Finovis

AI FOR THE FINANCIAL INDUSTRY

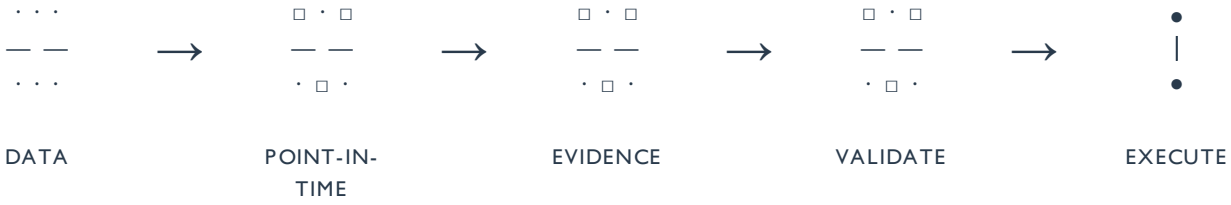
FINOVIS RESEARCH

DATA & SIGNALS · RP-002

From Noise to Signal

Identifying Information in Financial Market Data

Critical Literature Review and Analytical Synthesis



June 2026

finovis.ai

Executive Summary

Financial markets generate a continuous stream of observable data: prices, returns, trading activity, volatility, macroeconomic information and an expanding universe of derived variables. Yet the availability of observations does not by itself establish the presence of useful predictive information. For quantitative research, the central challenge is therefore not simply to identify patterns in historical data, but to determine whether an observed relationship contains information that is sufficiently distinct, robust and economically relevant to warrant being treated as a signal.

This distinction matters because financial datasets permit many relationships to be observed ex post. Some may reflect genuine predictive structure. Others may arise from random variation, imperfect measurement, sample construction, repeated hypothesis testing or excessive model flexibility. As the research space expands, both possibilities increase. Larger datasets and more flexible methods can reveal information that simpler approaches may miss, but they also enlarge the opportunity to fit historical noise. Evidence from high-dimensional U.S. equity-return prediction illustrates this tension: Gu, Kelly and Xiu (2020) find that flexible methods can improve out-of-sample forecasts when model complexity is appropriately controlled, while an unrestricted expansion of a linear predictor set performs poorly in their empirical setting.

The forecasting literature likewise demonstrates why historical fit cannot be equated with predictive reliability. Welch and Goyal (2008) documented weak out-of-sample performance for many established U.S. equity-premium predictors. Goyal, Welch and Zafirov (2024), extending the evidence through 2021 and evaluating both earlier and more recently proposed variables, again find that many predictors weaken when samples are extended, although a smaller subset retains useful evidence. Other research reaches more constructive conclusions. Campbell and Thompson (2008) show that economically motivated restrictions can improve forecasts relative to the same historical-average benchmark, while Li et al. (2025) develop an alternative forecasting approach and report substantially stronger out-of-sample evidence for equity-premium predictability. Taken together, these studies show that conclusions about predictability can depend materially on estimation procedure, forecast construction and evaluation design. They do not support a simple conclusion that financial prediction is either generally reliable or generally impossible.

The evidentiary threshold becomes particularly important when many potential relationships are examined. Data-snooping concerns in financial research have been documented for decades. Lo and MacKinlay (1990) show that tests of financial asset-pricing models can produce misleading inference when properties of the data are used in constructing the test statistics themselves. Harvey, Liu and Zhu (2016) subsequently argue that conventional significance thresholds become inadequate when the effective number of tested return factors is large. These findings do not establish that every discovered relationship is false. They establish that the strength of evidence required to support a predictive claim depends partly on the research process through which that claim was discovered.

Statistical significance itself must also be interpreted carefully. As emphasised by Wasserstein and Lazar (2016) in the American Statistical Association's statement on p-values, a p-value can indicate how incompatible observed data are with a specified statistical model under its assumptions; it does not measure the probability that a hypothesis is true, the probability that a result occurred by random chance alone, or the economic importance of an effect. Statistical evidence is therefore one component of signal qualification rather than its endpoint.

This paper treats signal identification as a sequence of evidentiary tests. The underlying data must first represent information that was actually available at the relevant decision time. The predictive problem must then be defined: what is being forecast, over what horizon, using which information set and relative to which benchmark. A candidate relationship must subsequently demonstrate more than contemporaneous association or in-sample fit. It must provide incremental forecasting information, withstand appropriate robustness and out-of-sample evaluation, and retain economic relevance once the conditions required to act on it are considered.

The distinction between statistical and economic significance is particularly important. A statistically detectable relationship may be too small to matter economically, while a modest improvement in forecasting accuracy may potentially be valuable in a particular investment setting. Whether predictive information is ultimately usable depends additionally on turnover, transaction costs, liquidity, market impact, capacity, execution timing and the speed with which the information may decay.

Accordingly, this paper uses the following conceptual progression:

RAW MARKET OBSERVATIONS

→ POINT-IN-TIME DATA

→ DEFINED PREDICTION PROBLEM

→ MEASUREMENT & PREPROCESSING

→ CANDIDATE PATTERN

→ STATISTICAL EVIDENCE

→ INCREMENTAL PREDICTIVE INFORMATION

→ ROBUSTNESS & REPLICATION

→ ECONOMIC RELEVANCE

→ EXECUTABILITY

The framework does not prescribe a universal research algorithm. Different markets, frequencies and forecasting problems require different methods. Its purpose is instead to separate stages that are frequently conflated. An observed pattern is not automatically a statistical signal; a statistically supported relationship is not automatically predictive; a predictive relationship is not automatically robust; and a robust predictor is not automatically economically executable.

Key Findings

1. More data expand both the opportunity to identify information and the opportunity to fit noise. Increasing data volume and model flexibility therefore increases the importance of research discipline rather than reducing it.

2. A signal can only be assessed relative to a defined predictive problem. The target, forecast horizon, available information set and benchmark form part of the signal definition itself.

3. Data construction is part of inference. Point-in-time integrity, sampling, preprocessing and universe construction can materially affect the relationships subsequently identified.

4. Statistical association is not equivalent to incremental predictive information. Correlation or in-sample significance may disappear once existing information, alternative specifications or the broader search process are considered.

5. Financial relationships may be genuine yet unstable. Changes in economic conditions, market structure, participation and regimes can alter the strength or relevance of previously documented relationships.

6. Out-of-sample evidence and independent replication materially strengthen a predictive claim, but neither is automatically conclusive. Their value depends on whether validation remains genuinely separated from model discovery and modification.

7. Predictive information becomes economically usable only when implementation is considered. Assessment must therefore extend beyond statistical evidence to transaction costs, turnover, liquidity, market impact, capacity, timing and signal decay.

The central implication is that quantitative research should not ask only whether a relationship can be found in financial data. It should ask what evidence is required for that relationship to survive the transition from observation to executable information.

Table of Contents

1. The Signal Identification Problem

1.1 Markets Produce Observations, Not Automatically Information

Financial markets are observable through many dimensions. Prices and returns provide a direct record of market outcomes, while volume, volatility, spreads, order flow, fundamentals, macroeconomic variables and derived measures describe additional aspects of market conditions. These observations can contain economically meaningful structure, but they also contain random variation, measurement imperfections and relationships that may not persist beyond the period in which they are identified.

The existence of a historical pattern is therefore an observation, not a conclusion about predictability. Financial research provides numerous examples in which relationships detectable within one sample become substantially weaker when evaluated using subsequent observations or alternative specifications. Welch and Goyal (2008), for example, find limited out-of-sample forecasting success among many established U.S. equity-premium predictors. Goyal, Welch and Zafirov (2024) later reassess 29 variables proposed in subsequent research, alongside the earlier predictor set, and extend the underlying data through 2021. More than one-third of the newer variables no longer retain in-sample significance in the extended data. Among those that do, approximately half exhibit poor out-of-sample performance, while a small number perform reasonably well both in and out of sample.

These findings do not imply that return prediction is inherently impossible. The literature contains competing evidence. Campbell and Thompson (2008) show that economically motivated restrictions on predictive regressions can produce improvements relative to the historical-average forecast, despite relatively small gains in out-of-sample explanatory power. More recently, Li et al. (2025) propose a forecasting methodology that uses a conservative nonzero predictive slope to reduce bias relative to the historical-average forecast while limiting the estimation variance associated with conventional predictive regressions. Their results provide substantially stronger real-time out-of-sample evidence across a broad set of equity-premium predictors.

The contrast is important. It demonstrates that evidence about predictability can depend not only on the predictor itself, but also on how forecasts are constructed, estimated and evaluated. Robust signal identification should therefore examine sensitivity to reasonable methodological choices rather than treating the outcome of a single specification as definitive.

A useful distinction for this paper is between observation, pattern and predictive information.

An observation is a recorded market or economic quantity.

A pattern is a systematic relationship identified among observations under a specified analytical procedure.

Predictive information exists when a relationship provides incremental forecasting information for a defined target and horizon relative to a stated benchmark and evaluation criterion.

This distinction is necessary because a pattern can be descriptively real without being useful for prediction. Two variables may move together contemporaneously without one containing information about the future value of the other. A historical coefficient may be statistically distinguishable from zero in one sample while providing no improvement in subsequent forecasts. A variable may also contain information that is already captured by another predictor or benchmark and therefore offer no incremental value.

For this reason, the term signal is used cautiously throughout this paper. A candidate signal is not simply a variable associated with market outcomes. It is a measurable relationship proposed to contain information about a defined future outcome. Whether that relationship deserves further qualification depends on the evidence accumulated through the research process.

1.2 Defining the Predictive Problem

Signal identification begins before a model is estimated. A predictive claim has meaning only after the forecasting problem has been specified.

Four elements are fundamental:

Target. What quantity is being predicted?

Horizon. At what future interval is that target evaluated?

Information Set. What information was legitimately observable when the forecast would have been formed?

Benchmark. Relative to what existing forecast or information standard is predictive improvement assessed?

The target determines the object of prediction. Forecasting returns is a different problem from forecasting volatility or liquidity. Even within return prediction, forecasting the aggregate market risk premium and forecasting relative returns across individual securities represent distinct research problems. Gu, Kelly and Xiu (2020), for example, separately examine cross-sectional stock-return prediction and time-series prediction of the aggregate U.S. equity premium.

The forecast horizon also forms part of the hypothesis. Information need not remain equally relevant across intervals. A relationship that contains information at one horizon may contain little at another. Changing the horizon therefore changes the predictive question rather than merely adjusting a technical parameter.

The information set determines what could actually have been known at forecast formation. A relationship identified using data that became available only after the relevant decision time cannot constitute genuine historical predictive evidence for that decision. Reporting lags, revisions, historical index membership and corporate-event timestamps can therefore materially alter the validity of a backtest. These issues are examined in Section 2.

Finally, the benchmark determines whether the candidate adds forecasting information. A positive historical fit is insufficient if a simpler forecast performs equally well or better. Welch and Goyal (2008), Campbell and Thompson (2008) and Li et al. (2025) all evaluate equity-premium prediction relative to the historical-average forecast, yet reach different conclusions under different estimation and forecast-construction procedures. The comparison illustrates an important point: benchmark selection and forecasting methodology are separate components of the evaluation design.

Consider a purely illustrative case in which a variable is positively associated with next-period returns in historical data. That association does not by itself establish a useful signal. It may fail to improve upon a simple benchmark. Its apparent information may disappear once another predictor is included. It may rely on observations unavailable at forecast formation, or its predictive content may exist only at a particular horizon. None of these outcomes necessarily invalidates the original descriptive relationship; each changes the predictive claim that can reasonably be made from it.

Defining the predictive problem before evaluating the evidence therefore limits the scope for retrospectively adjusting the target, horizon, information set or benchmark to accommodate whichever historical specification happens to perform best.

1.3 From Pattern to Signal

Once the predictive problem has been defined, a candidate pattern can be evaluated progressively.

The first transition is from pattern to statistical evidence. A relationship visible in a chart, summary statistic or historical comparison may motivate investigation, but descriptive regularity alone does not establish how compatible the observed result is with an appropriate statistical model.

This distinction requires careful interpretation. As emphasised by Wasserstein and Lazar (2016), a p-value can indicate how incompatible observed data are with a specified statistical model, conditional on the assumptions underlying the calculation. It does not measure the probability that the tested hypothesis is true, nor the probability that the data resulted from random chance alone. Statistical significance also does not measure effect size or economic importance. Statistical inference should therefore be understood as one source of evidence within a broader research design rather than as a binary certification that a signal exists.

The second transition is from statistical evidence to incremental predictive information. A relationship may be statistically detectable while contributing little forecasting value relative to existing information. Redundancy becomes particularly important in high-dimensional settings, where candidate features may be strongly related to one another.

Gu, Kelly and Xiu (2020) provide an instructive empirical example from U.S. equity-return prediction. In their specific dataset and research design, expanding an ordinary least-squares specification to more than 900 baseline predictors produces poor out-of-sample performance,

while methods incorporating regularisation, dimension reduction and nonlinear structure perform substantially better. The finding should not be interpreted as a universal ranking of modelling approaches. It illustrates the more general problem that expanding the predictor set can increase both the information available to a model and its capacity to fit sample-specific variation.

The third transition is from predictive information to robustness. A relationship discovered in one sample is exposed to sampling variation, specification choices and the market conditions prevailing during the research period. Confidence increases when the result survives appropriately separated data, reasonable alternative specifications and, where feasible, independent replication.

Robustness does not require a relationship to remain numerically identical across every period or environment. Financial relationships may themselves be conditional. The relevant question is whether the accumulated evidence remains consistent with the scope of the predictive claim being made.

The final stages concern economic relevance and executability. A forecast can contain measurable predictive information without being economically useful. Effect size, turnover, transaction costs, liquidity, market impact, execution timing and capacity can determine whether information survives implementation. These considerations answer a different question from statistical discovery: not merely whether information exists, but whether it can be acted upon under realistic conditions.

The resulting hierarchy is:

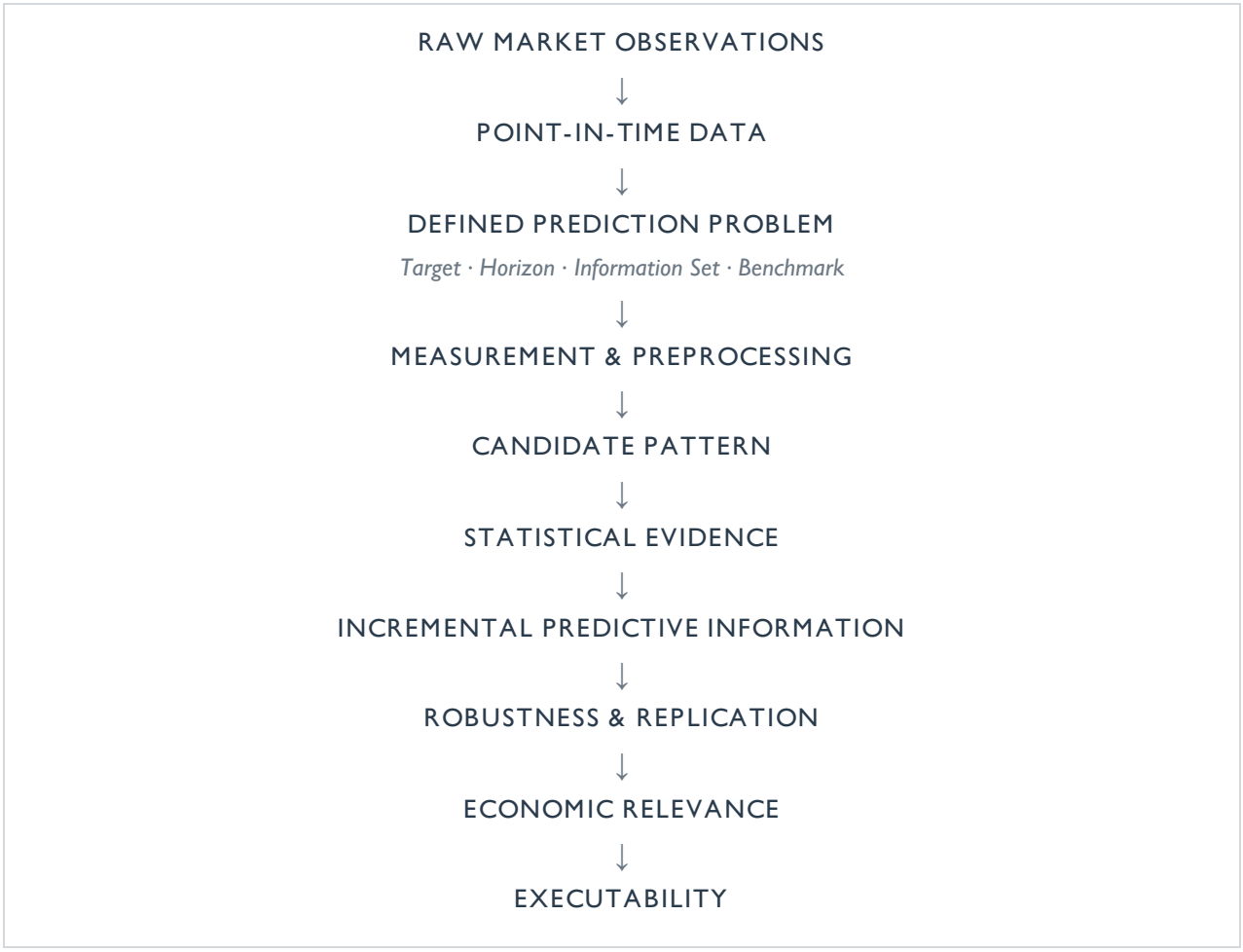


Figure 1. Signal Qualification Framework.

