Algorithmic Trading

Arthur Li

February 1, 2025

Contents

1 Preface and Prerequisites

1.1 Brief Overview

Lorem Ipsum
To be completed once most of the book is done

1.2 Reading Roadmap

Lorem Ipsum. To be completed once most of the book is done

This content builds upon the foundational works of Rishi K. Narang (2013), Raja Velu (2020), and Marcos Lopez Prado (2018), among others, whose insights form the backbone of our discussion.

1.3 Overview of Systematic Investments

The figure below illustrates a live, production-level trading strategy. (Note that the diagram does not include ancillary components—such as research tools—that are also essential for building a complete strategy.)

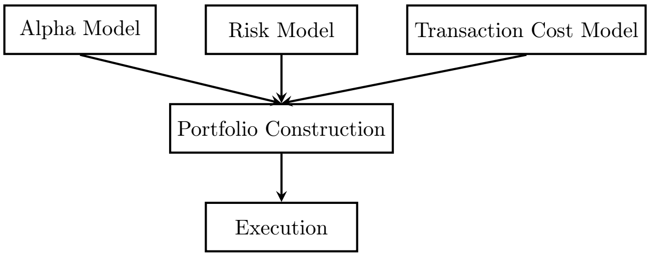

Figure 1: Live Production Trading Strategy Overview

At its core, the trading system is organized into three primary modules:

i. **Alpha Model:** Predicts the future behavior of the instruments under consideration, thereby generating directional alpha.

ii. **Risk Model:** Limits exposure by controlling risk factors that might not generate returns but could lead to losses (i.e., by setting directional exposure limits on asset classes).

iii. **Transaction Cost Model:** Assesses whether the trading costs required to transition from the current portfolio to a new one are acceptable within the portfolio construction framework.

These models feed into a portfolio construction model that balances profitability and risk to determine the optimal portfolio configuration. This model identifies the necessary trade adjustments, which are then executed by the system.

The execution model receives the required trades and, based on inputs such as trade urgency and current market liquidity, carries out the transactions in an efficient and cost-effective manner.

Method 1.3.1. *Chains of Production for Alpha Signals*
The production process for generating alpha signals consists of the following stages:

i. **Data Curation:** Involves collecting, cleaning, indexing, storing, adjusting, and delivering data to the production pipeline. This stage requires expertise in market microstructure and data protocols (e.g., FIX).

ii. **Feature Analysis:** Converts raw data into meaningful signals through techniques in information theory, signal extraction, visualization, labeling, weighting, classification, and feature importance assessment. Feature analysts compile and catalogue these insights.

iii. **Strategy Development:** Transforms informative features into actionable investment algorithms. Strategists mine feature libraries for ideas, blending deep financial market knowledge with data science to develop transparent (white-box) strategies—even if some features originate from black-box methods.

iv. **Back-Testing:** Evaluates the investment strategy under various scenarios. This process demands rigorous empirical analysis and includes metadata on how the strategy was formulated.

v. **Deployment:** Integrates the strategy code into the production system. This step is managed by algorithm specialists and mathematical programmers to ensure fidelity to the prototype while minimizing latency.

vi. **Portfolio Oversight:** Monitors the strategy post-deployment throughout its lifecycle:

   1. **Embargo:** Initially, the strategy is run on post-backtest data. If performance is consistent with backtesting, the strategy advances to the next phase.
   2. **Paper Trading:** The strategy is executed on a live feed, simulating real trading conditions. This phase accounts for data parsing, calculation delays, execution lags, and other operational latencies.
   3. **Graduation:** The strategy is allocated real capital, either individually or as part of an ensemble, with detailed evaluations of risk, returns, and costs.
   4. **Re-allocation:** Allocations are dynamically adjusted based on live performance. Initial allocations are small and may be increased as the strategy proves itself, with a subsequent reduction if performance declines.
   5. **Decommission:** If the strategy underperforms over an extended period, it is phased out.

1.3.1 Alpha Models Overview

Alpha models are designed to test and exploit theories about market behavior. These fall into two categories:

- **Theory-Driven Models**, which derive signals from economic or behavioral principles, and
- **Data-Driven Models**, which rely on complex data mining and statistical techniques.

In practice, practitioners often blend several models together to capture multiple facets of market dynamics.

Definition 1.3.2. *Theory-Driven Models*
Theory-driven models use fundamental principles to explain and predict market behavior. These include:

i. **Trend Following:** Assumes that once a market trend is established, it will persist long enough to be identified and exploited. As supporting data accumulate for a bullish or bearish outlook, more participants join the trend, shifting the asset price to a new equilibrium. Note that strategies such as moving average crossovers often yield less than a one-to-one return relative to the downside risk because market behavior can be unstable and episodic.

ii. **Mean Reversion:** Based on the idea that asset prices will revert to their historical average after deviating. Short-term liquidity imbalances can force prices to move abruptly, but these are typically followed by corrective movements. Statistical arbitrage, which bets on the convergence of prices between similar stocks that have diverged, is a common application of this strategy.

iii. **Value/Yield:** Evaluates securities by comparing fundamental ratios to their market price (often using an inverted ratio for consistency). A higher yield indicates a relatively cheaper instrument. This approach underpins *carry trades*, where investors buy undervalued assets and sell overvalued ones. In *Quant Long Short (QLS)* strategies, stocks are ranked on factors like value, and long positions are taken in the most attractive stocks while shorting the least attractive.

iv. **Growth:** Focuses on the future or historical growth potential of an asset. Forward-looking growth expectations are central, with the belief that companies with strong growth will increasingly dominate their sectors. Macro-level growth factors can drive foreign exchange decisions, while micro-level factors are key for individual companies.

v. **Quality:** Emphasizes capital preservation by favoring high-quality assets—those with robust earnings, sound balance sheets, and low leverage—over lower quality ones.

While theory-driven models incorporate clear economic rationale, **data-driven models** use advanced statistical and machine learning methods to extract patterns directly from data. These techniques, though more mathematically complex and often applied in high-frequency environments, can identify subtle market signals without relying on traditional economic theory.

Method 1.3.3. *Strategy Parameters*
Implementing an alpha strategy requires careful specification of several parameters:

i. **Forecast Target:** Define what the model predicts – whether it is the direction, magnitude, duration of a move, or even the probability of a particular outcome. A stronger signal (in terms of higher expected return or likelihood) typically warrants a larger position.

ii. **Time Horizon:** Forecast horizons can range from microseconds to years. Short-term strategies usually involve a high frequency of trades, whereas long-term strategies trade less frequently.

iii. **Bet Structure:** Models may generate absolute forecasts or relative forecasts (comparing one instrument to another). For relative forecasts, grouping assets into pairs or clusters (e.g., sectors) is common. While pair trading allows for precise comparisons, larger groupings can isolate idiosyncratic movements from overall market trends. Grouping can be achieved through statistical methods or heuristic industry classifications, each with its own tradeoffs.

iv. **Investment Universe:** Selection is based on factors such as geography, asset class, and instrument type. High liquidity is essential for reliable transaction cost estimation and consistent behavior. Consequently, common stocks, futures (on bonds and equity indices), and forex are frequently modeled, while more volatile instruments (e.g., certain biotech stocks) are typically excluded.

v. **Model Specification:** This defines the mathematical structure of the strategy, including any machine learning or data mining techniques used. Regular model refitting is necessary to adapt to changing market conditions, although it may increase the risk of overfitting.

vi. **Run Frequency:** This parameter sets how often the model is executed, ranging from monthly to real time. A higher run frequency increases transaction costs and sensitivity to noise, whereas a lower frequency can lead to larger trades that might themselves influence market prices.

