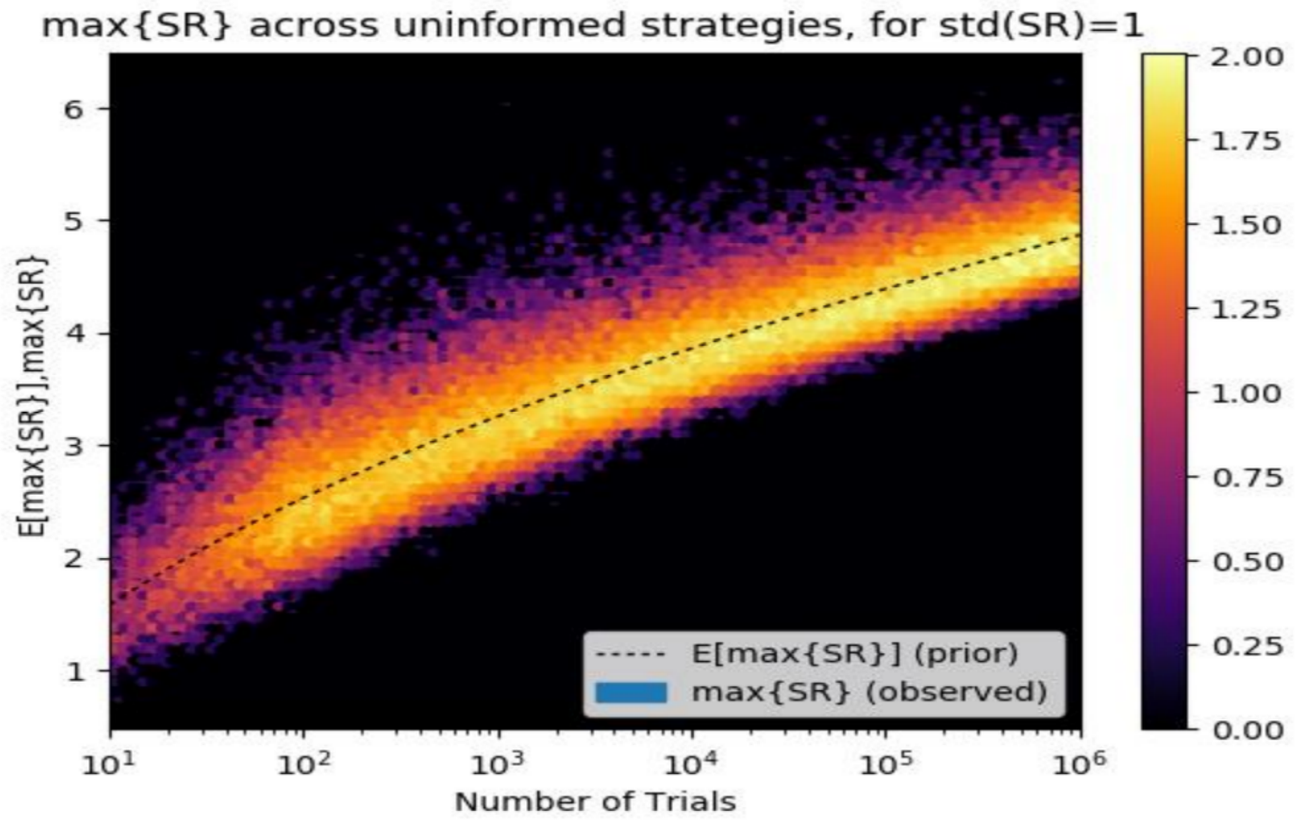


This plot shows one partition of the K-Fold CV. The test set is surrounded by two train sets, generating two overlaps that must be purged to prevent leakage.

To further prevent leakage, the train observations immediately after the testing set are also embargoed.

Pitfall #7:
Backtest Overfitting

The Most Important Plot in Finance



The y-axis displays the distribution of the maximum Sharpe ratios (max{SR}) for a given number of trials (x-axis). A lighter color indicates a higher probability of obtaining that result, and the dash-line indicates the expected value. For example, after only 1,000 independent backtests, the expected maximum Sharpe ratio (E[max{SR}]) is 3.26, even if the true Sharpe ratio of the strategy is zero!

The reason is *Backtest Overfitting*: When selection bias (picking the best result) takes place under multiple testing (running many alternative configurations), that backtest is likely to be a false discovery. **Most quantitative firms invest in false discoveries**.

The "False Strategy" Theorem [2014]

- Given a sample of IID-Gaussian Sharpe ratios, $\{\widehat{SR}_k\}$, $k = 1, \ldots, K$, with $\widehat{SR}_k \sim \mathcal{N}\left[0, \mathrm{V}\left[\{\widehat{SR}_k\}\right]\right]$, then

$$\mathrm{E}\left[\max_k\{\widehat{SR}_k\}\right]\left(\mathrm{V}\left[\{\widehat{SR}_k\}\right]\right)^{-1/2} \approx$$
$$(1-\gamma)Z^{-1}\left[1-\frac{1}{K}\right] + \gamma Z^{-1}\left[1-\frac{1}{Ke}\right]$$

where $Z^{-1}[.]$ is the inverse of the standard Gaussian CDF, e is Euler's number, and γ is the Euler-Mascheroni constant.

- Corollary: Unless $\max_k\{\widehat{SR}_k\} \gg \mathrm{E}\left[\max_k\{\widehat{SR}_k\}\right]$, the discovered strategy is likely to be a *false positive*.

Source: López de Prado et al. (2014): "The effects of backtest overfitting on out-of-sample performance." *Notices of the American Mathematical Society*, 61(5), pp. 458-471.