Conceptual progression from observable market data to economically executable information. The framework represents stages of qualification rather than a requirement that every research problem use an identical methodology.

Failure at a later stage does not necessarily imply that an earlier observation was false. A relationship may exist historically but fail to generalise. A predictor may contain information that is too weak to implement economically. A previously informative relationship may become less useful when market conditions change. The framework therefore distinguishes the nature and strength of the evidence at each stage rather than reducing signal identification to a single pass-or-fail test.

It also distinguishes prediction from explanation. A variable may improve forecasts without establishing why the relationship exists. Conversely, an economically plausible narrative does not establish predictive power. Economic reasoning can guide hypothesis formation, interpretation and robustness testing, but it cannot substitute for empirical validation.

For this paper, a qualified predictive signal is therefore defined as:

a measurable relationship that provides incremental forecasting information for a defined target and horizon, using only information available at forecast formation, and that demonstrates sufficient robustness to justify evaluation of its economic relevance.

A qualified predictive signal is not yet an executable signal. The latter requires the additional economic and implementation tests represented by the final stages of the framework.

1.4 Why More Data Can Produce More Opportunity—and More False Discovery

The expansion of financial datasets creates genuine research opportunities. Additional observations can improve estimation. New variables can capture dimensions of markets that were previously difficult to measure. Greater computational capacity enables researchers to investigate interactions, nonlinear relationships and high-dimensional information sets at substantially greater scale.

The same expansion also enlarges the search problem.

When researchers evaluate more variables, transformations, horizons, asset universes, parameter combinations or model specifications, the effective number of candidate relationships increases. Even when individual tests are correctly calculated, a sufficiently broad search can produce apparently successful historical results that do not generalise.

This problem predates contemporary large-scale computing. Lo and MacKinlay (1990) show that tests of financial asset-pricing models can produce misleading inference when properties of the data are used in constructing the test statistics themselves. Harvey, Liu and Zhu (2016) later examine the problem in the context of the expanding cross-section of expected-return factors and argue that conventional significance thresholds should be raised when the effective number of hypotheses under consideration is large.

The implication is not that large datasets or flexible models are inherently unreliable. High dimensionality creates both an opportunity and a risk. It may expose predictive structure omitted by simpler specifications, but it also increases the capacity of the research process to reproduce sample-specific variation.

Gu, Kelly and Xiu (2020) illustrate both sides of this trade-off in their U.S. equity application. Their results show that a broad predictor set can contain useful forecasting information when complexity is appropriately controlled, while unrestricted linear expansion performs poorly out of sample. The appropriate response to increased data availability is therefore not automatically to prefer either simpler or more complex models. It is to strengthen the procedures used to determine whether the resulting performance generalises.

More data can also mean different things. More observations of the same phenomenon may increase precision. More features can broaden the information set while introducing redundancy. More frequent observations can change the statistical and measurement characteristics of the data. More model specifications enlarge the opportunity to represent complex relationships but also expand the effective search space.

These distinctions are especially relevant in many return-prediction settings, where the predictable component can be small relative to unpredictable variation in realised returns. Under

such conditions, useful information may be difficult to isolate while sample-specific relationships remain comparatively easy to discover.

More observable data therefore do not automatically create more predictive information or more economically usable investment opportunities. They enlarge the opportunity set from which information might be identified. Whether an observed relationship survives as evidence depends on the integrity of the data, the definition of the predictive problem, the research search process and the strength of subsequent validation.

The next stage of the analysis begins with the first of these requirements: the data themselves.

2. Data Integrity, Measurement and Sampling

A predictive model can only be as credible as the information set on which its historical evaluation is based. This requirement extends beyond whether individual observations are numerically accurate. Quantitative research must also establish whether those observations existed in the recorded form at the relevant decision time, whether the historical sample represents the opportunity set that would actually have been available, and whether the transformation and sampling of the data preserve rather than distort the object being studied.

These issues are consequential because a backtest reconstructs a decision process retrospectively. The researcher observes the past with knowledge that a market participant living through that period did not possess. Securities that subsequently failed are known; macroeconomic series may have been revised; final corporate disclosures are available; index membership is documented retrospectively; and dense market data can be reorganised at frequencies or timestamps that may not correspond to the original information flow. Unless this informational advantage is controlled, the historical research environment can become materially different from the environment in which the hypothetical decisions would actually have been made.

Data quality in quantitative finance should therefore be understood as `information-set integrity`, not merely database accuracy.

2.1 Point-in-Time Data and Look-Ahead Bias

The fundamental point-in-time question is simple:

Was this information available, in this form, at the moment when the hypothetical forecast or decision was made?

A dataset can contain internally accurate or currently recorded historical values and still fail this test.

Macroeconomic data provide a clear example. Measures of output, employment and other economic quantities are frequently revised after their initial publication. A current historical database may therefore contain values that differ from those available to market participants in real time. Croushore and Stark (2001) construct a dataset of successive vintages of U.S. macroeconomic information precisely to preserve this distinction and demonstrate that data revisions can affect forecasting results. Their work illustrates that the historical value eventually recorded in a database and the value observable at the original forecast date are not necessarily the same informational object.

The same principle extends beyond revised macroeconomic series. Corporate accounting information relates to an economic reporting period, but the period end is not the same as the date on which the information became public. Index or universe membership observed retrospectively may differ from the membership known at an earlier date. Corporate actions may be recorded in databases according to effective dates that differ from announcement dates.

Historical data can therefore be chronologically accurate while remaining unsuitable for a point-in-time simulation if the availability timestamp is wrong.

Look-ahead bias occurs when information that would not yet have been known is allowed to influence an earlier simulated decision. Its most obvious form is direct use of a future outcome. More subtle forms arise when later revisions, corrected records or retrospectively reconstructed datasets are treated as though they had existed contemporaneously.

Consider an illustrative research design in which a model uses quarterly company fundamentals to forecast subsequent returns. Assigning the information to the fiscal quarter-end date would implicitly assume that the financial statements were available at that point. A point-in-time implementation instead requires the researcher to determine when the relevant disclosure became observable and to ensure that the forecast uses it only thereafter. The numerical values may be identical in both datasets; the information sets are not.

Point-in-time integrity therefore requires more than attaching a date to every observation. Where relevant, researchers must distinguish between several different concepts of time:

- the period or event to which the observation relates;
- the time at which the information was first released;
- the time at which it became available in the research or trading system;
- and the first point at which it could reasonably have influenced the decision being simulated.

The required precision depends on the research horizon. A difference of several hours may be immaterial for some low-frequency studies but decisive for an intraday strategy. Conversely, the appropriate response is not to impose maximum timestamp granularity on every dataset. It is to ensure that the temporal representation is sufficiently accurate for the predictive claim being evaluated.

A useful operational principle follows:

Historical research should reconstruct the information state that existed at the forecast date rather than the database state available to the researcher today.

This requirement forms the first barrier between retrospective pattern recognition and legitimate predictive evidence.

2.2 Survivorship and Sample Selection

Point-in-time integrity concerns what was known. A related problem concerns what was included.

Historical datasets can become selective because assets, companies, investment funds or other observations disappear from the population over time. If a research sample includes only entities that survived until a later date, the resulting history is conditioned on information that was not available ex ante: survival itself.

Brown, Goetzmann, Ibbotson and Ross (1992) demonstrate the potential severity of this problem in investment-performance research. In a sample truncated by survivorship, they show

theoretically and through numerical examples that the interaction between volatility and survival can create an appearance of return predictability. Their result is important because survivorship does not merely alter an average historical return. Under some research designs, it can alter the apparent relationships between variables and thereby generate misleading evidence of predictability.

Shumway (1997) documents a related but distinct problem in CRSP equity data. He finds that correct delisting returns were unavailable for many firms delisted for bankruptcy and other negative reasons and shows, using over-the-counter prices, that the omitted delisting returns were economically large. A return history that fails to account appropriately for the final losses associated with unsuccessful securities can therefore misrepresent the realised experience of the historical universe.

These examples illustrate two related concepts that should remain distinct.

Survivorship bias arises when inclusion in the historical sample depends on continued survival to a later date.

Sample-selection bias is broader. It arises when the mechanism determining which observations enter the analysis is systematically related to the outcome or relationship being investigated.

Survivorship is therefore one form of sample selection, but not the only one.

For signal research, the appropriate universe is generally the universe that would have satisfied the stated eligibility rules at each historical decision date, not the set of entities that remain visible or convenient to obtain at the end of the sample.

An illustrative equity study makes the distinction clear. Suppose a researcher identifies the current constituents of an index and obtains their ten-year histories to test a predictive characteristic. Even if every return and accounting observation is correct, the exercise conditions the sample on companies that ultimately remained in or reached the selected universe. A point-in-time study would instead reconstruct historical membership, including securities subsequently removed, acquired, delisted or otherwise absent from the current constituent set.

Importantly, this principle does not imply that every study must include every security that ever existed. Research universes may legitimately impose liquidity, market-capitalisation, listing-history or data-availability requirements. The methodological requirement is that these criteria be defined independently of outcomes that became known only later and, where they vary through time, be applied using contemporaneously available information.

The distinction can be summarised as follows:

A clean dataset is not necessarily an unbiased dataset.

Removing incomplete, failed or inconvenient observations may improve apparent data consistency while simultaneously changing the economic population represented by the sample.

2.3 Sampling Frequency, Dependence and Market Microstructure

The amount of information in a financial dataset also depends on how observations are sampled.

At first sight, higher-frequency data appear to offer an unambiguous advantage. If a price is observed thousands of times rather than once during a trading day, the researcher has more observations from which to infer market behaviour. High-frequency data have indeed enabled substantially richer measurement of quantities such as realised volatility and correlation. Andersen, Bollerslev, Diebold and Ebens (2001), for example, construct realised daily equity volatilities and correlations from intraday transaction prices and document strong temporal dependence in these measures.

However, increasing frequency also changes the measurement problem.

Observed transaction prices need not equal a continuously evolving latent efficient price at every instant. Bid–ask effects, price discreteness and other features of market microstructure can affect very high-frequency observations. Consequently, treating increasingly frequent price changes as though they represented independent and frictionless observations of the underlying process can distort statistical estimates.

Aït-Sahalia, Mykland and Zhang (2005) formalise this problem in the estimation of volatility. If market-microstructure noise is present but ignored, they show that an optimal finite sampling frequency arises: using every recorded observation naively is not necessarily optimal. Crucially, however, their conclusion is not that researchers should generally discard high-frequency information. When the noise is modelled explicitly, they show that using all available observations can be preferable, even under some misspecification of the noise distribution.

Hansen and Lunde (2006) provide further evidence that microstructure noise itself can have nontrivial properties. In their analysis of Dow Jones Industrial Average stocks, the noise is time-dependent and correlated with changes in the efficient price, with implications for volatility estimation from high-frequency observations.

The methodological implication is therefore more precise than the common proposition that "higher-frequency data contain more noise."

Increasing sampling frequency changes both the quantity of observations and the statistical structure of what is observed.

More frequent data can contain additional useful information, but extracting it may require explicitly accounting for dependence, microstructure effects and the construction of the sampled variable.

Where financial observations are serially dependent, the nominal number of observations should not be equated with an equal number of independent pieces of information. Andersen et al. (2001) provide a concrete example: the realised equity volatilities and correlations in their sample exhibit strong temporal dependence and long-memory characteristics.

Sampling decisions also define the economic object being investigated. Calendar-time, business-time and transaction-time sampling need not provide statistically equivalent representations of market activity. Oomen (2006), studying realised variance under alternative sampling schemes, shows that the statistical properties of the estimator can depend on how high-frequency observations are sampled.

The appropriate sampling frequency is consequently not a universal constant. It depends on:

- the market and instrument;
- the variable being measured;
- the forecast horizon;
- the underlying data-generating process;
- the characteristics of market microstructure;
- and the estimation method used to separate information from measurement effects.

The relevant question is therefore not simply "How much data are available?" but "At what resolution and under what measurement model do those observations represent the phenomenon being studied?"

2.4 Preprocessing and Data Leakage

Financial data rarely enter a predictive model in their original database form. Researchers may align timestamps, handle missing observations, construct returns, standardise variables, winsorise outliers, estimate rolling quantities, combine databases or transform raw measurements into candidate features.

These operations are often described as preprocessing, a term that can make them appear methodologically neutral. They are not.

Every transformation establishes a rule about which observations are retained, how information is combined and what historical information contributes to each model input. Preprocessing is therefore part of the statistical specification of the research.

A particularly important failure mode is data leakage. Kaufman, Rosset, Perlich and Stitelman (2012) define leakage in data mining as the introduction of information about the prediction target that should not legitimately be available for model construction. The concept is broader than direct look-ahead bias: information can enter indirectly through feature construction, preprocessing, dataset assembly or other data-dependent transformations.

In a financial time-series context, leakage can occur when a transformation applied to an earlier observation is estimated using information from later periods. Consider, for example, standardising a candidate feature using the mean and standard deviation calculated over the entire historical dataset before separating the sample into research and evaluation periods. The transformation applied to the early observations then depends partly on future data. Similarly, an imputation rule, dimensionality-reduction procedure or feature-construction step estimated using the full sample can allow information from the evaluation period to influence the representation seen by the model.

The issue is not that standardisation, imputation or dimension reduction are inherently problematic. The issue is when and on which information set their parameters are estimated.

A robust research pipeline therefore treats data transformations that learn from the sample as part of the model. Their parameters should be estimated using only information legitimately available within the relevant research window and then applied prospectively to later observations.

The broader design of training, validation and out-of-sample evaluation is examined separately in Section 6. At the data stage, the relevant principle is narrower:

No operation used to construct a historical predictor should incorporate information that would have been unavailable when that predictor was formed.

This applies not only to the original variables but also to the transformations used to create them.

Table 1. Research Failure Modes at the Data Stage

Failure Mode	How It Enters the Research Process	Potential Consequence	Principal Research Control
Look-ahead bias	Future or subsequently released information is assigned to an earlier decision point	Artificial predictive relationship	Point-in-time availability timestamps
Vintage substitution / revision leakage	Later or final data vintages replace the observations actually available at the historical decision date	Unrealistic historical information set	Vintage-preserving data
Survivorship bias	Failed or disappeared entities are excluded retrospectively	Distorted distributions or apparent predictability	Point-in-time universe reconstruction
Delisting omission	Final losses associated with unsuccessful securities are absent	Upwardly distorted historical returns	Complete delisting treatment
Sample-selection bias	The mechanism determining inclusion is systematically related to the outcome or relationship under study	The analysed sample may not represent the intended population or opportunity set	Predefined eligibility rules and explicit assessment of the selection mechanism
Sampling distortion	Observation frequency or scheme is inconsistent with the measurement process	Biased or inefficient estimation	Sampling method matched to market structure and estimator
Microstructure contamination	Very high-frequency observations are treated as frictionless efficient prices	Distorted estimates of latent market quantities	Explicit microstructure modelling or appropriate estimator
Preprocessing leakage	Transformation parameters use information from future	Contaminated evaluation and potentially optimistic	Fit transformations only on permissible historical

Failure Mode	How It Enters the Research Process	Potential Consequence	Principal Research Control
	or evaluation periods	estimates of predictive performance	information

The table does not imply that each failure mode can be eliminated through a single mechanical procedure. The appropriate control depends on the market, dataset and research objective. Its purpose is to make explicit that data integrity involves several distinct dimensions that can otherwise be hidden within the general label of "data cleaning."