Method 1.3.4. *Blending of Models*
Integrating multiple alpha signals can be accomplished using different approaches:

i. **Linear Models:** Assume that factors are independent and additive. Multiple regression techniques are typically employed to determine the weight of each factor.

ii. **Nonlinear Models:** Used when factor relationships are interdependent or evolve over time. Conditional models adjust the weight of one factor based on the performance of another, while rotational models dynamically shift weights according to recent performance metrics.

iii. **Machine Learning Models:** Although developing machine learning strategies requires substantial effort in data curation, infrastructure, feature engineering, and backtesting, these models can capture micro-level signals that traditional econometric methods may miss. Despite the complexity, ML approaches are becoming more prevalent as computational resources improve.

1.3.2 Risk Models

Risk models are indispensable tools in algorithmic trading, providing a quantitative framework to assess, monitor, and manage the inherent risks of financial markets. They not only serve to measure exposure but also guide portfolio construction, hedging strategies, and overall risk control.

Method 1.3.5. *Factor-Based Models*
Factor-based models decompose asset returns into contributions from systematic factors and idiosyncratic components. The most common factors include:

i. **Market Factor:** Captures the overall movement of the market.

ii. **Size Factor:** Reflects the differential risk associated with companies of varying market capitalizations.

iii. **Value Factor:** Accounts for risk due to discrepancies between market prices and fundamental valuations.

iv. **Momentum Factor:** Measures the tendency of asset prices to continue in their current trajectory.

This allows traders to understand which elements drive portfolio risk and adjust exposures accordingly.

Method 1.3.6. *Statistical Models*
Statistical risk models leverage historical data and probabilistic techniques to quantify risk parameters.

i. **Historical Simulation:** Directly computing risk metrics from past return distributions.

ii. **Monte Carlo Simulation:** Generating a large number of potential future return scenarios to estimate risk under diverse conditions.

iii. **Parametric Methods:** Employing analytical formulas based on assumed return distributions to calculate key risk measures.

These are useful for dynamically updating risk assessments as new market data become available.

Method 1.3.7. *Limiting Size of Risk*
Quantitative risk models are designed to limit the size of exposures to enhance return consistency.

i. **Constraint Mechanisms:**

   a. **Hard Constraints:** Absolute limits imposed on position sizes, which may be set arbitrarily.

   b. **Penalty Functions:** Flexible constraints where positions can exceed the limit if the alpha model forecasts significantly higher returns, with penalties applied for surpassing the prescribed levels.

ii. **Risk Measurement Approaches:**

   a. **Longitudinal Analysis:** Evaluates risk by assessing the volatility of an instrument over time.

   b. **Dispersion Analysis:** Measures risk by analyzing the correlation or covariance between assets.

These methods can be applied to single positions, groups of positions (such as sectors or asset classes), different types of risk exposures, and overall portfolio leverage.

Method 1.3.8. *Limiting the Types of Risk*
To eliminate unintentional exposure as there is no expectation of being compensated sufficiently for accepting them. This can be achieved the following measures:

i. **Market Exposure Restriction:** Focus on specific market segments to avoid undue exposure to volatile or unpredictable markets.

ii. **Leverage Management:** Limit the use of leverage by enforcing conservative leverage ratios to mitigate amplified losses.

iii. **Stop-Loss Policies:** Implement strict stop-loss rules to automatically exit positions that breach predetermined loss thresholds.

iv. **Position Size Controls:** Cap individual position sizes relative to overall portfolio risk, ensuring no single trade unduly influences portfolio performance.

v. **Regular Risk Assessments:** Continuously monitor risk metrics and adjust risk parameters to reflect evolving market conditions.

1.3.3 Transaction Cost Models

A trade is executed only if it improves the probability or magnitude of a return (as predicted by alpha model) or reduces likelihood or extent of a loss (as determined by risk model). However, this improvement must be greater than cost of trading. Note that transaction cost model is intended not to minimize trading costs directly, but rather to inform the portfolio construction engine of the costs associated with executing a given trade.

Remark 1.3.9. *Transaction Cost Components*

i. **Commissions and Fees:** These are payments made to brokerages (for accessing other market participants), exchanges (for enhanced transaction security), and regulators (for maintaining operational infrastructure). In the context of quantitative trading, where bank infrastructure is utilized, brokerage commissions tend to be minimal on a per-trade basis.

   Brokers also charge clearing and settlement fees. *Clearing* involves activities such as regulatory reporting and monitoring, tax handling, and failure management, all of which occur before settlement. *Settlement* is the final delivery of securities in exchange for full payment.

ii. **Slippage:** This refers to the change in price between the moment the trading system decides to execute a transaction and the time the order is actually sent to the exchange. Trend-following strategies tend to experience more slippage because the assets are already moving in the desired direction, whereas mean-reverting strategies are less affected. Reduced latency to market minimizes slippage, while higher asset volatility increases it.

iii. **Market Impact:** This measures the extent to which an order influences the market through its demand for liquidity. The market impact remains uncertain until after the trade is executed. Additionally, there can be an interaction between slippage and market impact (i.e., selling during an upward trending market).

Definition 1.3.10. *Types of Transaction Cost Models*

i. **Flat Model:** Assumes that the cost of trading remains constant regardless of the order size. This model is appropriate when the traded size is nearly uniform and liquidity remains stable.

ii. **Linear Model:** In this model, the trading cost increases at a constant rate relative to the order size. It provides a better estimate than the flat model.

iii. **Piece-Wise Linear Model:** This approach employs piece-wise linear functions to model costs. It strikes a balance between simplicity and accuracy, offering improved precision compared to flat or linear models.

iv. **Quadratic Model:** Although the most computationally intensive, the quadratic model delivers the highest accuracy in cost estimation.

1.3.4 Portfolio Construction Models

Portfolio construction models generally come in two major forms: rule-based models and optimisers. Rule-based models are built on heuristics—they can be very simple or quite complex—and are often derived from human experience and trial-and-error. In contrast, optimiser-based models rely on an objective function and use algorithmic methods to achieve the desired portfolio outcome.

Definition 1.3.11. *Rule-Based Models*

i. **Equal Position Weighting:** Applied when portfolio manager believes that once a position is deemed good enough to own, no further information is needed to decide its size. Strength of signal does not influence the weighting. Model assumes there is enough statistical power to predict not only direction but also magnitude relative to other forecasts within the portfolio. As a result, portfolio may place a few large bets on the strongest forecast and many smaller bets on less dramatic signals; however, this can lead to taking excessive risk in an idiosyncratic event on an apparently attractive position, which may cause adverse selection bias.

ii. **Equal Risk Weighting:** Strategy adjusts position sizes inversely to their volatility or another measure of risk. More volatile positions has smaller allocations, whereas less volatile positions has more allocation. Because the unit of risk is typically a backward-looking measure, such as historical volatility, if volatility shifts over time, the model might be misled.

iii. **Alpha-Driven Weighting:** Position size is primarily determined by alpha model. Alpha signal guides size of position, usually subject to predetermined size limits. Additional constraints often include limits on the total bet size within a group. A function may also be used to relate the forecast's magnitude to the position size. In futures trend following, it might suffer from sharp drawdowns, as it heavily relies on the accuracy of the alpha signals.

iv. **Decision-Tree Weighting:** A decision-making process is used to determine the allocation for each instrument based on both the type of alpha model and the type of instrument. Constraints, such as percentage limits for allocation, may also be imposed. However, as more alpha models or instrument types are introduced, the decision tree can grow significantly in complexity.