Data Integrity as an Evidentiary Requirement

The central implication of this section is that data cannot be separated cleanly from inference.

A statistical relationship is conditional on the dataset from which it was estimated, and the dataset is itself the product of choices about timing, inclusion, sampling and transformation. If those choices introduce information that was unavailable ex ante, remove observations through an outcome-related selection mechanism or alter the measurement process in a way that is inconsistent with the research question, later statistical sophistication cannot by itself restore the validity of evidence constructed from a contaminated information set.

This does not imply that a perfect reconstruction of historical markets is always possible. Databases can be incomplete, release timestamps can be uncertain, and market conventions change. The appropriate standard is therefore not the elimination of every possible measurement imperfection. It is the transparent identification of data limitations, the use of point-in-time information where it is material, and research controls proportionate to the predictive claim being made.

For quantitative research, the relevant sequence is consequently:

DATA AVAILABILITY

→ POINT-IN-TIME INTEGRITY

→ REPRESENTATIVE UNIVERSE

→ APPROPRIATE SAMPLING

→ LEAKAGE-FREE TRANSFORMATION

→ MODEL INPUT

Only after these conditions have been considered does the analysis move to the next question: whether relationships among those model inputs represent merely statistical association or provide incremental predictive information.

3. From Association to Predictive Information

Once the integrity of the underlying data has been established, the next question is whether an observed relationship contributes information about a future outcome.

This distinction is more demanding than determining whether two variables are statistically associated. Financial variables can move together contemporaneously, share exposure to a common underlying driver or display a stable historical relationship without one providing useful information about the future value of the other. Even when a relationship is predictive in some sense, it may add little once a benchmark forecast or existing information set is taken into account.

For signal research, the relevant progression is therefore not:

CORRELATION → SIGNAL

but:

ASSOCIATION

→ TEMPORALLY VALID PREDICTIVE RELATIONSHIP

→ INCREMENTAL INFORMATION

→ ROBUST CANDIDATE SIGNAL

The distinction becomes particularly important as the number of available features increases. A large information set may contain meaningful structure while also containing variables that are redundant, highly correlated or useful only through combinations with other information. Feature selection is therefore not simply a search for variables that are individually significant. It is an attempt to determine which information contributes something that is not already represented elsewhere.

3.1 Correlation versus Prediction

Correlation describes statistical association. Prediction imposes an additional temporal and informational requirement.

A contemporaneous relationship between two quantities can be useful for description without establishing that one helps forecast the other. If two financial variables respond to the same market development at approximately the same time, they may be strongly correlated even though neither contains information that would have allowed the other to be forecast in advance.

Prediction instead asks whether information available at one point in time improves the forecast of a subsequently realised target.

This distinction is closely related to the predictive framework introduced by Granger (1969), which evaluates whether information from one process improves forecasts of another conditional on the specified information set. In this paper, the framework is used only to

illustrate conditional predictive content; it should not be interpreted as establishing structural economic causation.

This leads to three distinctions.

First, contemporaneous association is not forward-looking information.

Second, temporal precedence alone is not sufficient. A variable may occur earlier without improving prediction.

Third, predictive value is conditional on the existing information set. A relationship that appears informative in isolation may become redundant once other variables are incorporated.

Forecasting research consequently evaluates predictions comparatively rather than assessing each forecast in isolation. Diebold and Mariano (1995), for example, develop tests for whether competing forecasts differ in predictive accuracy under a specified loss function. Their framework makes explicit that forecasting quality depends on the criterion used to evaluate forecast errors and on the performance of the relevant alternative.

A purely illustrative example makes the distinction clear. Suppose that variable X has a positive historical association with the next-period value of target Y. This establishes a candidate predictive relationship. But if a simple benchmark forecast of Y performs equally well without X, the association has not demonstrated incremental forecasting value. Alternatively, if X improves forecasts only because it proxies for information already contained in variable Z, then its apparent predictive content may not represent an independent source of information.

Signal identification therefore requires more than asking whether a relationship exists.

The relevant question is:

What can be forecast using this information that could not be forecast as well without it?

3.2 Incremental Information versus a Benchmark

The concept of incremental information provides the bridge from statistical association to a candidate predictive signal.

A predictor is incremental when adding it to the relevant information set improves the forecast of the defined target according to the evaluation criterion established by the research design.

This definition has three implications.

First, predictive information is relative rather than absolute. The appropriate benchmark may be a historical average, an existing forecasting model, a market-implied quantity or another forecast appropriate to the research question.

Second, the benchmark should be specified before the candidate relationship is evaluated. Changing the benchmark after observing results can change the apparent contribution of a predictor.

Third, improvement must be judged according to a defined criterion. A model can outperform another under one loss function or investment objective while offering little improvement under another.

The equity-premium forecasting debate discussed in Section 1 provides a useful example. Campbell and Thompson (2008) compare predictive regressions with the historical-average excess-return forecast and show that economically motivated restrictions can allow several predictive models to improve upon that benchmark. Importantly, the out-of-sample explanatory improvements they report are small, while the authors nevertheless find them economically meaningful for a mean-variance investor. This illustrates two distinct questions that should not be conflated: whether a forecast improves statistically upon a benchmark and whether the magnitude of that improvement matters economically.

Forecast evaluation becomes more subtle when the candidate model contains the benchmark model as a special case. Clark and West (2007) show that, when forecasting models are nested, the raw difference in mean-squared prediction error can be biased against the larger model under the null of equal population predictive ability because estimation of additional parameters introduces forecast noise. They propose an adjustment designed for this setting.

Forecast encompassing provides a related way to assess incremental information. If an optimally formed combination of two forecasts assigns no useful incremental role to one of them, the other forecast may be said to encompass its predictive information. Harvey, Leybourne and Newbold (1998) develop tests for this setting.

Incremental information should therefore not be reduced to the question:

"Is the coefficient statistically significant?"

A more demanding sequence is:

Does the candidate contain predictive information?

Does it improve upon the benchmark?

Does it add information beyond what existing predictors already provide?

Is the improvement material under the relevant evaluation criterion?

The last question eventually leads to economic significance and implementation, which are examined in Section 7. At the current stage, the purpose is narrower: to determine whether the candidate contributes information that is genuinely new to the forecasting problem.

3.3 Feature Selection and Redundancy

As the number of candidate predictors grows, evaluating information one variable at a time becomes increasingly inadequate.

A variable may appear strongly related to returns when examined individually but contribute little when other correlated variables are considered. Several features may capture different

manifestations of the same underlying phenomenon. Conversely, information may be distributed across combinations of variables such that no single feature provides an adequate representation on its own.

Related evidence from cross-sectional asset-pricing research illustrates the problem of conditional redundancy, although these studies should not be interpreted as direct out-of-sample forecasting tests.

Green, Hand and Zhang (2017) simultaneously examine 94 firm characteristics in U.S. equity data to determine which provide independent information about average monthly returns. For non-microcap stocks over 1980–2014, they identify 12 characteristics as reliable independent determinants in their specification; after 2003, only two remain independent determinants. Their results illustrate how a large universe of individually proposed characteristics can contract substantially when the question changes from whether each characteristic is associated with returns to whether it contributes independent information in a joint setting.

In a related asset-pricing rather than forecasting setting, Feng, Giglio and Xiu (2020) examine whether newly proposed factors contribute explanatory power beyond a high-dimensional set of existing factors while accounting for model-selection errors. In their empirical application, most of the newly considered factors are redundant relative to the existing factor set, while a smaller number retain statistically significant incremental explanatory power.

These results should not be interpreted as evidence that financial information is necessarily sparse or that a very small number of original variables must always be optimal.

Kozak, Nagel and Santosh (2020) provide an important counterpoint. In a high-dimensional cross-sectional asset-pricing setting, they find that models based on only a few characteristics-based factors do not adequately summarise the cross-section of expected stock returns, whereas a representation based on a small number of principal components extracted from a much broader factor universe performs well out of sample.

Their result illustrates that dimensionality and informational redundancy depend on the representation of the predictor space. Information can remain distributed across many primitive characteristics while becoming more compact after those characteristics are transformed into combinations such as principal components.

This leads to an important distinction:

Feature selection is not equivalent to selecting a small number of original variables.

More generally:

Low dimensionality of an effective representation does not imply that the original information set is itself sparse.

Feature selection should therefore be understood as an attempt to control how information enters the model rather than simply to minimise the number of variables.

Depending on the research problem, this may involve:

- retaining variables with demonstrably distinct information;

- shrinking weak or unstable contributions;
- combining correlated features;
- representing a high-dimensional information set through lower-dimensional components;
- or excluding variables whose apparent contribution disappears conditional on existing information.

These approaches answer related but different questions. Variable selection asks which individual features should remain. Shrinkage limits the influence of weakly supported estimates. Dimension reduction asks whether a smaller number of latent combinations can represent a larger information set.

No single method is universally preferable. The appropriate approach depends on the structure of the data, the forecasting objective and the research design.

The relevant principle is instead:

The number of available features is not the same as the number of independent sources of predictive information.

Nor is the number of original features necessarily the same as the effective dimensionality of the information they collectively contain.

3.4 Economic Interpretation as Supporting Evidence

Statistical evidence alone does not explain why a relationship might exist.

Economic reasoning can therefore play an important role in signal research. A plausible mechanism can help define the expected direction of a relationship, identify relevant horizons, distinguish competing interpretations and design robustness tests. It can also provide disciplined restrictions that reduce reliance on unconstrained empirical fitting.

Campbell and Thompson (2008) provide one example. Their forecasting improvements arise partly from restrictions motivated by economic reasoning, including restrictions on coefficient signs and return forecasts, rather than from unconstrained optimisation of historical fit.

Kozak, Nagel and Santosh (2020) provide another. Their high-dimensional stochastic-discount-factor approach imposes an economically motivated prior that shrinks the contribution of low-variance principal components, improving robustness in their empirical setting.

Economic structure can therefore act as a form of research discipline.

It cannot, however, substitute for predictive evidence.

A compelling narrative does not demonstrate that a variable improves forecasts. Nor does predictive success by itself establish the mechanism that generates the relationship. Cochrane (2011) highlights the rapid proliferation of proposed expected-return factors, illustrating the broader challenge of reconciling an expanding empirical literature with coherent asset-pricing explanations.

The distinction can be stated directly:

Economic reasoning can motivate a signal hypothesis.

Statistical evidence can support a predictive relationship.

Neither alone establishes a robust signal.

This separation is particularly important when an economic explanation is developed after the empirical relationship has already been observed. An ex post narrative may make a historical pattern appear intuitive without providing independent evidence that the relationship will persist. Economic interpretation is most informative when it generates restrictions, predictions or robustness implications that can themselves be evaluated rather than merely providing a retrospective story.

A strong signal assessment should therefore consider two separate questions:

Does the relationship add predictive information?

and

Is there a coherent economic basis that can help explain when and why that information might persist?

The first question is indispensable to a predictive claim. A satisfactory answer to the second can strengthen interpretation, generate additional testable implications and increase confidence in persistence, but the absence of a complete economic explanation does not by itself invalidate predictive evidence.

Even after predictive information has been established, however, an additional problem remains. A researcher who examines many predictors, specifications, transformations or models can discover apparently compelling relationships simply because the search process itself creates opportunities for false positives.

The next section therefore moves from the information contained in an individual candidate relationship to the evidentiary consequences of searching across many such relationships.

4. Discovery Under Model Search

The evidentiary meaning of a statistical result depends not only on the result itself, but also on the process through which it was discovered.

A relationship identified after examining a single prespecified hypothesis is not statistically equivalent to the best-performing relationship selected from hundreds or thousands of alternatives. In the latter setting, the observed result reflects both the underlying data-generating process and the search procedure that selected it.

This creates a central problem for quantitative research. Modern financial datasets permit researchers to evaluate large numbers of variables, transformations, horizons, portfolio definitions, parameter values and model specifications. Such flexibility can help identify genuine structure that would otherwise remain hidden. But it also increases the probability that at least some candidate relationships will appear unusually successful because of sampling variation alone.

The relevant question therefore changes from:

Does this individual result appear significant?

to:

How strong is this result relative to the number, dependence and structure of the alternatives that could have produced it?

This distinction underlies multiple testing, data snooping, model-selection bias and backtest overfitting. These concepts are related, but they are not interchangeable.

4.1 Multiple Testing and False Discovery

Classical hypothesis testing is easiest to interpret when a single prespecified hypothesis is evaluated.

If many hypotheses are examined, the probability of obtaining at least one apparently significant result can increase even when some or all of the underlying null hypotheses are true. The statistical problem is therefore no longer defined solely by the significance level of each individual test. It also depends on the collection of tests considered jointly.

Different multiple-testing procedures address different error concepts.

The familywise error rate (FWER) concerns the probability of making at least one false rejection within a family of hypotheses.

The false discovery rate (FDR) addresses a different quantity: the expected proportion of false rejections among the hypotheses that are rejected.

Procedures that control the familywise error rate can be more conservative than procedures targeting the false discovery rate, particularly in large testing problems, although their relative power depends on the testing procedure and the dependence structure among hypotheses.

Benjamini and Hochberg (1995) introduced a procedure for controlling the FDR under the assumptions of their original framework, providing an influential alternative to familywise-error control in large multiple-testing problems.

Neither framework implies that one universal significance threshold should apply to every research setting. The appropriate error criterion depends on the objective of the research, the number of hypotheses, their dependence structure and the relative consequences of false discoveries and missed discoveries.

Romano and Wolf (2005), for example, develop a stepwise multiple-testing procedure that controls the familywise error rate while incorporating the joint dependence structure of the test statistics. Their framework, presented in the context of comparing multiple strategies with a common benchmark, illustrates an important point for financial research: tests of related strategies are often statistically dependent, so treating every candidate as an unrelated experiment can misrepresent the structure of the testing problem.

The distinction between an individual p-value and the evidence produced by a larger search process is fundamental.

Suppose, purely illustratively, that a researcher evaluates many candidate predictors and retains the one with the strongest historical result. The reported significance of the selected predictor cannot be interpreted as though it had been nominated independently before the search began. The selection process itself has influenced which statistic is being observed.

Multiple-testing adjustment attempts to account for this broader evidentiary context.

It does not imply that a significant result discovered within a large research programme must be false. Nor does it imply that every candidate should be subjected to the same mechanical correction. It implies that the amount and structure of searching undertaken are relevant information when assessing the credibility of the selected result.

This principle becomes especially important when the true size of the search space is difficult to observe. A published study may report a small number of final specifications while the underlying research process involved alternative variables, transformations, horizons or portfolio definitions that never appear in the final document.

The effective research search space can therefore be larger than the number of formal tests eventually reported.

4.2 Data Snooping and Research Selection

The repeated use of the same historical data creates a closely related problem.

White (2000) defines data snooping as the reuse of a given dataset for purposes of inference or model selection. When the same history is repeatedly searched, the best-performing model encountered may owe part of its apparent success to the search itself rather than to persistent predictive superiority. White develops a Reality Check for testing the null hypothesis that the

best model encountered in a specification search has no predictive superiority over a specified benchmark, while accounting for the model search.

The problem is especially relevant for financial time series because researchers often possess only one realised historical path of the market being studied. Once that history has been used to generate hypotheses, select variables, reject unsuccessful specifications and refine successful ones, it becomes increasingly difficult to regard the same observations as independent evidence for the resulting model.

Sullivan, Timmermann and White (1999) demonstrate the practical importance of this issue in their study of technical trading rules. They expand the universe of rules beyond those examined in earlier work and apply a bootstrap procedure designed to account for data snooping across the broader strategy universe. Their analysis illustrates why the performance of a selected trading rule should be evaluated relative to the larger set of rules from which it could have been selected.

For clarity, several related forms of research selection should remain distinct.

Multiple testing describes a statistical environment in which many hypotheses are evaluated. It may be entirely transparent and prespecified; the methodological problem arises when the multiplicity of the testing environment is ignored in the interpretation of the results.

Data snooping refers to repeated use of the same data for inference or model selection.

Model-selection bias arises when a preferred model is selected from multiple alternatives and its observed performance is interpreted without adequately accounting for the selection process.

Selective reporting occurs when unsuccessful or less favourable analyses are omitted while successful results are emphasised.

p-hacking refers to analytical or reporting choices that are responsive to the statistical results obtained and that increase the probability of producing or highlighting conventionally significant findings, particularly when such flexibility is not adequately disclosed or incorporated into inference.

Publication bias concerns the possibility that the research ultimately visible in the published literature differs systematically from the broader set of research that was undertaken or submitted.

Multiple testing and p-hacking should therefore not be treated as synonyms. Multiple testing can be legitimate, transparent and prespecified. P-hacking concerns result-responsive analytical or reporting flexibility.

Similarly, publication bias should not be treated solely as a journal-selection problem. Selection can occur when researchers decide which analyses to pursue, which findings to report, whether to submit a study and when journals or reviewers decide which work to publish.

These mechanisms can reinforce one another, but they operate at different stages of the research process.

The appropriate response therefore also differs. Multiple-testing procedures can address some consequences of simultaneous hypothesis search. Transparent reporting can reveal the number and nature of specifications examined. Independent validation can reduce reliance on the discovery sample. None of these controls alone eliminates every form of research selection.

A key methodological principle is therefore:

The evidentiary history of a result includes the unsuccessful alternatives that helped produce it, even when those alternatives are not ultimately reported.

4.3 Model Selection and Backtest Overfitting

Model search does not require explicit hypothesis testing to create optimistic results.

A researcher may instead evaluate many model architectures, parameter combinations, signal thresholds, portfolio rules or weighting schemes and select the specification with the best simulated historical performance. The final backtest is then the maximum of a larger set of noisy estimates.

This creates model-selection bias.

The selected model may genuinely be superior. But even when candidate models have similar underlying predictive ability, the model with the highest historical performance will tend to have benefited from unusually favourable sample-specific variation.

When the research process becomes sufficiently flexible, this can lead to backtest overfitting: the selected specification captures features of the historical sample that do not generalise to unseen data.

Bailey and López de Prado (2014) examine this problem in the context of investment-strategy evaluation. They argue that searching across large numbers of strategy variants can inflate observed Sharpe ratios and propose the Deflated Sharpe Ratio to adjust performance assessment for selection under multiple testing and for non-normal returns. Their contribution illustrates the broader principle that observed backtest performance should be interpreted in light of the number and characteristics of the trials from which the selected strategy emerged.

This issue is broader than mechanical parameter optimisation.

Consider a research process in which the researcher repeatedly changes:

- the predictor definition;
- the estimation window;
- the forecast horizon;
- the asset universe;
- the portfolio-construction rule;
- the transaction filter;
- or the model architecture.

Even if each change is individually defensible, the accumulated flexibility creates a larger effective model search. If only the final specification is evaluated as though it had been defined independently of the discovery process, the resulting evidence can be overstated.

This does not imply that iterative research is invalid. Scientific research necessarily involves learning from data, modifying hypotheses and improving models.

The relevant distinction is between discovery and confirmation.

A historical dataset can legitimately be used to explore ideas and identify candidate models. The resulting model becomes evidentially stronger only when its performance is assessed in a manner that acknowledges how much of the available information has already influenced its design.

This is why a highly optimised backtest and a test of the same model specified independently of the evaluation evidence are not equivalent forms of evidence, even if the final performance statistics are identical.

The problem is therefore not model complexity by itself.

A complex model can generalise well.

A simple model can also be overfit if it was selected after enough alternative simple models were examined.

What matters is the combination of model flexibility, search intensity and independent evidence.

The independent-evidence problem is addressed in detail in Section 6.

4.4 Publication Bias and the Visibility of Research

The research process creates another selection problem: the evidence available to later researchers and investment professionals may itself be a selected subset of the evidence that was originally generated.

Positive, novel or statistically significant results can influence author, submission, reviewer and journal decisions. The published literature may therefore reflect several stages of selection rather than a neutral record of all hypotheses that were investigated.

Harvey (2017) argues that competition for publication and preferences for statistically significant findings can encourage unreported testing and both direct and indirect forms of p-hacking in financial economics. He cautions that increasing conventional statistical thresholds alone may not fully resolve the problem because the reliability of a finding depends on broader features of the research process.

The magnitude and source of publication-related distortion are nevertheless contested.

Brodeur et al. (2023) provide useful evidence from economics more broadly by examining journal submissions rather than only published papers. They find substantial bunching of test statistics already in initial submissions. This implies that the distribution observed in published research

cannot be attributed solely to editorial or peer-review selection and is consistent with researcher-side analytical or reporting selection occurring before submission.

In their sample, peer review has relatively little effect on the overall distribution of test statistics, leading the authors to conclude that publication bias may be less prominent than sometimes feared. Their study concerns economics broadly rather than financial asset pricing specifically, but it illustrates why researcher selection, submission selection and journal selection should not be treated as the same mechanism.

The asset-pricing literature contains a related disagreement over the extent to which published factor evidence could plausibly be generated by p-hacking.

Chen (2021) examines deliberately extreme thought experiments in which all published asset-pricing factors are assumed to be spurious and asks how much p-hacking would be required to reproduce the distribution of observed published t-statistics. Within these stylised experiments, Chen concludes that p-hacking alone is unlikely to account for the entire published factor universe.

The analysis does not estimate the actual prevalence of p-hacking. Rather, it examines whether particular p-hacking mechanisms could plausibly generate the observed distribution of published t-statistics under deliberately strong assumptions.

This distinction is important.

Evidence of publication or selection bias does not imply that the published literature contains no genuine information.

Conversely, evidence that many findings replicate does not imply that the publication process is unbiased.

The appropriate research response is therefore not to assume either complete reliability or complete contamination of the visible literature. It is to treat the research-publication pipeline as a collection of potential selection mechanisms and to place greater weight on transparent designs, complete specification reporting, independent replication and evidence obtained after the original discovery.

4.5 The Factor-Zoo and Replication Debate

The debate over equity-return factors provides one of the clearest empirical illustrations of the issues discussed in this section.

Harvey, Liu and Zhu (2016) examine the rapid expansion of variables proposed to explain the cross-section of expected returns. They argue that conventional significance standards become too permissive when hundreds of potential factors have been investigated and develop multiple-testing approaches that imply substantially higher evidentiary hurdles than the traditional single-test standard.

Chordia, Goyal and Saretto (2020) reach a similarly cautious conclusion using a different methodology. They generate information from more than two million trading strategies using

real data and use the resulting distribution, together with strategies surviving the publication process, to infer properties of the broader strategy universe. In their setting, multiple-testing-adjusted t-statistic thresholds are approximately 3.8 for time-series regressions and 3.4 for cross-sectional regressions, and they estimate that failure to account for multiple testing would generate an expected false-rejection proportion of approximately 45%.

These figures are results of their particular research design and should not be interpreted as universal thresholds or universal estimates of false discovery in financial research.

Hou, Xue and Zhang (2020) conduct a large replication exercise covering 452 published anomalies. Using NYSE breakpoints and value-weighted returns to reduce the influence of microcaps, they report that 65% of the anomalies fail to reach an absolute t-statistic of 1.96. Applying their multiple-testing hurdle of 2.78 increases the failure rate to 82%. They also find that the economic magnitudes of many replicated anomalies are smaller than originally reported.

Taken together, these studies provide substantial evidence that the apparent strength of a large anomaly literature can decline when construction choices, multiple testing and replication standards are tightened.

They do not, however, settle the question.

Jensen, Kelly and Pedersen (2023) develop a Bayesian framework for factor replication and construct a global dataset containing 153 factors across 93 countries. They reach materially more favourable conclusions. In their framework, the majority of factors replicate, the factors can be organised into 13 correlated economic themes, and much of the evidence extends internationally and into observations outside the samples used by the original studies.

The headline replication rates reported by Hou, Xue and Zhang (2020) and Jensen, Kelly and Pedersen (2023) are not directly comparable.

Jensen, Kelly and Pedersen explicitly decompose part of the difference. Their baseline raw-return replication rate is 55.6%, compared with the 35% rate they report from Hou, Xue and Zhang. Excluding factors for which the original studies themselves did not establish a significant factor-return or alpha result raises the comparable rate to 61.3%.

Changing the replication object from raw factor returns to CAPM alpha increases their estimate further to 82.4%. Applying a conventional frequentist multiple-testing correction produces a replication rate of 75.6%, whereas their Bayesian joint-factor framework produces a rate of 82.4%.

These differences demonstrate that reported replication rates can change materially with sample construction, factor construction, the statistical object being tested, risk adjustment and the inferential framework.

The studies also differ more broadly in:

- the factor universes examined;
- portfolio-construction choices;
- the treatment of small stocks;

- the statistical object being tested;
- frequentist versus Bayesian inference;
- the treatment of dependence among factors;
- the use of raw returns versus risk-adjusted factor alphas;
- and the definition of successful replication.

Jensen, Kelly and Pedersen argue in particular that factors should not necessarily be treated as unrelated hypotheses because many belong to highly correlated themes. Their Bayesian framework uses dependence across factors and countries as information rather than ignoring that structure.

The methodological conclusion is therefore more important than choosing a single side of the debate.

Reported replication rates are not determined by the underlying literature alone. They can depend materially on sample and factor construction, the object being tested, risk-adjustment choices, inferential methodology and the definition of successful replication.

For RP-002, the evidence supports three narrower conclusions.

First, conventional single-test significance thresholds can be inadequate when the effective research search space is large.

Second, large-scale replication and multiple-testing adjustments can materially weaken the apparent evidence for some published relationships.

Third, neither observation establishes that the majority of financial predictors are necessarily false. Alternative methodologies and broader datasets can produce materially different assessments of the same general literature.

The appropriate response to the factor-zoo debate is therefore not a universal higher t-statistic or a presumption that every published anomaly is either valid or spurious.

It is a higher standard of research transparency.

Researchers should seek to understand:

How many hypotheses or model variants were effectively considered?

How dependent were those alternatives?

How was the final specification selected?

What evidence exists outside the discovery process?

And how sensitive is the conclusion to reasonable alternative definitions of the test itself?

These questions shift the focus from the apparent success of the selected model to the evidentiary process that produced it.

Search-Aware Evidence

The central implication of this section can be expressed simply:

A result should be evaluated in the context of the search that produced it.

Conventional statistical significance is most straightforward to interpret when the hypothesis and inferential procedure were specified independently of the data used for the test. As the same evidence plays a larger role in hypothesis generation, specification choice and model selection, the resulting test requires increasingly search-aware interpretation or independent confirmation.

This does not require eliminating exploratory research. Exploration is an essential part of quantitative discovery.

It requires distinguishing clearly between:

EXPLORATION

→ CANDIDATE SELECTION

→ SEARCH-AWARE INFERENCE

→ INDEPENDENT VALIDATION

The first stages can identify potentially useful relationships. They cannot alone determine whether those relationships will survive beyond the environment in which they were discovered.

That question becomes increasingly important because financial relationships are not necessarily stable through time. Even a relationship that survives appropriate search controls may weaken as market conditions, participants and economic regimes change.

The next section therefore examines non-stationarity, changing regimes and signal decay.

5. Non-Stationarity, Regimes and Signal Decay

A relationship can survive careful data construction, demonstrate incremental predictive information and withstand controls for model search without remaining equally informative through time.

This possibility is particularly important in financial markets because the environment in which a relationship is estimated can change. Economic policy, market participation, liquidity, institutional structures, technology, risk preferences and the behaviour of other market participants are not fixed. A predictor that captures information under one set of conditions may therefore weaken, strengthen or change character under another.

Instability does not have a single interpretation. A declining relationship may indicate that the original finding was partly sample-specific. It may instead reflect a structural change in the underlying economic process, a transition between market regimes, adaptation by market participants, or a change in the costs and constraints associated with exploiting the information. Several mechanisms may operate simultaneously.

For this reason, the observation that a signal has weakened should be separated from the explanation for why it has weakened.

The relevant progression is:

PREDICTIVE RELATIONSHIP → TEMPORAL STABILITY → REGIME DEPENDENCE → POST-DISCOVERY BEHAVIOUR → DECAY DIAGNOSIS

A robust research process should therefore evaluate not only whether a relationship exists, but also how its strength and interpretation vary across time and market conditions.

5.1 Parameter and Relationship Instability

In time-series analysis, stationarity is a property of the stochastic process being studied, whereas parameter stability concerns the constancy of a specified model relationship through time. The concepts are related in many forecasting applications but are not interchangeable.

Evidence that predictive coefficients change does not by itself establish that every property of the underlying variables is non-stationary. Conversely, non-stationarity in an observed series does not by itself determine whether a particular predictive relationship is unstable.

For signal research, the operational question is therefore narrower:

Does the relationship used for prediction remain sufficiently stable over the period for which the predictive claim is intended?

Evidence from return forecasting demonstrates that the answer can change through time. Pesaran and Timmermann (1995), examining U.S. stock-return predictability, find that the predictive power of several economic variables changes over time and tends to vary with return volatility. Their findings illustrate that a predictor need not have a single, time-invariant relationship with future returns across the entire historical sample.

Forecast instability is not confined to financial returns. Rossi (2021), reviewing a broad forecasting literature, documents the importance of instabilities for forecast evaluation and discusses methods designed to detect and accommodate changes in predictive relationships. The central implication for signal research is that performance measured over a full historical sample can conceal meaningful variation in performance across subperiods.

A full-sample estimate can therefore represent an average of relationships that differed substantially through time.

Consider a purely illustrative predictor whose estimated relationship with returns is strongly positive during one part of a sample and approximately zero during another. The full-period coefficient may remain positive and statistically significant. That estimate describes an average historical relationship, but it does not establish that the same relationship was continuously present.

Longer histories provide more observations and may improve estimation precision, but they may also span economic or market environments in which the relevant relationship differs. More history is therefore not automatically more representative history.

Structural change provides one possible explanation. Lettau and Van Nieuwerburgh (2008), for example, examine return predictability using valuation ratios when the steady-state relationship represented by the dividend-price ratio may shift over time. In their dividend-price-ratio application, allowing for shifts in the estimated steady-state relationship materially changes the return-predictability evidence and helps reconcile otherwise conflicting in-sample, out-of-sample and parameter-stability results.

This example is important because instability does not necessarily imply that the original predictive relationship was spurious.

A predictor can fail under a constant-parameter specification because the relevant economic relationship itself has changed.

Conversely, allowing for structural change can create additional modelling flexibility and therefore additional opportunities to fit historical data. Breaks and time variation should not be introduced merely because they rescue an otherwise unsuccessful predictor. Evidence for instability must itself be subjected to disciplined estimation and subsequent validation.

The appropriate conclusion is therefore not that stable models are inherently naive or that time-varying models are inherently superior.

It is that temporal stability is an empirical assumption that should be examined rather than silently imposed.

5.2 Changing Market Regimes

One way to represent time variation is through the concept of regimes.

A regime describes a state in which the statistical or economic behaviour of relevant market variables differs from that prevailing in another state. Regimes may differ, for example, in

expected returns, volatility, correlations or other parameters relevant to the forecasting problem.

Hamilton (1989) provides a foundational econometric framework for modelling discrete regime changes. In his approach, parameters of a time-series model can depend on an unobserved state governed by a Markov process, with the state inferred probabilistically from observed data.

The importance of the framework for financial research lies not in assuming that markets literally occupy a small number of objectively observable regimes, but in providing a disciplined way to represent relationships whose behaviour may change between persistent states.

Financial evidence provides economically meaningful examples of such state dependence.

Ang and Bekaert (2002), studying international asset allocation, model a time-varying investment opportunity set with regime shifts. Their analysis allows correlations and volatilities across international equity markets to rise in highly volatile bear-market states. The study illustrates how relationships that appear comparatively stable under ordinary conditions can behave differently when the market enters a materially different state.

For signal research, this creates an important distinction between instability and conditional stability.

A predictor may appear unstable when evaluated unconditionally across all observations but be more stable conditional on a particular market environment. Alternatively, a predictor may perform well in one identified regime and fail in another.

Neither observation automatically establishes a usable regime-dependent signal.

A regime classification itself can be uncertain. In Hamilton-type models, the state is latent and must be inferred. A regime identified retrospectively may also be easier to classify than one observed in real time. If a model recognises a crisis regime only after sufficient crisis observations have occurred, a historical regime-conditioned result can overstate the information that would actually have been available at the beginning of that regime.

This reconnects regime research to the point-in-time principles established in Section 2.

A valid regime-dependent model must therefore answer at least three separate questions:

Are the states statistically distinguishable?

Was the information required to identify the state available at the relevant forecast time?

Does conditioning on the state improve genuine predictive performance rather than merely historical fit?

Regime models should consequently be treated as one possible representation of instability rather than as evidence that every market relationship is best understood through discrete states.

The broader principle is:

A relationship that is unstable unconditionally may still contain conditional information, but conditioning creates an additional hypothesis that must itself be validated.

5.3 Signal Decay

Even when a predictive relationship was genuinely present historically, its subsequent strength need not remain constant.