Remark 1.3.12. *Optimisers Models Parameters*
The pioneering model in optimiser-based portfolio construction is Harry Markowitz's mean variance optimisation (MVO), which is founded on the principles of modern portfolio theory (MPT). The main inputs to these models include the expected return (mean), asset variance, and the expected correlation matrix. Other inputs typically involve the portfolio's size in currency terms, the desired risk level (such as volatility or expected drawdown), and additional constraints like liquidity and universe limits. Model uses an objective function paired with an algorithm that seeks to maximise the portfolio's return relative to its volatility.

i. **Expected Return:** Derived from alpha models, this input captures not only the direction but also the magnitude of the expected returns.

ii. **Expected Volatility:** Typically estimated using stochastic volatility forecasting methods (i.e., GARCH), this input accounts for periods of high and low volatility along with occasional jumps.

iii. **Expected Correlation:** Given that instrument correlations can fluctuate over time, it is often more effective to group similar assets together before calculating the correlations within each group.

Method 1.3.13. *Optimisation Techniques*

i. **Unconstrained Optimisation:** Most basic form of optimisation with no constraints applied. Might result in a portfolio that invests all available capital in a single instrument—specifically, the one with the highest risk-adjusted return.

ii. **Constrained Optimisation:** Constraints such as position limits or limits on groups of instruments are applied. These constraints can sometimes have a greater influence on the portfolio construction than the optimiser itself.

iii. **Black-Litterman Optimisation:** This combines investor forecasts with a measure of confidence in those forecasts, blending them with historical data. It adjusts the historically observed correlation levels by incorporating the investor's return expectations for various instruments.

iv. **Grinold and Kahn's Approach:** Instead of directly sizing positions, this constructs a portfolio of signals. It creates factor portfolios, each typically based on a single type of alpha forecast. After backtesting these factor portfolios, their return series are then treated as instruments for the optimiser. Since the number of factor portfolios is usually limited (typically no more than 20), the optimisation problem becomes more manageable. This method also allows for the inclusion of a risk model, transaction cost model, overall portfolio size, and risk targets as additional inputs.

v. **Resampled Efficiency:** Seeks to improve the input parameters for optimisation by reducing the oversensitivity to estimation errors. It does so by employing Monte Carlo simulation to resample data, thereby mitigating the estimation error in the inputs.

vi. **Data-Mining Approaches:** Leverage machine learning techniques—such as supervised learning or genetic algorithms—to explore a wide range of potential portfolios and identify the optimal one.

1.3.5 Execution Model

There are two primary methods to execute a trade: electronically or via a human intermediary. In electronic execution, direct market access (DMA) is employed, allowing traders to leverage the brokerage firm's infrastructure and exchange connectivity to trade directly on electronic markets.
Execution algorithms can be developed in-house, sourced from brokers, or obtained from third-party vendors. Brokerages also provide portfolio bidding services. In these arrangements, the "blind" portfolio for the transaction is characterized by features such as valuation ratios of long and short positions, sector allocation, market

capitalization, and similar metrics. The broker then quotes a fee—expressed in basis points relative to the gross market value of the portfolio traded—which offers the trader a measure of certainty. Once an agreement is reached, the broker collects the fee and assumes the risk of executing the portfolio at future market prices, which may turn out to be more or less favorable than the initially guaranteed prices..

Remark 1.3.14. *Order Execution Algorithm Parameters*

i. **Aggressive vs Passive:** Algorithm decides whether to use an aggressive or passive order based on how immediately the trade must be executed. Market orders are inherently aggressive. Limit order placed at current best quote is relatively aggressive; limit order positioned below current bid is considered passive. Many exchanges reward liquidity providers for placing passive orders, while charging traders for consuming liquidity. When an order crosses the spread, it effectively uses liquidity from another trader's passive order, thereby reducing available liquidity. In return, if the passive order is executed, the trader can benefit from a better transaction price along with a commission rebate from the exchange.
   Momentum-based strategies typically favour aggressive orders, whereas mean reversion strategies lean towards passive orders. A strong, reliable signal justifies a more aggressive execution, while a weaker or less certain signal may call for a more passive approach. An intermediate strategy might involve placing limit orders between the best current bid and offer.

ii. **Large vs Small Order:** Large order may be divided into several smaller orders to reduce market impact, though this carries risk of adverse price movements. The optimal size of each order segment is determined by estimates from a transaction cost model and an analysis of the appropriate level of aggressiveness.

iii. **Hidden vs Visible Order:** Visible order discloses certain trading intentions, whereas a hidden order conceals this information, helping to prevent market imbalances; however, hidden orders typically suffer from lower execution priority. In algorithmic trading, the practice of using hidden orders—where a large order is segmented into many smaller parts, with most being placed as hidden—is known as "iceberging."

iv. **Order Routing:** When multiple liquidity pools exist for same instrument, smart order routing systems are used to determine most suitable pool for executing a given order. These systems evaluate factors such as depth of liquidity on various electronic communication networks (ECNs) and connectivity speeds.

v. **Cancelling and Replacing Orders:** Traders may submit a large number of orders without the expectation of execution, only to rapidly cancel and replace them. This practice helps gauge the market's response to changes in order book depth, offering insights into potential profitable patterns. For example, if a trader intends to buy a significant number of shares, they might place numerous small orders at prices further from the market and then cancel them, thereby improving market perception.

Definition 1.3.15. *High Frequency Trading*
High Frequency Trading (HFT) involves alpha-driven strategies that focus on extremely short-term bets—often executed in seconds or less—referred to as *microstructure alphas*. These strategies analyze liquidity patterns in the order book. Larger quantitative groups may also incorporate these insights into their execution models to reduce the costs associated with entering trades. Even marginal improvements per trade can accumulate significantly over time. However, pursuing microstructure alpha as a standalone high frequency strategy necessitates substantial investments in both infrastructure and research.
Machine learning techniques may be applied to detect patterns in how other market participants execute orders. When competitors' execution models are less refined, their patterns become more apparent, allowing an ML strategy to exploit these patterns in the future. Short-term patterns tend to exhibit a degree of stability.

Definition 1.3.16. *HFT Shark Strategy*
This is designed to detect large orders that have been fragmented (or "iceberged") by observing the rapid filling of a series of very small trades. Quick execution of these small orders can indicate the presence of a large hidden order. The strategy then involves front-running this iceberg order by placing visible trades ahead of it. Consequently, the iceberg order is forced to move market prices upward in order to execute its trades. Once the iceberg order is fully executed, the resulting favorable price movement enables the shark to exit its position quickly, thereby securing a relatively risk-free profit.

Remark 1.3.17. *HFT Trading Infrastructure*
By using a broker that acts as a trading agent, traders can offload infrastructure requirements and avoid dealing directly with regulatory and other operational constraints. High frequency strategies may also employ colocation or sponsored access. In a colocation setup, traders place their trading servers in close physical proximity to the exchange to minimize latency.
The Financial Information eXchange (FIX) protocol is the standard for real-time electronic communication among market participants. Although open source FIX engines are available, high frequency traders often develop their own proprietary FIX engines to ensure optimal execution speeds.

1.3.6 Research

Definition 1.3.18. *Scientific Method*

1. Researcher observe a phenomenon in the market and construct a theory.
2. Researcher seeks out information to test the theory.
3. Researcher tests the theory, and with enough confidence, risk some capital on the validity of the theory.