The term signal decay is used here to describe a reduction through time in the measured predictive or economic usefulness of a previously identified relationship.

This definition is deliberately descriptive.

Signal decay identifies an outcome; it does not identify its cause.

Several different processes can generate similar observed deterioration.

A relationship may initially have been overestimated because the selected historical sample was unusually favourable. A structural change may alter the economic relationship between predictor and target. Competing investors may begin to act on the same information. Market structure may change. The implementation costs associated with capturing the information may increase. Or the relationship may remain statistically detectable while becoming too weak to matter economically.

For analytical purposes, this paper distinguishes two dimensions of deterioration.

Statistical decay refers to a decline in the magnitude, reliability or incremental predictive content of a relationship.

Economic decay refers to a decline in the value obtainable from acting on the relationship, even if some statistical predictability remains.

The two can occur together, but they need not.

McLean and Pontiff (2016) provide unusually useful evidence because their research design separates different stages of predictor performance. They study 97 variables previously documented to predict the cross-section of stock returns and compare performance within the original samples, outside those samples but before publication, and after publication.

They report portfolio returns that are 26% lower out of sample than in sample and 58% lower after publication than in sample.

McLean and Pontiff interpret the 26% out-of-sample decline as an upper-bound estimate of data-mining effects and estimate an additional 32% reduction in returns associated with publication-informed trading. Their additional findings—including larger post-publication declines for stronger in-sample predictors and rising correlations among published-predictor portfolios—are consistent with investors learning about documented return patterns.

The decomposition is important because it distinguishes two different questions.

The first asks whether the original relationship survives observations not used in its initial discovery.

The second asks whether the relationship changes further after information about it becomes publicly available.

These findings should nevertheless be interpreted within the scope of the study.

They do not establish a universal rate at which financial signals decay. They concern a particular set of published cross-sectional U.S. equity predictors, constructed and evaluated under the authors' methodology. Nor does the observed post-publication decline establish that every underlying relationship was arbitraged away.

Subsequent international evidence cautions further against treating this post-publication pattern as universal. Jacobs and Müller (2020) examine 241 cross-sectional anomalies across 39 stock markets and find a reliable post-publication decline in long-short returns only in the United States. Their results indicate that the effects associated with publication and subsequent arbitrage can depend materially on market structure and barriers to implementation.

The combined evidence therefore supports a conditional rather than universal interpretation of publication-related signal decay.

For signal research, post-discovery behaviour can consequently provide evidence that is qualitatively different from ordinary historical validation, but its interpretation remains market- and design-specific.

5.4 False Discovery, Structural Change or Market Learning?

When a signal deteriorates, several explanations may fit the same observed pattern.

The first possibility is initial overestimation or false discovery.

If the original relationship was selected from a large search space, part or all of its apparent strength may have reflected sample-specific variation. Subsequent deterioration then does not require the market to have changed; it may simply reveal that the initial estimate was too optimistic. This mechanism connects signal decay directly to the search problems examined in Section 4.

The second possibility is structural change.

The relationship may have been genuine under the conditions prevailing during the discovery sample, but the economic process connecting the predictor and target may subsequently have changed. The evidence in Pesaran and Timmermann (1995) and Lettau and Van Nieuwerburgh (2008) demonstrates why parameter instability and changes in underlying relationships are credible possibilities in return-prediction settings.

The third possibility is regime dependence.

The relationship may remain informative only under particular market states. What appears to be decay across the complete sample may therefore reflect changing exposure to states in which the predictor is more or less informative.

The fourth possibility is market learning or adaptation.

Once information about a return pattern becomes known and sufficiently exploitable, market participants may change their behaviour. Trading on the relationship can alter prices, reduce the

original return opportunity or change the timing through which the relevant information is incorporated into prices.

McLean and Pontiff's post-publication evidence is consistent with such an interpretation for the predictors in their U.S. sample. Jacobs and Müller's broader international evidence, however, demonstrates why the same mechanism should not automatically be assumed across markets.

Publication is also not a controlled intervention that permits every alternative explanation to be eliminated.

Market adaptation should therefore be treated as one possible economic mechanism consistent with some empirical evidence, not as an inference that follows automatically whenever predictability declines.

A fifth possibility is changing economic implementability.

A relationship may retain forecast information while becoming harder or more expensive to exploit. Increased competition, lower available capacity, changes in liquidity or altered execution conditions can reduce net economic value even without eliminating the underlying statistical relationship. These implementation effects are examined in Section 7.

The central diagnostic problem can therefore be represented as:

OBSERVED SIGNAL DETERIORATION

↓

INITIAL OVERESTIMATION / FALSE DISCOVERY?

STRUCTURAL CHANGE?

REGIME DEPENDENCE?

MARKET LEARNING / ADAPTATION?

CHANGING IMPLEMENTATION ECONOMICS?

These explanations are not mutually exclusive.

A predictor can have been initially overestimated, subsequently encounter a structural change and later face greater competition from other investors. Historical data may not permit these components to be separated perfectly.

This uncertainty places an important limit on the interpretation of signal decay:

Observed deterioration is evidence about stability. It is not, by itself, evidence about mechanism.

A rigorous research process should therefore resist retrospective narratives that automatically explain weakening performance as market adaptation or arbitrage.

Stability as an Ongoing Evidentiary Requirement

The broader implication of this section is that robustness cannot be treated as a one-time property established at the end of model development.

Predictive relationships exist within evolving market environments. Evidence supporting a relationship at one point in time does not guarantee identical behaviour indefinitely, while later deterioration does not necessarily demonstrate that the original evidence was invalid.

A useful analytical sequence is:

DISCOVERY → INITIAL QUALIFICATION → TEMPORAL VALIDATION → REGIME SENSITIVITY → POST-DISCOVERY MONITORING → DECAY DIAGNOSIS → RETAIN · ADAPT · REJECT

This sequence is conceptual rather than prescriptive. It does not imply that a model should be modified whenever recent performance weakens.

Repeated adaptation to every new observation can itself create overfitting.

The purpose of monitoring is instead to determine whether incoming evidence remains consistent with the assumptions and predictive claims under which the signal was originally qualified.

This creates a tension that quantitative research cannot eliminate entirely.

A model that never adapts may fail to respond to genuine structural change.

A model that adapts too readily may fit transient noise.

The appropriate balance depends on the research problem, horizon, available data and economic mechanism. What matters is that adaptation rules form part of the research design and are themselves evaluated rather than introduced retrospectively whenever historical performance deteriorates.

The next stage of the paper therefore addresses the primary mechanism for separating discovery from subsequent evidence: out-of-sample validation, independent replication and the hierarchy of evidence available after a candidate signal has been identified.

6. Validation: From Discovery to Evidence

The discovery of a candidate predictive relationship does not complete the research process. It changes the question.

During discovery, the objective is to identify potentially informative structure. During validation, the objective is to determine whether the evidence supporting that structure survives observations and decisions that were not responsible for creating it.

This distinction matters because a model can fit historical data for several reasons. It may capture persistent predictive information. It may capture relationships specific to the estimation period. It may have benefited from model selection, parameter tuning or repeated researcher interaction with the same sample. Historical performance alone cannot determine which explanation dominates.

Validation therefore attempts to introduce evidentiary separation between the process that generated a model and the observations used to assess it.

A useful principle is:

Evidence provides a stronger test of generalisation when it is less involved in the process that generated the model being evaluated.

The implementation is less simple. Financial data are temporally dependent, market relationships can change, observations are limited, and researchers may repeatedly learn from results that were originally intended to remain outside the discovery process. A nominally out-of-sample test can consequently become part of the effective research process even if the underlying observations are never entered directly into the model-estimation routine.

This section distinguishes several forms of validation and evidentiary separation:

IN-SAMPLE EVIDENCE

HISTORICAL VALIDATION

Temporal Holdout · Walk-Forward / Recursive Evaluation

INDEPENDENT EVIDENCE

Independent Reconstruction · Replication · Extension

PROSPECTIVE EVIDENCE

Post-Discovery / Live Observations

Across all of these stages, test-set integrity and control of adaptive reuse remain essential.

The stages represent different forms of evidentiary separation. They should not be interpreted as a strict ranking in which each successive procedure is statistically superior to the preceding one. Temporal holdouts and walk-forward designs primarily separate estimation from subsequent historical evaluation. Independent reconstruction, replication and extension introduce different

forms of separation in implementation, data or context. Prospective evidence introduces temporal separation from the original historical discovery itself.

6.1 In-Sample Evidence

In-sample evidence is generated from observations that contribute directly to model estimation, feature selection, parameter choice or research design.

Such evidence is indispensable.

Researchers require historical data to identify relationships, compare specifications, estimate parameters and understand the behaviour of candidate predictors. In-sample diagnostics can reveal whether a proposed relationship is internally coherent, whether model assumptions are obviously violated and whether the estimated effect is large enough to justify further investigation.

What in-sample evidence cannot establish by itself is generalisation.

The distinction follows directly from the search problem examined in Section 4. Once a dataset has contributed to the identification of a variable, the choice of its transformation, the specification of a model or the rejection of alternatives, measured performance on that same dataset reflects both the underlying relationship and the research decisions made in response to the data.

This is true even when the final model is statistically simple.

A single-factor model selected after examining hundreds of candidate factors has a different evidentiary history from the same model specified before those observations were examined.

In-sample significance and fit should therefore be treated primarily as discovery evidence.

They can support questions such as:

Is there a candidate relationship worth investigating?

Is the direction and magnitude consistent with the proposed hypothesis?

Does the model describe relevant aspects of the historical sample?

They cannot alone answer:

Will the relationship retain predictive information in observations that did not shape its discovery?

That question requires separation between estimation and evaluation.

6.2 Temporal Holdout and Out-of-Sample Testing

A basic form of separation is the temporal holdout.

Historical observations are divided chronologically so that an earlier period is used for model development and a later period is reserved for evaluation. The later observations therefore

simulate, subject to the limitations of historical data, information arriving after the model has been specified.

When the evaluation objective is to reconstruct a genuinely forward-looking historical decision process, a chronological holdout answers a different question from a random partition because it preserves the information ordering faced by the hypothetical forecaster.

This does not imply that random cross-validation is statistically invalid in every time-series setting.

Tashman (2000), in a broad review of out-of-sample forecast evaluation, emphasises that OOS testing involves several design choices rather than a single standardised procedure. These include how the sample is split, whether forecast origins are fixed or rolling, whether model coefficients are updated, whether estimation windows expand or roll, and whether evaluation uses one or multiple forecast periods.

The phrase out-of-sample therefore describes a family of evaluation designs rather than one mechanical test.

A simple fixed holdout can be useful when the model is developed entirely on an earlier sample and then evaluated once on a later period. If the later period genuinely played no role in model development, successful performance provides evidence that the relationship extends beyond the observations that generated it.

But temporal holdouts introduce trade-offs.

Allocating data to a holdout reduces the amount of information available for estimation. A short test period may contain too few relevant market environments to provide a meaningful evaluation. A very long test period may itself span structural changes. The result can also depend materially on where the sample split is placed.

Out-of-sample performance should therefore not be interpreted as a binary property:

passed OOS / failed OOS.

The more informative questions are:

How was the boundary chosen?

How many genuinely new forecasts were evaluated?

Were model parameters re-estimated as new information arrived?

Did the evaluation period contain market conditions relevant to the intended use of the model?

How did information from the holdout enter the research process?

Out-of-sample evidence also remains conditional on the evaluation criterion. As discussed in Section 3, different models can be compared using different forecast-loss functions, and nested-model comparisons can require specialised statistical treatment.

Giacomini and White (2006) extend forecast evaluation further by developing tests of conditional predictive ability in settings where forecasting models may be misspecified. Their framework emphasises that forecasting performance can depend on the information available at the forecast date rather than being adequately summarised by a single unconditional average.

The purpose of temporal holdout evaluation is therefore not to prove that a model is universally correct.

It is to ask whether its predictive claim survives a controlled transition from historical estimation to subsequently observed data.

6.3 Walk-Forward and Recursive Validation

A single train-test split captures only one historical transition.

For many financial applications, a more informative design repeatedly recreates the forecasting process through time.

In recursive or expanding-window evaluation, a model is initially estimated using an early historical sample. A forecast is generated for a subsequent observation or period. Once that outcome has occurred, the estimation sample is expanded, the model may be re-estimated according to the prespecified procedure, and another forecast is generated.

The process continues sequentially:

ESTIMATE → FORECAST → OBSERVE → UPDATE → FORECAST → OBSERVE → UPDATE

A rolling-window design follows a similar sequence but maintains a fixed or otherwise limited estimation window. As new observations enter, older observations leave.

These procedures answer different questions.

An expanding window assumes that earlier history remains sufficiently relevant to retain in the estimation sample.

A rolling window allows older observations to lose influence, which may be appropriate when structural change is a concern but introduces an additional choice: the window length itself.

Neither design is universally superior.

The appropriate procedure should reflect how the model would actually be updated and deployed, as well as the researcher's assumptions about the relevance of older information.

Walk-forward validation is therefore useful because it evaluates not merely a static equation but a forecasting procedure:

- what data are available at each point;
- how parameters are estimated;
- whether features are recalculated;
- when the model is updated;
- and how forecasts are generated before the relevant outcome becomes known.

The procedure should be specified before the evaluation is interpreted.

Otherwise the researcher can retrospectively alter window lengths, update frequencies or recalibration rules in response to historical results, recreating the model-selection problem within the validation process itself.

Time-series dependence also requires nuance in the use of conventional cross-validation.

It is sometimes asserted that ordinary K-fold cross-validation is inherently invalid for time-series data because observations are temporally dependent. That statement is too broad.

Bergmeir, Hyndman and Koo (2018) show that, for purely autoregressive models with uncorrelated errors, standard K-fold cross-validation can be valid and may perform favourably relative to some alternative procedures.

Their result does not imply that random K-fold splitting is appropriate for every financial forecasting problem.

Rather, it illustrates a broader principle:

Validation design should follow from the dependence structure, model class and forecasting objective, not from a universal rule that one splitting procedure is always valid or always invalid.

Where the purpose is to reconstruct a genuinely forward-looking trading or investment decision process, chronological walk-forward evaluation remains particularly transparent because the information ordering corresponds directly to the decision ordering being simulated.

6.4 Test-Set Integrity and Adaptive Reuse

A holdout sample provides independent evaluation only to the extent that its use remains consistent with the inferential procedure being claimed.

Suppose a researcher creates a training sample and a test sample.

The model performs poorly on the test sample.

The researcher then changes the features, modifies parameters or selects another model after observing the result and evaluates the new version on the same test sample.

If this process is repeated without accounting for the adaptivity, information from the test sample is progressively incorporated into model development.

The test data may never have entered the estimation formula directly.

They have nevertheless influenced the model that is ultimately selected.

The distinction is fundamental:

A test set can be technically excluded from estimation while becoming informationally involved in model development.

Dwork et al. (2015), studying adaptive data analysis, emphasise that conventional statistical inference often assumes an analysis procedure fixed independently of the data on which it is

evaluated, while real research frequently adapts subsequent analyses to results obtained from earlier evaluations.

Their work also demonstrates that adaptive reuse does not imply that a holdout may literally be inspected only once. They develop a specialised reusable-holdout framework that permits repeated interaction under additional statistical controls.

For the purposes of this paper, a genuine holdout therefore refers to evaluation evidence whose use remains consistent with the inferential procedure being claimed.

Under ordinary one-shot holdout logic, this means that evaluation results did not inform the specification being tested.

Specialised procedures can permit adaptive reuse under additional statistical controls.

The relevant distinction is therefore not simply whether the data were ever inspected, but whether information obtained from them was incorporated into model development without the inferential procedure accounting for that adaptivity.

Investment backtesting provides an important application of this problem because a finite historical record may be used repeatedly to compare and select strategies. Bailey et al. (2017) develop a framework for estimating the probability of backtest overfitting in precisely this setting.

Test-set integrity should therefore be evaluated by asking:

How many times was the evaluation result observed?

What decisions changed after it was observed?

Were new features, models or parameters introduced in response?

Did the inferential procedure account for that adaptive use?

This leads to a more precise interpretation of out-of-sample evidence:

> Out-of-sample is not merely a label attached to a data segment; its evidentiary meaning depends on how that segment entered the research process.

Creating another test set after the first has been repeatedly inspected also does not automatically restore independence. If researchers continue to iterate through successive historical holdouts, the broader historical record can eventually become part of the effective discovery environment.

There is no single mechanical boundary at which a dataset becomes permanently invalid for evaluation.