Remark 1.3.19. *Sources of Alpha Idea Generation*

1. Observing the market, using the scientific method to test the theory
2. Academic literature, requiring significant time to read academic journals, working papers, and conference presentations for ideas. Literature from other fields such as astronomy, physics, or psychology, may provide ideas relevant to quant finance problems.
3. Migration of a researcher or portfolio manager from one quant shop to another.
4. Lessons from activities of discretionary traders

Remark 1.3.20. *Model Quality Assessment*

i. **Cumulative Profit Graph:** A smooth profit curve is ideal; if the graph shows long periods of inactivity or exhibits sharp, erratic losses and gains, it may signal issues with the model.

ii. **Average Annual Rate of Return:** Indicates the historical performance level of the strategy.

iii. **Variability of Returns:** Lower variability in returns is preferable, as it suggests consistency. Examining the "lumpiness"—the share of total returns derived from periods significantly above average—can further measure return consistency.

iv. **Peak-to-Valley Drawdowns:** Measures maximum decline from any cumulative peak in the profit curve. Lower drawdowns, along with shorter recovery periods after drawdowns, reflect a more robust strategy.

v. **Predictive Power:** The R-squared statistic can be employed to assess how much of the variability in the predicted asset is explained by the model. For example, an exceptionally high R^2 (around 0.05 out of sample) may warrant further scrutiny. Additionally, bucketing instrument returns by deciles can help verify whether the model categorizes them accurately.

vi. **Percentage of Winning Trades and Winning Time Periods:** Determines whether the strategy relies on a small number of highly profitable trades or on a larger volume of moderately successful trades.

vii. **Risk-Adjusted Return Ratios:** Evaluate statistics such as the Sharpe ratio, information ratio, Sterling ratio, Calmar ratio, and Omega ratio to assess the balance between returns and risk.

viii. **Relationship with Other Strategies:** Consider the incremental value provided by a new strategy by comparing its performance with existing strategies, both independently and in combination.

ix. **Time Decay:** Examine how the returns of the strategy are affected if trades are initiated on a lagged basis after receiving a trading signal. This helps to determine the sensitivity of the strategy to the timeliness of information and the degree of market saturation.

x. **Sensitivity to Specific Parameters:** A high-quality strategy should show only minor variations in outcomes with small changes in its parameters; large fluctuations may indicate overfitting.

xi. **Overfitting:** By plotting parameter values against the corresponding outcomes, one should observe a relatively flat curve with no abrupt jumps. Models that are parsimonious—that is, those that rely on fewer parameters—tend to be less prone to overfitting.

Remark 1.3.21. *Other Considerations in Model Testing*
It is crucial to note that overestimating trading costs can lead to holding positions longer than optimal, whereas underestimating these costs might result in excessive portfolio turnover and a detrimental bleed from trading expenses. Moreover, assumptions regarding the availability of short positions must be carefully considered, especially with respect to hard-to-borrow lists.

1.3.7 Risk Assessment

Definition 1.3.22. *Model Risks*
Quantitative models inherently carry model risk—the risk that a model fails to accurately describe, match, or predict real-world phenomena. Each element of a quant model is subject to its own potential for error.

i. **Inapplicability of Modelling:** This risk arises when a quant model is applied to a problem for which it was not designed, or when a particular technique is misapplied to a given scenario.

ii. **Model Misspecification:** This occurs when the model does not properly reflect the real world. Although a model may perform adequately under normal conditions, it can fail under extreme circumstances.

iii. **Implementation Errors:** These include mistakes in coding or system design. Additionally, errors can occur if the model components are executed in an incorrect sequence.

Definition 1.3.23. *Regime Change Risk*
Quant models are built on historical data, relying on relationships that have prevailed over time. When a regime change occurs, these historical relationships and behaviors can shift, leading the model to lose its effectiveness.

Definition 1.3.24. *Exogenous Shock Risk*
This type of risk is driven by external events that are not inherent to the market itself, such as terrorist attacks, the outbreak of wars, bank bailouts, or regulatory changes (e.g., modifications to shorting rules). In such cases, discretionary overrides may be necessary.

Definition 1.3.25. *Contagion Risk*
Contagion risk occurs when multiple investors follow similar strategies. There are two components to this risk: first, the extent to which a quant strategy is crowded; and second, the additional positions held by other investors that might force them to exit the strategy in a panic (often referred to as the ATM effect).
Quant liquidation crisis may be triggered by factors such as the sheer size and popularity of quantitative strategies, suboptimal returns by operators leading up to a crisis, the cross-collateralisation of many strategies within funds, and risk targeting—where risk managers aim to maintain a specific level of volatility across their funds or strategies.

Method 1.3.26. *Risk Monitoring Methods*

i. **Exposure Monitoring Tools:** These tools aggregate current positions by various exposures (e.g., valuation, momentum, volatility) to monitor both gross and net exposures across sectors, industries, market capitalisation buckets, and style factors.

ii. **Profit and Loss Monitors:** By comparing the current portfolio against the previous day's closing prices, these monitors utilize intraday performance charts and also examine the source of profits and the hit rate (i.e., the percentage of positions where the strategy is profitable).

iii. **Execution Monitors:** These displays track the status of orders—indicating which are being processed and which have been completed—along with details such as transaction sizes and prices. They also measure fill rates for limit orders in passive strategies, and monitor slippage and market impact.

iv. **System Performance Monitors:** These monitors are responsible for detecting software or infrastructure errors, assessing CPU performance, measuring the speed of various stages of automated processes, and tracking communication latency.

1.4 Exploratory Data Analysis

1.4.1 Data Taxonomy

A brief overview of the types of data used in systematic trading.

Four essential types of financial data

Fundamental Data	Market Data	Analytics	Alternative Data
Assets	Price/Yield/IV	Analyst Recommendation	Satellite/CCTV
Liabilities	Volume	Credit Ratings	Google Searches
Sales	Dividend/Coupons	Earnings Expectations	Twitter/Chats
Costs/Earnings	Open Interest	News Sentiment	Metadata
Macro Variables	Quotes/Cancellations	. . .	. . .
. . .	Aggressor Side		
	. . .		

Remark 1.4.1. *Fundamental Data Characteristics*

i. Data is published with an index corresponding to last date in the report, which precedes the release date.

ii. Data is frequently backfilled or corrected, with the data vendor overwriting initial values as needed.

iii. The data is highly regularized and available at low frequency.

Remark 1.4.2. *Market Data Characteristics*

i. The raw feed consists of unstructured information, such as FIX messages (which allow full reconstruction of the trading book) or complete collections of BWIC (bids wanted in competition) responses.

ii. Processing FIX data is non-trivial, with approximately 10TB generated daily.

Remark 1.4.3. *Analytics Data Characteristics*

i. This is derivative data processed from the original source, with the relevant signal already extracted.

ii. It is costly to produce, and the methodology used in production may be biased or opaque.

Remark 1.4.4. *Alternative Data Characteristics*

i. This data is generated by individuals, business processes, and sensors.

ii. It provides primary information that is not available from traditional sources.

iii. Cost and privacy concerns; it may be particularly valuable if it challenges existing data infrastructure.