The relevant issue is whether its role in model development is compatible with the statistical interpretation attached to the reported result.

6.5 Reproducibility, Replication and Extension

Terminology in this area varies across disciplines.

To avoid ambiguity, this paper follows the distinction adopted by the U.S. National Academies of Sciences, Engineering, and Medicine (2019), while adding two categories useful for quantitative-finance research.

Computational reproducibility asks whether a reported computational result can be recovered using the same input data and analytical procedure.

Independent reconstruction asks whether an independent implementation can recover the central empirical result when the original computational environment is not simply rerun.

Replication asks whether a new study using new data and aimed at the same scientific question obtains a result consistent with the original claim.

Extension or generalisability evidence asks whether the relationship persists when evaluated in different markets, populations, periods or materially different contexts.

These forms of evidence answer different questions.

Computational reproducibility can expose coding errors, undocumented transformations or inconsistencies between reported methods and implemented procedures. It does not itself demonstrate that the relationship generalises beyond the original evidence.

Independent reconstruction adds separation in implementation.

Chen and Zimmermann (2022) provide an important example of large-scale independent reconstruction. They assemble data and code covering 319 characteristics and report that their implementations recover nearly all cross-sectional stock-return predictors. Among 161 characteristics that were clearly significant in the original studies, 98% of their reproduced long-short portfolios have t-statistics above 1.96.

This evidence primarily addresses whether published empirical relationships can be reconstructed under a transparent and standardised implementation.

It is distinct from replication with genuinely new observations.

Replication introduces new data while retaining the underlying scientific question. It therefore provides evidence about whether the original claim survives beyond the observations on which it was initially established.

Extension goes further in a different direction by asking whether the relationship remains informative in materially different settings.

Jensen, Kelly and Pedersen (2023) provide an example of external-validity and extension evidence. Their global dataset evaluates 153 factors across 93 countries, allowing U.S.-centred findings to be examined in materially different geographic observations. The authors interpret this analysis as evidence of external validity across countries.

These categories should not be treated as a simple ranking.

A successful computational reproduction can establish that an original result was implemented as reported.

A failed independent reconstruction may reveal ambiguity or hidden methodological dependence.

A successful replication with new data provides evidence of generalisation to another sample.

A cross-market extension evaluates whether the relationship survives in a different economic environment.

Each strengthens or challenges a different part of the original claim.

6.6 Post-Discovery and Prospective Evidence

The cleanest form of temporal separation from the original historical discovery occurs when the evaluation observations did not yet exist when the model was developed.

Future observations cannot have influenced the original discovery because they had not yet occurred.

This makes post-discovery evidence conceptually important.

McLean and Pontiff (2016), as discussed in Section 5, distinguish performance within the original samples from performance after those samples and from performance after publication. Their research illustrates that these periods can produce materially different evidence about the same documented predictors.

Post-discovery data have a clear informational advantage over historical observations already accessible during model development:

they could not have been inspected before they occurred.

But even this advantage requires qualification.

A model can continue to be modified after deployment. Once post-discovery performance has been observed, changes made in response to that evidence become part of a new research cycle. Subsequent observations can remain prospective relative to the modified model, but earlier live evidence no longer constitutes untouched validation for that new specification.

The relevant evidentiary unit is therefore not simply the strategy name or model family.

It is the specific research specification evaluated against observations that did not influence that specification.

Prospective evaluation can also introduce complications that a historical backtest does not fully capture:

- data may arrive imperfectly and later be corrected;
- implementation decisions must be made in real time;
- market conditions may change;
- operational failures may affect realised outcomes;

- and transaction costs and execution can separate a predictive forecast from an economically realised result.

These complications do not make prospective evidence uninformative, but they broaden the object being evaluated and can make attribution more difficult. Prospective outcomes may reflect both the predictive relationship and the operational process through which it is implemented.

A historical backtest can isolate aspects of a predictive hypothesis under controlled assumptions.

Prospective observation tests the forecasting and decision process under conditions closer to those in which the information would actually be used.

A strong research programme should therefore distinguish:

historical discovery performance;

historical validation;

independent reconstruction or replication;

cross-market or cross-period extension;

and

prospective post-discovery evidence.

Each answers a different question.

The Validation Hierarchy

The framework developed in this section can be summarised as three principal forms of separation from the original discovery process.

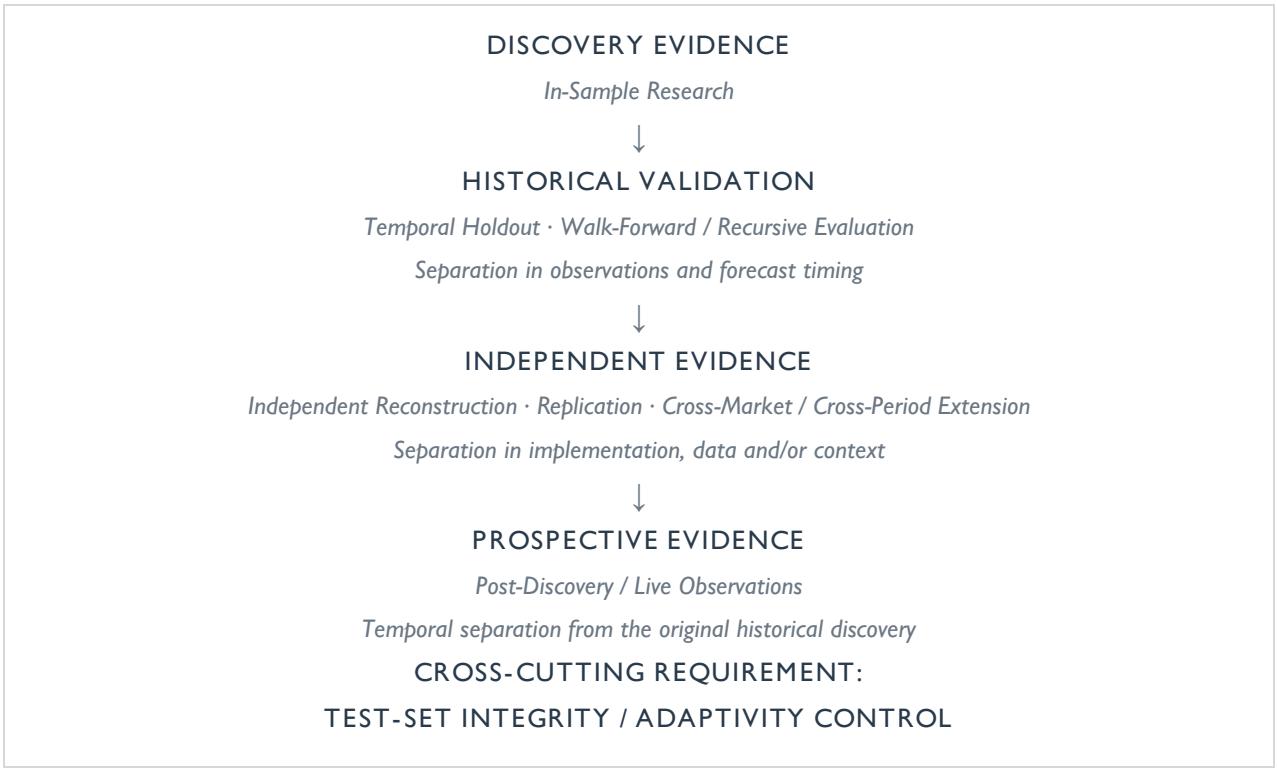


Figure 2. Validation Hierarchy.

Conceptual framework describing different forms of evidentiary separation between a predictive relationship and the process that generated it. The stages should not be interpreted as a strict ranking of statistical quality. Each addresses a different dimension of generalisation, independence or external validity.

The framework does not imply that a candidate must pass every category before any conclusion can be drawn.

Different research problems provide different validation opportunities. A long-horizon macroeconomic hypothesis may accumulate genuinely new observations only slowly. A cross-sectional equity relationship may instead be tested across additional markets. A high-frequency signal may generate large numbers of subsequent observations while still requiring careful treatment of statistical dependence.

The relevant principle is not the mechanical number of validation procedures applied.

It is the type and degree of evidentiary independence achieved.

A useful validation process should therefore ask:

Which observations generated the hypothesis?

Which observations influenced feature and model selection?

Which observations were used only after the specification had been fixed?

How were evaluation results inspected and acted upon?

Has the result been reconstructed independently?

Has the central claim survived genuinely new data?

Does the evidence extend to materially different periods, markets or environments?

Has the model survived observations that did not exist at discovery?

These questions provide a more informative description of validation strength than the single label "out of sample."

Validation also cannot resolve every remaining issue.

A predictive relationship can survive rigorous historical validation, replication or prospective observation and still be economically irrelevant. It may generate an improvement that is too small, too costly, too capacity-constrained or too short-lived to use in practice.

The final stage of signal qualification therefore moves from the existence and robustness of predictive information to its economic usability.

7. From Statistical Evidence to Economic Usability

A predictive relationship can survive data-quality controls, model-search adjustments and genuine out-of-sample validation without becoming an economically usable signal.

This final distinction is essential because statistical, predictive and economic questions are different.

Statistical analysis evaluates the evidence for a relationship under a specified model, inferential procedure and set of assumptions. Predictive evaluation asks whether that relationship improves forecasts relative to an appropriate benchmark. Economic evaluation asks whether the magnitude and persistence of that improvement matter for the intended decision.

Implementation introduces a further requirement.

Implementation can introduce a broader set of costs and frictions, including explicit fees, bid–ask costs, market impact, financing and borrowing costs, and the risk or opportunity cost associated with executing a decision over time. The magnitude and timing of these frictions can depend on turnover, liquidity, position size, trading urgency and the execution process itself.

For this reason, the final transition in the Signal Qualification Framework is not:

STATISTICAL SIGNIFICANCE → TRADABLE SIGNAL

but:

STATISTICAL EVIDENCE → PREDICTIVE MAGNITUDE → ECONOMIC RELEVANCE → IMPLEMENTATION COSTS & FRICTIONS → CAPACITY & EXECUTION → EXECUTABILITY

The distinction also implies that economic usability is conditional. The same predictive relationship can be economically meaningful for one investor, portfolio size or trading horizon and unattractive for another.

7.1 Statistical Significance versus Effect Size

Statistical significance does not measure the magnitude or economic importance of a relationship.

As discussed in Section 1, the American Statistical Association explicitly cautions that a p-value does not measure effect size or the importance of a result. A highly precise estimate of a very small predictive relationship can therefore be statistically compelling while remaining economically unimportant.

The reverse distinction is also relevant. A potentially meaningful effect may be estimated imprecisely when the available sample is limited or the target variable contains substantial unpredictable variation. Statistical uncertainty should not be ignored, but neither should statistical significance be treated as a direct measure of practical value.

For signal research, the first economic question is therefore:

How large is the predictive improvement?

The answer cannot be expressed by a universal minimum coefficient, return or forecasting statistic. Magnitude must be evaluated relative to the target, horizon, risk undertaken, forecast benchmark and intended decision process.

Cederburg, Johnson and O'Doherty (2023) demonstrate this distinction directly in stock-market return prediction. Examining 26 predictors from the perspective of Bayesian investors, they show that statistically strong predictors can nevertheless have limited economic value when, for example, extreme forecasts occur primarily in periods of high volatility, predictors have low persistence or their distributions exhibit fat tails.

Their results illustrate that the economic value of predictability depends not only on statistical strength but also on the conditions under which the predictive variation occurs.

This point changes how effect size should be interpreted.

An identical expected-return forecast does not necessarily have identical economic value if one forecast occurs when risk is low and another when risk is exceptionally high. Similarly, a modest forecast improvement that is persistent and inexpensive to implement may potentially be more useful than a larger but unstable forecast requiring aggressive portfolio changes.

Effect size is therefore necessary for economic assessment but is not sufficient by itself.

The relevant object is the effect in the context of the decision it is intended to support.

7.2 Economic Significance

Economic significance concerns whether predictive information materially improves the objective relevant to the intended decision maker, conditional on the risks, preferences, constraints and horizon of that decision.

It is therefore investor- and objective-specific and should not be treated as synonymous with realised profitability.

For a portfolio investor, relevant measures might include expected utility, certainty-equivalent return, risk-adjusted performance, drawdown characteristics or another prespecified portfolio criterion. For a forecasting system that does not directly determine portfolio weights, economic relevance may instead concern whether the improvement is sufficiently large and stable to justify incorporating the information into a later decision process.

There is therefore no universal conversion from a statistical test statistic into economic significance.

The distinction can be expressed as:

STATISTICAL QUESTION

Is there statistical support for predictive information under the stated research design?

ECONOMIC QUESTION

Does the magnitude of that information materially improve the objective relevant to the intended decision maker?

The two questions can produce different answers.

Cederburg et al. (2023) provide an example of statistically strong return predictors whose economic value can remain limited under their investor-based framework. Conversely, Campbell and Thompson (2008), discussed earlier in this paper, demonstrate that relatively small improvements in out-of-sample forecasting performance can still produce meaningful gains for a mean-variance investor under their assumptions.

These findings are not contradictory. They reinforce the same principle:

economic significance is a property of the interaction between predictive information and the decision problem, not of a statistical metric in isolation.

This qualification becomes even more important once implementation is introduced.

A relationship can appear economically attractive before implementation costs and frictions and unattractive afterward.

Equally, a relationship that appears unattractive under a naive high-turnover implementation may become usable when the portfolio is constructed in a way that reduces unnecessary trading.

Economic evaluation should therefore proceed from gross potential value to net implementable value rather than assuming that the portfolio implied by an academic signal is unique.

7.3 Turnover and Transaction Costs

Turnover provides one of the most direct links between a predictive signal and the cost of acting on it.

A slowly changing predictor may require relatively infrequent portfolio adjustment. A rapidly changing signal, by contrast, can require repeated trading simply to maintain the desired exposure. Even moderate per-trade costs can therefore become economically important when turnover is high.

Transaction costs themselves are multidimensional.

They can include:

- commissions and fees;
- bid–ask costs;
- taxes or levies where applicable;
- financing and securities-borrowing costs;
- price impact;
- and the opportunity cost associated with incomplete or delayed execution.

The relevant cost is not necessarily the value of one isolated trade. It is the cumulative implementation burden implied by the entire strategy.

The momentum literature provides a useful illustration of why this distinction matters.

Lesmond, Schill and Zhou (2004) examine whether previously documented momentum profits survive estimates of transaction costs and conclude that, under their methodology, the high trading costs associated with momentum portfolios place severe limits on their apparent profitability. Their analysis is particularly concerned with the concentration of trading in relatively costly securities.

Korajczyk and Sadka (2004), using intraday data and explicit models of proportional and price-impact costs, reach a more conditional conclusion. They find that momentum profitability declines with portfolio size, but liquidity-aware portfolio construction can materially improve implementability. In their historical application, some liquidity-weighted momentum strategies support estimated break-even fund sizes of approximately USD 5 billion or more, measured relative to December 1999 market capitalisation, before abnormal returns are eliminated.

Novy-Marx and Velikov (2016) broaden the analysis across a larger anomaly universe. They report that transaction-cost mitigation techniques materially affect net performance. In their study, most anomalies with monthly turnover below 50% generate significant net spreads when designed to reduce trading costs, whereas few higher-turnover anomalies do. They also find that the extent to which additional capital reduces profitability is strongly related to turnover.

These findings are not directly contradictory because they evaluate different implementations, cost models, securities, periods and scales.

In particular, the USD 5 billion break-even figures reported by Korajczyk and Sadka are estimates from their historical momentum application and are scaled relative to December 1999 market capitalisation. They should not be interpreted as current or universal capacity thresholds.

Similarly, the turnover thresholds reported by Novy-Marx and Velikov describe the anomaly universe and implementation procedures examined in their study rather than a general rule for financial strategies.

The combined evidence demonstrates instead that implementation depends on:

the signal; the turnover it induces; the securities traded; the portfolio construction method; the assumed cost model; the investor; and the scale of implementation.

Patton and Weller (2020) provide an additional perspective by estimating the gap between academic factor returns and the returns actually achieved by mutual funds with corresponding factor exposures. Their method is designed to capture broad implementation costs without requiring an explicit microstructure model. In their application, typical mutual funds earn low returns to value and no returns to momentum after estimated implementation costs, with substantial heterogeneity across investors and time.

Patton and Weller also emphasise important limits to their approach. Their estimates can represent lower bounds on implementation costs for a representative mutual fund, and the method is not designed to estimate the carrying capacity of counterfactual factor strategies.