Definition 1.4.5. *Reference Data*

i. **Trading Universe:** Evolves daily to incorporate new listings and de-listings. Knowing when a stock ceases trading is crucial to avoid survivor bias.

ii. **Symbology Mapping:** Involves identifiers such as ISIN, SEDOL, RIC, and Bloomberg Tickers. Since symbols may change over time, mapping must persist as point-in-time data to support historical 'as-of-date' analyses, often requiring a bi-temporal data structure.

iii. **Ticker Changes:** Maintain a historical table of ticker changes (as described in symbology mapping) to ensure seamless continuity in time series data.

iv. **Corporate Actions Calendars:** Include events such as stock and cash dividends (with announcement and execution dates), stock splits, reverse splits, rights offers, mergers and acquisitions, spin-offs, adjustments in free float or shares outstanding, and quotation suspensions.
For dividends, announcements may coincide with increased volatility and price jumps, enabling strategies to capitalize on the added volatility.For splits and rights offers, historical data must be adjusted backward (both volume and price) to reflect these actions. For M&A and spin-offs, adjustments are needed to account for valuation changes, which are important in merger arbitrage strategies. Suspensions can create data gaps that impact backtesting.

v. **Static Data:** Comprises attributes like country, sector, primary exchange, currency, and quote factor. This data is used to group instruments based on fundamental similarities (e.g., for pairs trading), and maintaining a table of quotation currencies per instrument is essential for portfolio aggregation.

vi. **Exchange Specific Data:** Each exchange has unique features that must be considered.
First Group: Hours and Dates of Operations

1. **Holiday Calendar:** Different markets have their own holiday schedules. For strategies that trade across multiple markets, discrepancies in holiday closures can affect correlation.
2. **Exchange Session Hours:** Different sessions (Pre-Market, Continuous Core, After-Hours, etc.), auction times, cutoff times for order submission, lunch breaks, and pre-/post-lunch auctions. This also includes DST adjustments and variations in trading hours by venue.
3. **Disrupted Days:** Records of exchange outages or trading disruptions, which are important to filter out when building or testing strategies.

Second Group: Trading Mechanics

1. **Tick Size:** The minimum eligible price increment, which may vary by instrument and price level.
2. **Trade and Quote Lots:** The minimum size increments for trades or quotes.
3. **Limit-Up and Limit-Down Constraints:** Maximum daily price fluctuations and the rules for trading pauses or restrictions at threshold levels.
4. **Short Sell Restrictions:** Rules that may prevent short sales from trading at prices worse than the last trade, or from generating new quotes below the lowest prevailing price. These restrictions impact liquidity sourcing.

vii. **Market Data Condition Codes:** Vary by exchange and asset class. Each market event may have multiple codes (e.g., auction trade, lit/dark trade, cancelled/corrected trade, regular trade, off-exchange reporting, block trade, or multi-leg order such as an option spread). It is essential to build a mapping table for these codes so that trades published solely for reporting purposes can be excluded from liquidity updates in aggregated daily data.

viii. **Special Day Calendars:** Identify days with distinct liquidity characteristics that impact both execution strategies and alpha generation. Examples (non-exhaustive) include:

1. Half trading days before Christmas and after Thanksgiving (US).
2. Ramadan effects in Turkey.
3. Taiwan market opening on a weekend to compensate for holiday closures.
4. Adjusted trading hours in Korea on nationwide university entrance exam days.
5. Late openings in the Brazilian market following Carnival.
6. Last trading days of months and quarters, when portfolio rebalancing occurs.
7. Index rebalancing dates, where intraday volume skews toward the end of day.
8. Options and futures expiry dates (e.g., quarterly/monthly expiry, Triple Witching in the US, Special Quotations in Japan) that cause excess trading volume and altered intraday patterns due to hedging and portfolio adjustments.

Model normal days first; special days either modelled independently or by adjusting normal day baseline.

ix. **Futures-Specific Reference Data:** Essential for determining which contract was live at any point via an expiry calendar and identifying the most liquid contract. For example, equity index futures are generally most liquid for the front month, while energy futures (such as oil) might be more liquid in the second contract. Note that there is no standardised expiry frequency across markets. When computing rolling-window metrics, potential roll dates must be accounted for; volume data may be blended before and after a roll. Additionally, futures markets exhibit different intra-day phases with distinct liquidity characteristics, so market data metrics (volume profiles, average spreads, bid-ask sizes) should be computed for each session based on a schedule.

x. **Options-Specific Reference Data (Options Chain):** Consists of expiry date and strike price combinations. Mapping equity tickers to their corresponding option tickers, with strike and expiry information, facilitates more complex investment and hedging strategies (e.g., assessing distance to strike or changes in open interest between puts and calls).

xi. **Market-Moving News Releases:** Includes macroeconomic announcements (central bank statements such as FED/FOMC, ECB, BOE, BOJ, SNB; Non-Farm Payrolls; PMIs; Manufacturing Indices; Crude Oil Inventories) and stock-specific events like earnings releases. Maintaining a calendar of these events helps assess their impact on strategies.

xii. **Related Tickers:** Tickers that represent the same underlying asset, allowing for efficient opportunity identification. This includes mappings for primary tickers to composite tickers (in fragmented liquidity markets), dual-listed or fungible securities (e.g., in the US and Canada), ADRs/GDRs, or differences between local and foreign boards.

xiii. **Composite Assets:** Such as ETFs, indexes, and mutual funds, which are used to achieve desired exposures or as hedging instruments. They can also present arbitrage opportunities when deviating from their NAV. It is important to maintain data on their constituent time series, any cash component, the divisor for converting NAV to quoted price, and the constituent weights.

xiv. **Latency Tables:** For high-frequency trading, these tables record the distribution of latencies between different data centers for efficient order routing, as well as reordering data collected from different locations.

Definition 1.4.6. *Market Data Feed*

i. **Level I Data (Trade and BBO Quotes):** Contains trade executions and top-of-book quotes, sufficient to reconstruct the Best Bid and Offer (BBO). This data also includes trade status (e.g., cancelled, reported late) and qualifiers (e.g., odd lot, normal trade, auction trade, intermarket sweep, average price reporting, exchange details), which are useful for analyzing event sequences and deciding whether a print should update the last price and total traded volume.

ii. **Level II Data (Market Depth):** Provides full depth of the limit order book, including all updates (price changes, additions, or removals of shares) across all venues in fragmented markets.

iii. **Level III Data (Full Order View):** Provides detailed message data in which each incoming order is assigned a unique ID for tracking. Records precise details when an order is executed, cancelled, or amended, making it possible to reconstruct complete order book (with national depth) at any moment.

1. **Timestamp:** Milliseconds elapsed since midnight.
2. **Ticker:** Equity symbol (up to 8 characters).
3. **Order:** Unique order identifier.
4. **T (Message Type):** 'B' indicates an add buy order; 'S' an add sell order; 'E' a partial execution; 'C' a partial cancellation; 'F' a full execution; 'D' a full deletion; 'X' a bulk volume for a cross event; and 'T' an execution of a non-displayed order.
5. **Shares:** Order quantity for messages 'B', 'S', 'E', 'X', 'C', and 'T'. Zero for 'F' and 'D'.
6. **Price:** Order price for 'B', 'S', 'X', and 'T' messages; zero for cancellations and executions. The last 4 digits denote the decimal part (padded with zeroes), and the value should be divided by 1000 to convert to the currency unit.
7. **MPID:** A 4-character Market Participant ID associated with the transaction.
8. **MCID:** A 1-character Market Centre Code for the originating exchange.

iv. Special order types of note:

1. **Order Subject to Price Sliding:** Execution price may be one cent less favourable than display price (e.g., at NASDAQ). Such orders are ranked at locking price as hidden orders but displayed at one minimum price variation inferior; a new order ID is issued if order is replaced as a display order.
2. **Pegged Order:** Based on the NBBO, these orders are non-routable and receive a new timestamp upon repricing; display rules vary by exchange.
3. **Mid-point Peg Order:** A non-displayed order that may result in half-penny executions.
4. **Reserve Order:** The displayed size is treated as a displayed limit order, while the reserve size is subordinate to non-displayed and pegged orders. The minimum display quantity is 100 shares; when it falls below this threshold, the reserve is replenished, a new timestamp is generated, and the displayed size is re-ranked.
5. **Discretionary Order:** Displayed at one price while passively trading at a more aggressive discretionary price. It only becomes active when shares are available within the discretionary price range and is ranked last in priority. The execution price may be less favourable than the display price.
6. **Intermarket Sweep Order:** Can be executed without the need to verify the prevailing NBBO.

Using this comprehensive Level III data, one can model the inter-arrival times of various events, as well as the arrival and cancellation rates as functions of distance from the best bid/offer and other variables (e.g., order book imbalance, queue length). Subsequently, analysis may include assessing the impact of market orders on the limit order book, estimating the likelihood of a limit order advancing in the queue, determining the probability of capturing the spread, and forecasting short-term price movements.

Definition 1.4.7. *Binned Data*

i. **Open, High, Low, Close (OHLC) and Previous Close Price:** These values indicate trading activity and intraday volatility. The range between the low and high prices reflects market sentiment, and the previous close must be adjusted for corporate actions and dividends.

ii. **Last Trade before Close (Price/Size/Time):** Captures any jump in the close price during the final trading moments, serving as an indicator of the stability of the close as a reference for the next day.

iii. **Volume:** Acts as an indicator of trading activity, particularly when it deviates sharply from long-term averages. It is useful to collect volume breakdowns between lit and dark venues for execution strategies.

iv. **Auctions Volume and Price:** Reflects price discovery events marked by significant volume prints.

v. **VWAP:** The Volume Weighted Average Price offers an indication of daily trading activity and is particularly useful for algorithmically executing large orders.

vi. **Short Interest/Days-to-Cover/Utilisation:** These metrics serve as proxies for investor positioning. High short interest suggests a bearish view from institutional investors, while the utilisation of borrowable securities indicates the potential for additional shorting. Days-to-Cover helps assess the potential severity of a short squeeze; higher values imply a greater chance of sudden price surges in heavily shorted securities.

vii. **Futures Data:** Provides insights into market activity or positions of large investors through open interest data. Arbitrage opportunities may arise if the basis is mispriced relative to dividend estimates.

viii. **Index-Level Data:** Supplies relative measures for instrument-specific features (such as index OHLC and volatility). Normalised features can help identify instruments that deviate from their benchmarks.

ix. **Options Data:** Offers information on trader positioning via open interest and the Greeks.

x. **Asset Class Specific Data:** Includes yield or benchmark rates (such as repo rates, 2-year, 10-year, and 30-year yields), CDS spreads, and the US Dollar Index.

Definition 1.4.8. *Granular Intraday Microstructure Activity*

i. **Number and Frequency of Trades:** Serves as a proxy for market activity and continuity; a low number may indicate challenging execution and higher volatility.

ii. **Number and Frequency of Quote Updates:** Provides a similar measure of market activity.

iii. **Top of Book Size:** Liquidity; larger sizes allow for larger orders to be executed almost immediately.

iv. **Depth of Book (Price and Size):** Also reflects the liquidity available in the market.

v. **Spread Size (Average, Median, Time-Weighted Average):** Proxy for trading costs. Parameterised distribution of spreads helps in identifying when trading opportunities are relatively inexpensive or costly.

vi. **Trade Size (Average, Median):** Useful for identifying intraday liquidity opportunities by examining the volume available in the order book.

vii. **Ticking Time (Average, Median):** Represents how frequently the top level of the order book is updated. This measure is critical for execution algorithms that must adapt their update frequency (e.g., for adding or cancelling child orders) to the characteristics of the traded instrument.

Daily distributions of these metrics can be used as starting estimates at the beginning of the day and updated intraday via online Bayesian methods.
Derived daily data include:

i. X-day Average Daily Volume (ADV) / Average Auction Volume

ii. X-day Volatility (e.g., close-to-close, open-to-close)

iii. Beta with respect to an index or sector (using standard beta or asymmetric up-/down-day beta)

iv. Correlation Matrix

When binning data, intervals may range from a few seconds to 30 minutes. Minute-bar data is typically used for volume and spread profiles to reduce noise from market friction.

Definition 1.4.9. *Fundamental Data and Other Data*

i. **Key Ratios:** Such as Earnings Per Share (EPS), Price-to-Earnings (P/E), Price-to-Book (P/B), etc.

ii. **Analyst Recommendations:** Aggregated consensus valuations from analysts.

iii. **Earnings Data:** Quarterly earnings estimates provided by research analysts serve as indicators of a stock's performance prior to the actual published figures.

iv. **Holders:** Significant changes in ownership can indicate shifts in sentiment among sophisticated investors.

v. **Insider Purchase/Sale:** Reflects potential future price movements, as insiders typically possess the best available information about the company.

vi. **Credit Ratings:** Downgrades, which lead to higher funding costs, can adversely impact equity prices.

1.4.2 Financial Data Structures

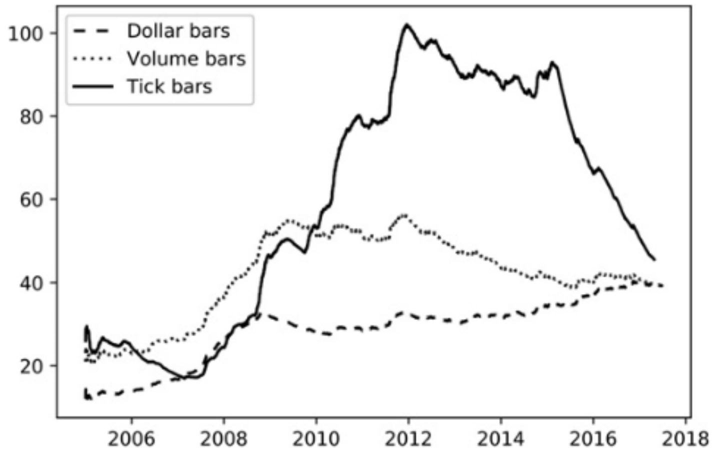

Figure 2: Average daily frequency of tick, volume, and dollar bars

Remark 1.4.10. *Standard BARS*
For converting a stream of irregularly timed observations into a homogeneous series through regular sampling.

i. **Time Bars:** Sample data at fixed time intervals. Typically record information such as the timestamp, volume-weighted average price (VWAP), open price, close price, high price, low price, and volume.
As markets do not operate in constant time intervals, time bars tend to oversample during periods of low activity and undersample during high-activity periods. Consequently, they often exhibit undesirable statistical properties such as serial correlation, heteroscedasticity, and non-normal return distributions.

ii. **Tick Bars:** Created by sampling each time a specified number of transactions occur, thereby aligning the sampling process with the arrival of market information.
This method often results in returns that are closer to IID normal (Thierry and Helyette (2000)). Nonetheless, care must be taken to address outliers since many exchanges conduct opening and closing auctions where the order book may accumulate unmatched bids and offers. Additionally, order fragmentation can introduce some arbitrariness in the tick count.

iii. **Volume Bars:** Generated by sampling whenever a predetermined number of security units are traded.
Volume bars generally yield better statistical properties than tick bars and serve as a useful tool for studying market microstructure theories.

iv. **Dollar Bars:** Formed by sampling each time a fixed monetary value is exchanged. This approach is particularly useful for analyses involving significant price fluctuations and is robust to corporate actions (e.g., splits, reverse splits, new share issuances, or share buybacks).
Moreover, the bar size can be adjusted dynamically based on factors such as a company's free-floating market capitalization or the total amount of issued debt.