Their evidence is therefore particularly informative about realised implementation gaps for the investor population they study rather than providing a universal transaction-cost or capacity model.

The combined literature argues against evaluating signals using a fixed universal transaction-cost deduction.

Implementation is not a fixed haircut applied to a theoretical return. It is part of the economic definition of the strategy being evaluated.

7.4 Liquidity, Market Impact and Capacity

Liquidity, market impact and capacity are related but distinct concepts.

Liquidity describes the trading environment of the instruments through which a strategy is implemented, including the ability to transact relevant quantities within a relevant time horizon at acceptable cost.

Market impact describes the effect that the investor's own trading can have on realised execution prices.

Capacity describes how the net economic value of a strategy changes as the amount of capital deployed through that strategy increases.

Market impact is one mechanism through which limited liquidity can constrain capacity, but capacity can also depend on turnover, portfolio breadth, shorting or financing constraints, signal persistence and the distribution of available opportunities.

Bertsimas and Lo (1998) formalise the execution problem by deriving dynamic trading strategies that minimise expected execution cost when execution prices depend on trade size and prevailing market conditions. Their framework makes explicit that the cost of implementing a position depends on the trading path rather than simply on a fixed fee per unit traded.

Almgren and Chriss (2001) analyse the trade-off between expected execution costs, including temporary and permanent market impact, and the risk created by allowing execution to occur over time. Their framework illustrates why trading faster can reduce exposure to subsequent price movements while increasing impact costs, whereas trading more slowly can reduce immediate market pressure while increasing execution risk.

These models provide theoretical frameworks for understanding the interaction between trade size, execution speed, impact and risk. They do not imply that a single functional form or parameterisation describes execution costs across all markets or investors.

For investment signals, market impact creates an important distinction between gross return opportunity and capacity.

A strategy may have positive expected net value at modest scale while that value declines as more capital is allocated.

Korajczyk and Sadka (2004) explicitly estimate this relationship for momentum strategies. Their price-impact models imply declining abnormal returns as portfolio size increases, allowing the authors to calculate break-even fund sizes at which estimated abnormal returns are reduced to zero.

Novy-Marx and Velikov (2016) reach a related conclusion across their anomaly set: strategy capacity varies substantially and is related to turnover; size, value and profitability strategies exhibit comparatively greater capacity in their empirical analysis.

Capacity should therefore not be interpreted simply as a maximum amount of assets that can be managed.

For analytical purposes, it is more useful to treat capacity as a relationship between scale and net economic opportunity.

As capital rises:

required trade size may rise; market participation may increase; market impact may increase; execution may need to become slower; gross exposure may need to deviate from the theoretical target; and net expected value may decline.

The point at which implementation becomes unattractive depends on the strategy, investor objective and market environment.

Capacity is consequently strategy-, investor-, implementation- and market-state-specific. It should not be treated as a permanent numerical characteristic of an underlying statistical signal.

A capacity estimate is conditional on the scale, market environment and implementation assumptions under which it was obtained.

7.5 Execution Timing and Signal Decay

Execution introduces a time dimension that is distinct from the longer-term signal decay discussed in Section 5.

A predictive relationship may be valid at the moment it is observed but lose economic value before the intended position can be established.

The relevant question is therefore not only:

Does the signal predict the future?

but:

Does sufficient predictive value remain at the point at which the signal can actually be acted upon?

This problem is particularly important for information with a short forecast horizon or rapid mean reversion.

Gârleanu and Pedersen (2013) develop a dynamic portfolio model in which returns are predicted by signals with different mean-reversion speeds and trading is costly. Their optimal policy places greater weight on predictors whose expected-return signals decay more slowly and trades only partially toward the desired portfolio because immediate adjustment incurs costs. In their commodity-futures application, the resulting dynamic policy achieves higher net returns than the naive benchmarks they consider.

The result provides a useful conceptual link between signal persistence and implementation.

A slowly decaying signal gives an investor more time to trade.

A rapidly decaying signal requires greater urgency if its forecast is to be captured, but greater urgency can itself increase execution cost and market impact.

The two forces operate in opposite directions:

WAIT LONGER → potentially lower trading pressure → greater risk that predictive value decays

TRADE FASTER → potentially preserve more of the signal → potentially higher market impact and execution cost

Execution is therefore an optimisation problem, not merely a subtraction performed after a theoretical portfolio return has been calculated.

Perold's (1988) concept of implementation shortfall provides a complementary perspective by distinguishing the performance of an investment decision on paper from the result achieved when that decision is implemented in actual markets.

For analytical purposes, this paper refers to the loss of predictive value between signal formation and feasible position establishment as decision-to-execution decay.

The term is not intended as a standard transaction-cost measure and is narrower than Perold's implementation shortfall, which measures the broader difference between a paper investment decision and its realised implementation.

Decision-to-execution decay is also distinct from the longer-term structural or post-publication signal decay examined in Section 5.

A signal can therefore remain persistent across years while being operationally difficult to capture if its useful predictive information decays within minutes or hours after each observation.

7.6 The Executability Test

The final stage of signal qualification asks whether the candidate retains sufficient economic value under an implementation that could plausibly be executed by the intended market participant.

This paper refers to that assessment as the Executability Test.

The test is conceptual rather than a universal numerical formula because different strategies, asset classes and investors face different constraints.

A candidate signal should be evaluated across at least six dimensions.

1. Predictive magnitude

Does the signal produce a sufficiently large improvement relative to the benchmark to justify further economic evaluation?

2. Risk-adjusted relevance

Does the forecast improve the intended decision once the risk associated with acting on it is incorporated?

3. Turnover and recurring costs

How frequently must the portfolio change to retain the intended exposure, and what cumulative costs and frictions does that imply?

4. Liquidity and capacity

Can the required positions be established at the intended scale without implementation effects materially reducing the economic opportunity?

5. Timing

Does sufficient predictive information remain after realistic decision, routing and execution delays?

6. Net robustness

Does the economic result remain sufficiently valuable under reasonable variation in cost, execution, liquidity and timing assumptions rather than depending on one favourable implementation estimate?

These dimensions separate signal validity from strategy usability.

A candidate can fail the Executability Test without invalidating the underlying statistical relationship.

For example, a signal may provide genuine incremental forecasts but induce too much turnover. Another may be statistically and economically attractive at modest capital but lose value when market impact is introduced at institutional scale. A third may possess substantial gross predictive value but decay faster than the required trades can reasonably be completed.

Conversely, implementation research can improve usability without changing the underlying signal.

Novy-Marx and Velikov's evidence that transaction-cost-aware portfolio design can materially improve the net performance of lower-turnover anomalies illustrates this point.

The final qualification chain is therefore:

PREDICTIVE INFORMATION

↓

STATISTICAL ROBUSTNESS

↓

ECONOMIC MAGNITUDE

↓

IMPLEMENTATION-COST RESILIENCE

↓

LIQUIDITY & CAPACITY

↓

TIMING & EXECUTION

↓

EXECUTABLE SIGNAL

Table 2. From Statistical Evidence to Economic Usability

Evaluation Stage	Principal Question	What a Positive Result Establishes	What It Does Not Establish
Statistical Evidence	Is the observed relationship supported under the stated statistical design?	Statistical support for the relationship under the stated model, assumptions and inferential procedure	Economic importance
Predictive Information	Does the relationship improve forecasting relative to the stated benchmark?	Incremental forecast content	Robustness across time or implementation
Robustness / Validation	Does the relationship survive evidence separated from discovery?	Greater confidence in generalisation	Economic usability
Economic Relevance	Is the magnitude valuable for the intended decision and risk objective?	Potential gross economic value for the defined decision maker	Net implementability
Implementation-Cost Resilience	Does value survive recurring implementation costs and frictions?	Potential net value under stated assumptions	Capacity at larger scale
Liquidity & Capacity	Does value survive position size, liquidity constraints and market impact at intended scale?	Scale-compatible opportunity under stated assumptions	Timely execution
Execution	Can the intended position be established before relevant predictive value dissipates?	Evidence that the intended exposure can be established within the tested timing and execution assumptions	Permanent future persistence
Executable Signal	Does sufficient net economic value remain across the complete implementation	Evidence of economic usability for the defined investor and setting	A guarantee of future profitability

Evaluation Stage	Principal Question	What a Positive Result Establishes	What It Does Not Establish
	chain?		

The final row requires the strongest qualification.

An executable signal is not a claim of certain profitability.

It is a statement that, under the information, assumptions, scale and implementation conditions tested, the predictive relationship has survived the evidentiary stages required for a plausible economic application.

Those conditions can change.

Transaction costs change. Liquidity changes. Capacity changes with assets deployed and market conditions. Competing participants adapt. Signals decay. New observations alter the evidence.

Executability should therefore be monitored rather than treated as a permanent property obtained once and retained indefinitely.

From Signal to Usable Information

The analysis developed throughout this paper now returns to the distinction introduced at the outset.

Financial-market data can contain patterns.

Some patterns survive statistical testing.

A smaller set may provide incremental predictive information.

Some of those relationships survive model search, temporal instability and genuinely separated validation.

Fewer still may remain economically relevant after the costs and constraints required to act upon them are incorporated.

The progression from data to signal is therefore incomplete until implementation is considered.

A statistically supported predictor identifies potential information. An executable signal identifies information that remains economically relevant under a defined implementation environment.

The distinction is critical because financial research can otherwise confuse three separate achievements:

finding a relationship;

forecasting an outcome;

and

capturing economic value from that forecast.

The final conclusion of this paper brings these stages together into a single framework for distinguishing noise, statistical evidence, predictive information and economically executable signal.

8. Conclusion — From Noise to Executable Information

Financial markets generate enormous quantities of observable data.

Prices, volumes, spreads, order flow, fundamentals, macroeconomic variables, textual information and increasingly complex derived datasets create a research environment in which large numbers of potential patterns can be identified and tested.

Identifying information remains a different problem.

The central argument of this paper is that observable structure should not be confused with predictive signal.

A relationship becomes a credible candidate for predictive information only after a sequence of increasingly demanding questions has been addressed:

Was the information genuinely available at the forecast date?

Was the prediction problem defined before the result was interpreted?

Is the relationship more than a consequence of measurement, sampling or preprocessing choices?

Does it add predictive information relative to a stated benchmark?

Has the evidence been interpreted in light of model search and multiple testing?

Does the relationship survive temporal instability, changing regimes or post-discovery deterioration?

Does it persist when evaluated using evidence sufficiently separated from the discovery process?

Is the magnitude economically relevant?

Can the information be acted upon after realistic implementation costs, liquidity constraints, capacity limits and execution delays are incorporated?

Each question addresses a different source of potential false confidence.

More Data, More Search and More False Discovery

The growth of financial data expands the opportunity to discover economically meaningful relationships.

It simultaneously expands the opportunity to discover relationships that appear meaningful only because researchers can search more variables, transformations, horizons, markets and model specifications.

This is the central paradox of modern quantitative research.

More observations can improve measurement and estimation.

More candidate variables can reveal information that would otherwise remain hidden.

Greater computational resources can expand the range of models, specifications and validation procedures that researchers are able to examine.

But none of these developments guarantees that the proportion of useful information rises with the volume of data available.

As the search space expands, the burden of evidence must expand with it.

A statistically unusual result drawn from one prespecified hypothesis does not have the same evidentiary meaning as the same result selected from thousands of unsuccessful alternatives.

The research process therefore matters as much as the final statistic.

Information Is Defined Relative to a Prediction Problem

There is no signal in the abstract.

Predictive information exists relative to a defined:

target; forecast horizon; information set; benchmark; and evaluation criterion.

A variable may contain information for one horizon and none for another.

A relationship may improve forecasts relative to a simple benchmark while adding little relative to a stronger competing model.

A predictor that appears valuable in one representation may become redundant when the information contained in other variables is considered.

For this reason, the relevant question is not whether a variable is correlated with an outcome.

It is whether the variable contains incremental predictive information for the defined forecasting problem.

That distinction separates pattern recognition from forecasting evidence.

Data Construction Is Part of Inference

Before statistical testing begins, the dataset has already embodied a series of research decisions.

Point-in-time availability, revisions, survivorship, sample selection, sampling frequency, market microstructure and preprocessing can all change the informational content of the observations being analysed.

A dataset can therefore be internally clean while remaining unsuitable for historical inference.

The required standard is not simply database accuracy.

It is information-set integrity.

Historical research should reconstruct what could have been known when the forecast was formed rather than what can be observed retrospectively in a modern database.

The same principle applies to transformations.

Preprocessing cannot legitimately use future information merely because the final transformed variable appears historically well behaved.

Measurement and preprocessing are consequently part of the inferential design, not administrative steps performed before the research begins.

Statistical Evidence Must Be Search-Aware

Even correctly constructed data can produce misleading evidence when researchers search extensively.

Multiple testing, data snooping, model selection, selective reporting and publication incentives can all increase the probability that apparently compelling results emerge from weak or nonexistent underlying relationships.

These mechanisms should not be treated as interchangeable.

Transparent multiple testing is not the same as p-hacking.

Model selection is not identical to publication bias.

Data snooping is not equivalent to every form of exploratory research.

But they share a common implication:

the evidentiary meaning of a result depends partly on the process through which it was selected.

Search is not itself a research failure.

Exploration is essential to scientific discovery.

The methodological requirement is that claims made after extensive search should be evaluated using procedures capable of recognising that search or through subsequent evidence sufficiently independent of it.

Predictive Relationships Are Not Necessarily Stable

A relationship that survives initial statistical scrutiny can still change through time.

Financial markets are not fixed systems.

Market structure, regulation, technology, participation, liquidity, risk preferences and economic conditions evolve. Predictive relationships can therefore weaken, strengthen or become conditional on different market environments.

Observed deterioration has several possible interpretations.

It may reveal an initially overstated relationship.

It may reflect structural change.

It may indicate regime dependence.

It may be consistent with market learning or adaptation.

It may arise because implementation economics changed even though some predictive information remains.

These mechanisms are not mutually exclusive.

The critical methodological principle is therefore:

Observed deterioration is evidence about stability. It is not, by itself, evidence about mechanism.

Signal decay should be diagnosed rather than narrated retrospectively.

Validation Is Separation, Not a Label

Out-of-sample testing is central to quantitative research, but the label itself is insufficient.

A test period can be excluded mechanically from model estimation while still influencing feature selection, parameter tuning or specification choice through repeated inspection.

Conversely, specialised statistical procedures can permit forms of adaptive reuse when that adaptivity is explicitly accounted for.

The evidentiary meaning of OOS therefore depends on how evaluation data entered the research process.

Validation can take several forms:

historical temporal holdout; walk-forward or recursive evaluation; independent reconstruction; replication with new data; cross-market or cross-period extension; and prospective post-discovery evidence.

These procedures do not form a universal ranking in which one method automatically dominates another.

They create different forms of separation between discovery and evaluation.

The evidentiary meaning of a validation result is clearest when the source and degree of separation from the discovery process are made explicit.

Economic Usability Is a Separate Test

A statistically robust predictor is not necessarily economically useful.

A predictor that improves forecasts is not necessarily implementable.

Economic significance depends on the decision maker, objective, horizon, risk, constraints and implementation environment.

Implementation then introduces a further set of conditions.

Turnover can convert small recurring costs into material economic losses.

Liquidity can constrain the quantities that can be traded efficiently.

Market impact can cause execution costs to rise with scale.

Net economic value can decline as the amount of capital deployed through a strategy increases.

Signal value can decay between observation and execution.

The same underlying predictive relationship can therefore be usable for one investor or portfolio size and unattractive for another.

The appropriate final question is not:

Does the signal work?