Remark 1.4.11. *Information-Driven Bars*
Sample more frequently when new (micro-structural) information enters the market.

i. **Tick Imbalance Bars:** Formed whenever the cumulative tick imbalance exceeds its expected threshold. Consider a sequence of ticks $\{(p_t, v_t)\}_{t=1}^T$, where p_t and v_t denote the price and volume of tick t, respectively. A tick rule defines the sequence $\{b_t\}_{t=1}^T$ by

$$b_t = \begin{cases} b_{t-1}, & \text{if } \Delta p_t = 0, \\ \frac{|\Delta p_t|}{\Delta p_t}, & \text{if } \Delta p_t \neq 0. \end{cases}$$

The tick imbalance at time T is then given by

$$\theta_T = \sum_{t=1}^T b_t.$$

Its expected value at the beginning of a bar is computed as

$$E_0[\theta_T] = E_0[T]\big(P[b_t = 1] - P[b_t = -1]\big) = E_0[T]\big(2P[b_t = 1] - 1\big),$$

where $E_0[T]$ is the expected tick bar size, and $P[b_t = 1]$ and $P[b_t = -1]$ are the unconditional probabilities of a tick being classified as a buy or a sell, respectively. In practice, $E_0[T]$ and $(2P[b_t = 1]-1)$ are estimated using an exponentially weighted moving average from previous bars.
Define the Tick Imbalance Bar (TIB) as the smallest contiguous subset of ticks such that

$$T^* = \arg\min_T \{|\theta_T| \geq E_0[T] \,|\, 2P[b_t = 1] - 1\}.$$

When the observed imbalance θ_T exceeds the expected value, a lower T will satisfy the condition.

ii. **Volume/Dollar Imbalance Bars:** These bars are triggered when the imbalance in volume or dollar amount deviates from its expectation. First, define the imbalance at time T as

$$\theta_T = \sum_{t=1}^{T} b_t v_t,$$

where v_t represents either the number of units traded (for Volume Imbalance Bars, VIB) or the dollar amount exchanged (for Dollar Imbalance Bars, DIB). The expected imbalance is calculated as

$$\begin{aligned} E_0[\theta_T] &= E_0\left[\sum_{t:b_t=1}^{T} v_t\right] - E_0\left[\sum_{t:b_t=-1}^{T} v_t\right] \\ &= E_0[T]\Big(P[b_t = 1]E_0[v_t|b_t = 1] - P[b_t = -1]E_0[v_t|b_t = -1]\Big) \\ &= E_0[T]\Big(v^+ - v^-\Big) \end{aligned}$$

or equivalently,

$$E_0[\theta_T] = E_0[T]\Big(2v^+ - E_0[v_t]\Big).$$

In practice, $E_0[T]$ and $(2v^+ - E_0[v_t])$ are estimated as an exponentially weighted moving average from previous bars. Next, define the VIB or DIB as the smallest contiguous subset of ticks satisfying

$$T^* = \arg\min_T \{|\theta_T| \geq E_0[T] \,|\, 2v^+ - E_0[v_t]\}.$$

Thus, when the observed imbalance exceeds the expected value, the bar is triggered.

iii. **Tick Runs Bars:** These bars are formed when the sequence of consecutive buys or sells (i.e., runs) deviates from expectation. This is particularly relevant when large traders sweep the order book, use iceberg orders, or split parent orders into multiple child orders, leaving a trace in the sequence $\{b_t\}_{t=1}^{T}$.
Define the run length as

$$\theta_T = \max\left\{\sum_{t:b_t=1}^{T} b_t - \sum_{t:b_t=-1}^{T} b_t\right\}.$$

The expected run length at the start of the bar is given by

$$E_0[\theta_T] = E_0[T]\max\{P[b_t = 1], 1 - P[b_t = 1]\}.$$

In practice, $E_0[T]$ and $P[b_t = 1]$ are estimated via an exponentially weighted moving average from prior bars. The Tick Runs Bar (TRB) is then defined as the smallest contiguous subset of ticks such that

$$T^* = \arg\min_T \Big\{\theta_T \geq E_0[T]\max\{P[b_t = 1], 1 - P[b_t = 1]\}\Big\}.$$

This definition implies that when θ_T exceeds the expected run length, a bar is formed. (Note: alternatively, one may count the number of ticks on each side without netting them off.)

iv. **Volume/Dollar Runs Bars:** These bars are triggered when the cumulative volume or dollar amount traded by one side exceeds its expected value.
First, define the run imbalance as

$$\theta_T = \max\left\{\sum_{t:b_t=1}^{T} b_t v_t - \sum_{t:b_t=-1}^{T} b_t v_t\right\},$$

where v_t is traded volume (for Volume Runs Bars, VRB) or dollar amount (for Dollar Runs Bars, DRB).

The expected imbalance is then computed as

$$E_0[\theta_T] = E_0[T] \max\Big\{P[b_t = 1]E_0[v_t|b_t = 1],\, (1 - P[b_t = 1])E_0[v_t|b_t = -1]\Big\}.$$

In practice, $E_0[T]$, $P[b_t = 1]$, $E_0[v_t|b_t = 1]$, and $E_0[v_t|b_t = -1]$ are estimated using exponentially weighted moving averages from prior bars. Define VRB or DRB as the smallest contiguous subset of ticks satisfying

$$T^* = \arg\min_T \Big\{\theta_T \geq E_0[T] \max\Big\{P[b_t = 1]E_0[v_t|b_t = 1],\ (1 - P[b_t = 1])E_0[v_t|b_t = -1]\Big\}\Big\}.$$

Thus, when the observed run imbalance exceeds the expected value, a bar is triggered.

Definition 1.4.12. *Multi-Product Series: ETF Trick*
A technique to model a basket of securities as if it were a single cash product, effectively transforming a complex multi-product dataset into a unified series that mimics a total return ETF.

Method 1.4.13. *ETF Trick*
The following steps produce a time series that reflects the value of a \$1 investment, with changes in the series representing profit and loss (PnL). The series remains strictly positive and incorporates implementation shortfall. The bars include:

i. The raw open price for each instrument $i = 1, \ldots, I$ at bar t, denoted by $o_{i,t}$.
ii. The raw close price for each instrument $i = 1, \ldots, I$ at bar t, denoted by $p_{i,t}$.
iii. The USD value of one point of instrument $i = 1, \ldots, I$ at bar t, denoted by $\varphi_{i,t}$ (includes forex rate).
iv. The trading volume for instrument $i = 1, \ldots, I$ at bar t, denoted by $v_{i,t}$.
v. Any carry, dividend, or coupon paid by instrument i at bar t, denoted by $d_{i,t}$ (this variable can also account for margin or funding costs).