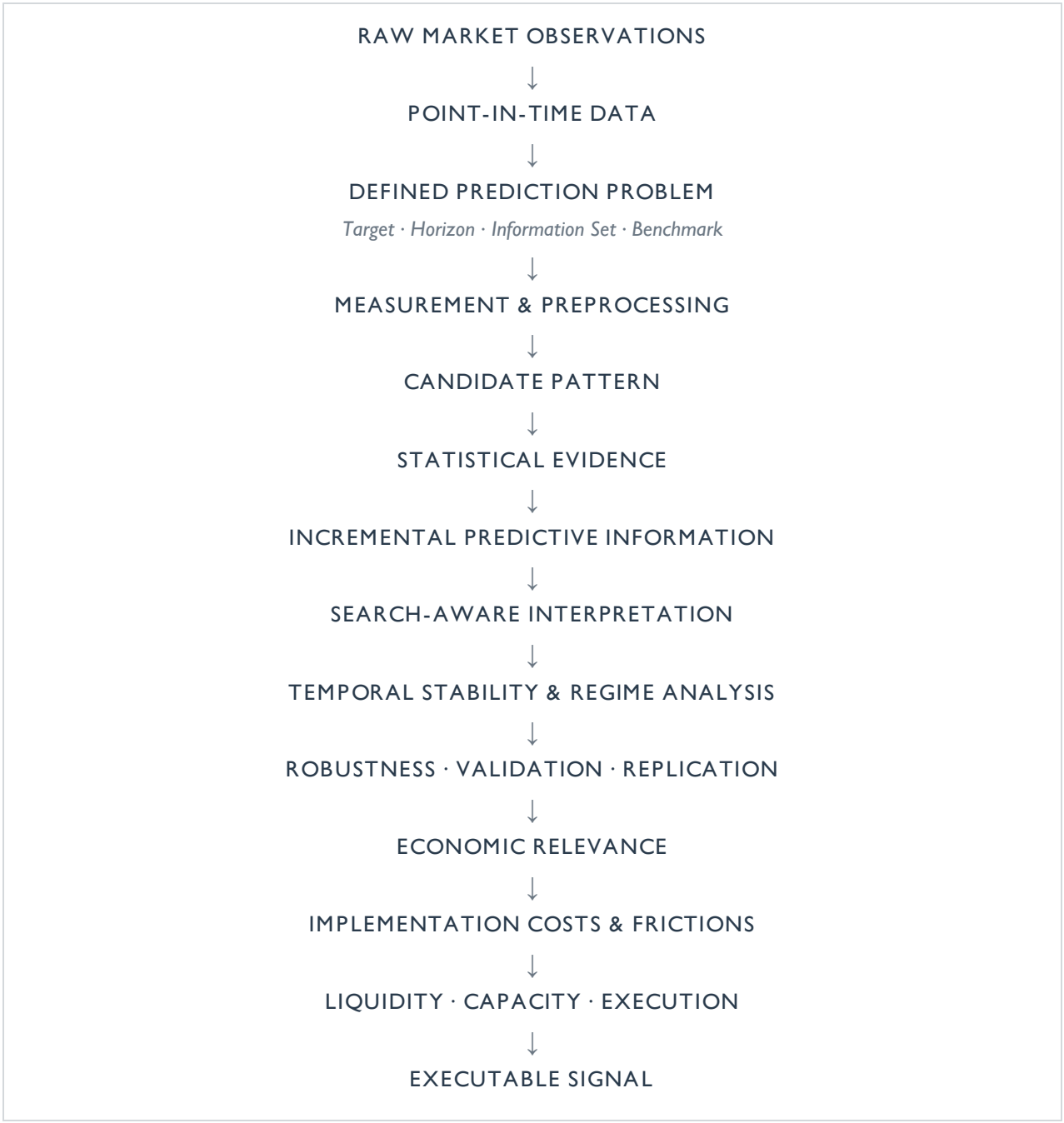
It is:

Under what defined conditions does sufficient predictive value remain after the complete implementation process is considered?

That is the purpose of the Executability Test developed in Section 7.

The Expanded Signal Qualification Sequence

The framework introduced at the outset of this paper can now be expanded to show the principal analytical stages developed across Sections 1–7:



This expanded sequence elaborates the Signal Qualification Framework introduced earlier in the paper rather than replacing it.

The framework does not assume that every observable pattern will reach its final stage. Its purpose is precisely to provide successive opportunities for weak or insufficiently supported candidates to be rejected before economic interpretation or capital is attached to them.

That is not a weakness of quantitative research.

It is the purpose of the research process.

A rigorous methodology should eliminate weak candidates before capital, confidence or interpretation is attached to them.

From Noise to Information

The distinction between noise and signal is therefore not determined by whether a pattern can be found in historical data.

Nor is it determined by whether the pattern generates an attractive chart, a statistically significant coefficient or a strong backtest.

A credible signal is the result of accumulated evidence.

It begins with valid data and a defined prediction problem.

It requires incremental predictive content rather than correlation alone.

It must be interpreted in light of the search process that produced it.

Its stability must be examined through time and across changing conditions.

Its evidence should survive evaluation sufficiently separated from discovery.

Its magnitude must matter economically.

And its value must remain after the constraints required to act upon it are incorporated.

This leads to the final distinction of the paper:

For the purposes of this paper, noise denotes observed variation or apparent pattern for which sufficient predictive information has not been established.

A predictive signal is information that improves a defined forecast relative to a stated benchmark and survives appropriate statistical and validation scrutiny.

An executable signal is predictive information that remains economically relevant under a defined, realistically implementable set of market and execution conditions.

These categories should not be treated as permanent labels.

New data can strengthen evidence.

New regimes can weaken it.

Implementation conditions can change.

A relationship that once appeared unusable may become economically relevant under a different cost or execution environment, while a historically successful signal may cease to justify action.

Within the framework developed in this paper, quantitative signal research is therefore treated as an ongoing process of evaluating whether observable relationships contain information that remains statistically defensible, predictively useful and economically actionable under the conditions in which decisions must actually be made.

The abundance of modern financial data does not remove this evidentiary burden.

It makes disciplined signal qualification increasingly important.

More data create more opportunities to search. Disciplined research is required to assess which of those opportunities contain sufficiently supported predictive information.

References

Aït-Sahalia, Y., Mykland, P. A., & Zhang, L. (2005). How often to sample a continuous-time process in the presence of market microstructure noise. *The Review of Financial Studies, 18*(2), 351–416. doi:10.1093/rfs/hhi016.

Almgren, R., & Chriss, N. (2001). Optimal execution of portfolio transactions. *The Journal of Risk, 3*(2), 5–39. doi:10.21314/JOR.2001.041.

Andersen, T. G., Bollerslev, T., Diebold, F. X., & Ebens, H. (2001). The distribution of realized stock return volatility. *Journal of Financial Economics, 61*(1), 43–76. doi:10.1016/S0304-405X(01)00055-1.

Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. *The Review of Financial Studies, 15*(4), 1137–1187. doi:10.1093/rfs/15.4.1137.

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2017). The probability of backtest overfitting. *The Journal of Computational Finance, 20*(4), 39–69. doi:10.21314/JCF.2016.322.

Bailey, D. H., & López de Prado, M. (2014). The deflated Sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management, 40*(5), 94–107. doi:10.3905/jpm.2014.40.5.094.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis, 120*, 70–83. doi:10.1016/j.csda.2017.11.003.

Bertsimas, D., & Lo, A. W. (1998). Optimal control of execution costs. *Journal of Financial Markets, 1*(1), 1–50. doi:10.1016/S1386-4181(97)00012-8.

Brodeur, A., Carrell, S. E., Figlio, D. N., & Lusher, L. R. (2023). Unpacking p-hacking and publication bias. *American Economic Review, 113*(11), 2974–3002. doi:10.1257/aer.20210795.

Brown, S. J., Goetzmann, W., Ibbotson, R. G., & Ross, S. A. (1992). Survivorship bias in performance studies. *The Review of Financial Studies, 5*(4), 553–580. doi:10.1093/rfs/5.4.553.

Campbell, J. Y., & Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *The Review of Financial Studies, 21*(4), 1509–1531. doi:10.1093/rfs/hhm055.

Cederburg, S., Johnson, T. L., & O'Doherty, M. S. (2023). On the economic significance of stock return predictability. *Review of Finance, 27*(2), 619–657. doi:10.1093/rof/rfac035.

Chen, A. Y. (2021). The limits of p-hacking: Some thought experiments. *The Journal of Finance, 76*(5), 2447–2480. doi:10.1111/jofi.13036.

Chen, A. Y., & Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review, 11*(2), 207–264. doi:10.1561/104.00000112.

Chordia, T., Goyal, A., & Saretto, A. (2020). Anomalies and false rejections. *The Review of Financial Studies, 33*(5), 2134–2179. doi:10.1093/rfs/hhaa018.

Clark, T. E., & West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics, 138*(1), 291–311. doi:10.1016/j.jeconom.2006.05.023.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance, 66*(4), 1047–1108. doi:10.1111/j.1540-6261.2011.01671.x.

Croushore, D., & Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics, 105*(1), 111–130. doi:10.1016/S0304-4076(01)00072-0.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics, 13*(3), 253–263. doi:10.1080/07350015.1995.10524599.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., & Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science, 349*(6248), 636–638. doi:10.1126/science.aaa9375.

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance, 75*(3), 1327–1370. doi:10.1111/jofi.12883.

Gârleanu, N., & Pedersen, L. H. (2013). Dynamic trading with predictable returns and transaction costs. *The Journal of Finance, 68*(6), 2309–2340. doi:10.1111/jofi.12080.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. *Econometrica, 74*(6), 1545–1578. doi:10.1111/j.1468-0262.2006.00718.x.

Goyal, A., Welch, I., & Zafirov, A. (2024). A comprehensive 2022 look at the empirical performance of equity premium prediction. *The Review of Financial Studies, 37*(11), 3490–3557. doi:10.1093/rfs/hhae044.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*(3), 424–438. doi:10.2307/1912791.

Green, J., Hand, J. R. M., & Zhang, X. F. (2017). The characteristics that provide independent information about average U.S. monthly stock returns. *The Review of Financial Studies, 30*(12), 4389–4436. doi:10.1093/rfs/hhx019.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33*(5), 2223–2273. doi:10.1093/rfs/hhaa009.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica, 57*(2), 357–384. doi:10.2307/1912559.

Hansen, P. R., & Lunde, A. (2006). Realized variance and market microstructure noise. *Journal of Business & Economic Statistics, 24*(2), 127–161. doi:10.1198/073500106000000071.

Harvey, C. R. (2017). Presidential address: The scientific outlook in financial economics. *The Journal of Finance, 72*(4), 1399–1440. doi:10.1111/jofi.12530.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *The Review of Financial Studies, 29*(1), 5–68. doi:10.1093/rfs/hhv059.

Harvey, D. I., Leybourne, S. J., & Newbold, P. (1998). Tests for forecast encompassing. *Journal of Business & Economic Statistics, 16*(2), 254–259. doi:10.1080/07350015.1998.10524759.

Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. *The Review of Financial Studies, 33*(5), 2019–2133. doi:10.1093/rfs/hhy131.

Jacobs, H., & Müller, S. (2020). Anomalies across the globe: Once public, no longer existent? *Journal of Financial Economics, 135*(1), 213–230. doi:10.1016/j.jfineco.2019.06.004.

Jensen, T. I., Kelly, B., & Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance, 78*(5), 2465–2518. doi:10.1111/jofi.13249.

Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data, 6*(4), Article 15. doi:10.1145/2382577.2382579.

Korajczyk, R. A., & Sadka, R. (2004). Are momentum profits robust to trading costs? *The Journal of Finance, 59*(3), 1039–1082. doi:10.1111/j.1540-6261.2004.00656.x.

Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics, 135*(2), 271–292. doi:10.1016/j.jfineco.2019.06.008.

Lesmond, D. A., Schill, M. J., & Zhou, C. (2004). The illusory nature of momentum profits. *Journal of Financial Economics, 71*(2), 349–380. doi:10.1016/S0304-405X(03)00206-X.

Lettau, M., & Van Nieuwerburgh, S. (2008). Reconciling the return predictability evidence. *The Review of Financial Studies, 21*(4), 1607–1652. doi:10.1093/rfs/hhm074.

Li, K., Li, Y., Lyu, C., & Yu, J. (2025). How to dominate the historical average. *The Review of Financial Studies, 38*(10), 3086–3116. doi:10.1093/rfs/hhaf010.

Lo, A. W., & MacKinlay, A. C. (1990). Data-snooping biases in tests of financial asset pricing models. *The Review of Financial Studies, 3*(3), 431–467. doi:10.1093/rfs/3.3.431.

McLean, R. D., & Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance, 71*(1), 5–32. doi:10.1111/jofi.12365.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. doi:10.17226/25303.

Novy-Marx, R., & Velikov, M. (2016). A taxonomy of anomalies and their trading costs. *The Review of Financial Studies, 29*(1), 104–147. doi:10.1093/rfs/hhv063.

Oomen, R. C. A. (2006). Properties of realized variance under alternative sampling schemes. *Journal of Business & Economic Statistics, 24*(2), 219–237. doi:10.1198/073500106000000044.

Patton, A. J., & Weller, B. M. (2020). What you see is not what you get: The costs of trading market anomalies. *Journal of Financial Economics, 137*(2), 515–549. doi:10.1016/j.jfineco.2020.02.012.

Perold, A. F. (1988). The implementation shortfall: Paper versus reality. *The Journal of Portfolio Management, 14*(3), 4–9. doi:10.3905/jpm.1988.409150.

Pesaran, M. H., & Timmermann, A. (1995). Predictability of stock returns: Robustness and economic significance. *The Journal of Finance, 50*(4), 1201–1228. doi:10.1111/j.1540-6261.1995.tb04055.x.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. *Econometrica, 73*(4), 1237–1282. doi:10.1111/j.1468-0262.2005.00615.x.

Rossi, B. (2021). Forecasting in the presence of instabilities: How we know whether models predict well and how to improve them. *Journal of Economic Literature, 59*(4), 1135–1190. doi:10.1257/jel.20201479.

Shumway, T. (1997). The delisting bias in CRSP data. *The Journal of Finance, 52*(1), 327–340. doi:10.1111/j.1540-6261.1997.tb03818.x.

Sullivan, R., Timmermann, A., & White, H. (1999). Data-snooping, technical trading rule performance, and the bootstrap. *The Journal of Finance, 54*(5), 1647–1691. doi:10.1111/0022-1082.00163.

Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*(4), 437–450. doi:10.1016/S0169-2070(00)00065-0.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129–133. doi:10.1080/00031305.2016.1154108.

Welch, I., & Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *The Review of Financial Studies, 21*(4), 1455–1508. doi:10.1093/rfs/hhm014.

White, H. (2000). A reality check for data snooping. *Econometrica, 68*(5), 1097–1126. doi:10.1111/1468-0262.00152.

Research Disclaimer

This publication has been prepared by Finovis Software Solutions LLC-FZ for general research, informational and educational purposes only.

The information, analysis and views contained in this publication do not constitute investment advice, financial advice, legal advice, tax advice or any other form of professional advice. Nothing in this publication constitutes or should be construed as investment advice, an investment recommendation, or an offer, solicitation or invitation to acquire, dispose of or otherwise transact in any financial instrument or to participate in any investment strategy.

This publication is thematic and methodological in nature. It does not provide a recommendation or suggestion concerning the purchase, sale, holding, valuation or expected performance of any specific financial instrument, issuer or investment strategy.

The research presented in this publication is not tailored to the investment objectives, financial circumstances, risk tolerance or particular needs of any individual or institution. Readers should conduct their own independent assessment and, where appropriate, consult qualified professional advisers before making investment or financial decisions.

This publication draws on academic research, publicly available information and analytical synthesis. External sources are identified where applicable; however, FINOVIS has not independently verified every underlying dataset, calculation or empirical result reported in third-party research.

Finovis Software Solutions LLC-FZ makes no representation or warranty, express or implied, as to the accuracy, completeness, timeliness or continued validity of third-party information or external empirical findings referenced in this publication.

The analytical frameworks, classifications and sequences developed in this publication represent conceptual research synthesis for the purposes of this paper. They do not disclose, describe or represent any proprietary FINOVIS trading model, investment strategy, forecasting system, signal architecture or production trading methodology.

Historical findings, backtested results, simulated outcomes and empirical relationships discussed in the academic literature are inherently dependent on the data, methodology, assumptions, market conditions and periods examined. They are not guarantees of future performance, continued predictability or economic profitability.

Financial markets and financial instruments involve risk. Market conditions, liquidity, transaction costs, execution conditions, regulatory environments and other relevant factors can change materially over time. Relationships that have been observed historically may weaken, disappear or behave differently in future periods.

The views and analytical conclusions expressed in this publication reflect the research framework and information considered as of the publication date and may be revised as new evidence becomes available. Finovis Software Solutions LLC-FZ undertakes no obligation to update this publication or any information contained in it.

To the fullest extent permitted by applicable law, Finovis Software Solutions LLC-FZ shall not be liable for any loss or damage arising from or in connection with the use of, or reliance on, this publication or any information contained herein.

This publication may not be interpreted as creating any advisory, fiduciary, contractual or other professional relationship between Finovis Software Solutions LLC-FZ and the reader.

Finovis

AI FOR THE FINANCIAL INDUSTRY

FINOVIS RESEARCH

DATA & SIGNALS · RP-002

From Noise to Signal

Identifying Information in Financial Market Data

June 2026

finovis.ai