All instruments $i = 1, \ldots, I$ must be tradable at each bar $t = 1, \ldots, T$. Even if some instruments are not continuously tradable throughout the interval $[t-1, t]$, they should be tradable at both times $t-1$ and t.
For a basket of securities with an allocation vector ω_t that is rebalanced (or rolled) on a set of bars $B \subseteq \{1, \ldots, T\}$, the \$1 investment value series $\{K_t\}$ is derived as follows:

$$h_{i,t} = \begin{cases} \dfrac{\omega_{i,t} K_t}{o_{i,t+1}\varphi_{i,t} \sum_{i=1}^{I} |\omega_{i,t}|}, & \text{if } t \in B, \\ h_{i,t-1}, & \text{otherwise,} \end{cases}$$

$$\delta_{i,t} = \begin{cases} p_{i,t} - o_{i,t}, & \text{if } (t-1) \in B, \\ \Delta p_{i,t}, & \text{otherwise,} \end{cases}$$

$$K_t = K_{t-1} + \sum_{i=1}^{I} h_{i,t-1}\varphi_{i,t}(\delta_{i,t} + d_{i,t})$$

with the initial asset under management (AUM) set as $K_0 = 1$. Here, $h_{i,t}$ denotes the holdings of instrument i at time t, and $\delta_{i,t}$ represents the change in market value for instrument i between $t-1$ and t. Profits or losses are reinvested on rebalancing days (i.e., when $t \in B$), thereby avoiding negative prices. Dividends $d_{i,t}$ are already incorporated within K_t.
The normalization factor $\omega_{i,t}\left(\sum_{i=1}^{I}|\omega_{i,t}|\right)^{-1}$ is used to de-lever the allocations.
Let τ_i be transaction cost associated with trading \$1 of the instrument. Three additional variables that the strategy needs to know for every observed bar t are:

i. Rebalance Costs: variable cost $\{c_t\}$ associated with allocation rebalance is

$$c_t = \sum_{i=1}^{I} (|h_{i,t-1}|\, p_{i,t} + |h_{i,t}|\, o_{i,t+1})\varphi_{i,t}\tau_i \quad \forall t \in B$$

Note c_t is not embedded in K_t, as shorting will generate fictitious proceeds when allocation is rebalanced. In code, $\{c_t\}$ is treated as a (negative) dividend.

ii. Bid-Ask Spread: the cost $\{\tilde{c}_t\}$ of buying or selling one unit of this ETF,

$$\tilde{c}_t = \sum_{i=1}^{I} |h_{i,t-1}|\, p_{i,t}\varphi_{i,t}\tau_i$$

When a unit is bought or sold, strategy must charge this cost $\tilde{c}_t$.

iii. Volume: volume traded $\{v_t\}$ is determined by least active member in the basket. Let $v_{i,t}$ be volume traded by instrument i over bar t. The number of tradable basket units is

$$v_t = \min_i \left\{ \frac{v_{i,t}}{|h_{i,t-1}|} \right\}$$

Transaction costs functions may not be linear, and can be simulated by the strategy.

Method 1.4.14. *ETF Trick: Computation of Allocation Vector with PCA*
Consider an IID multivariate Gaussian process characterized by a mean vector μ (of size $N \times 1$) and a covariance matrix V (of size $N \times N$). First, perform the spectral decomposition

$$VW = W\Lambda,$$

with the columns of W rearranged so that the diagonal entries of Λ are in descending order. Given an allocation vector ω, the portfolio risk is

$$\sigma^2 = \omega' V \omega = \omega' W \Lambda W' \omega = \beta' \Lambda \beta = (\Lambda^{1/2}\beta)'(\Lambda^{1/2}\beta)',$$

where β represents the projection of ω onto the orthogonal basis defined by W. Since Λ is diagonal, we have

$$\sigma^2 = \sum_{n=1}^{N} \beta_n^2 \Lambda_{n,n}.$$

The risk contribution from the nth component is then

$$R_n = \frac{\beta_n^2 \Lambda_{n,n}}{\sigma^2} = \frac{[W'n]_n^2 \Lambda_{n,n}}{\sigma^2},$$

with the property that $\sum_{n=1}^{N} R_n = 1$, where 1_N is an N-dimensional vector of ones.
To compute an allocation vector ω that satisfies a user-defined risk distribution R, note that

$$\beta = \left\{ \sigma \sqrt{\frac{R_n}{\Lambda_{n,n}}} \right\}_{n=1}^{N},$$

which represents the allocations in the new orthogonal basis. Allocation in original basis is then recovered by

$$\omega = W\beta.$$

Rescaling ω proportionally adjusts σ, thus preserving the risk distribution.

Method 1.4.15. *ETF Trick: Single Futures Roll*
To generate a non-negative, continuously rolled series for a \$1 investment, proceed as follows:

i. Compute the time series of rolled futures prices.

ii. Calculate the return r as the percentage change of the rolled price relative to the previous roll price.

iii. Construct the price series using these returns.

The methods described above enable the production of a continuous, homogeneous, and structured dataset from a collection of unstructured financial data. It is important to note, however, that many machine learning algorithms do not scale well with extremely large sample sizes; their accuracy typically improves when they are trained on the most relevant examples.

Method 1.4.16. *Sampling for Reduction*
To reduce volume of data used for fitting a machine learning algorithm, downsampling techniques can be applied:

i. **Sequential Sampling:** Sample the data at a constant step size (i.e., using linspace sampling).

ii. **Random Sampling:** Select data points randomly according to a uniform distribution.

Note that neither sampling method guarantees that the most relevant observations are retained.

Method 1.4.17. *Event-Based Sampling: CUMSUM Filter*
In many cases, decisions (or bets) are made after a significant event. To enable a machine learning algorithm to learn a predictive function in such scenarios, the CUSUM filter can be employed. The CUSUM filter is a quality-control method designed to detect shifts in the mean value of a measured variable away from a target. Let $\{y_t\}_{t=1}^{T}$ be IID observations from a locally stationary process. The cumulative sum is defined recursively by

$$S_t = \max\{0,\, S_{t-1} + y_t - E_{t-1}[y_t]\}, \quad S_0 = 0.$$

An action will be recommended at the first t satisfying $S_t \geq h$ for some threshold h (filter size).
Note $S_t = 0$ whenever $y_t = E_{t-1}[y_t] - S_{t-1}$, The zero floor means some downward deviations will be skipped. The filter is set up to identify a sequence of upside divergences from any reset level zero.
The threshold is activated when

$$S_t \geq h \Leftrightarrow \exists \tau \in [1, t] \mid \sum_{i=\tau}^{t} (y_i - E_{i-1}[y_t]) \geq h$$

The concept can be extended to capture symmetric movements by defining

$$\begin{aligned} S_t^+ &= \max\{0,\, S_{t-1}^+ + y_t - E_{t-1}[y_t]\}, \quad S_0^+ = 0, \\ S_t^- &= \min\{0,\, S_{t-1}^- + y_t - E_{t-1}[y_t]\}, \quad S_0^- = 0, \\ S_t &= \max\{S_t^+,\, -S_t^-\}. \end{aligned}$$

1.4.3 Data Labelling Techniques

Method 1.4.18. *Labelling with Fixed-Time Horizon Method*
Consider a features matrix X composed of I rows, where each observation $\{X_i\}_{i=1,\ldots,I}$ is drawn from a series of bars indexed by $t = 1, \ldots, T$ (with $I \leq T$). In this method, each observation X_i is assigned a label $y_i \in \{-1, 0, 1\}$ according to the following rule:

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

$$r_{t_{i,0}, t_{i,0}+h} = \frac{p_{t_{i,0}+h}}{p_{t_{i,0}}} - 1$$

Here, τ is a predetermined constant threshold, $t_{i,0}$ represents the index of the bar immediately after the occurrence of X_i, and $t_{i,0} + h$ denotes the h-th bar following $t_{i,0}$.

Remark 1.4.19. *Limitations of Fixed-Time Horizon Method*

i. Time bars do not exhibit desirable statistical properties (as discussed previously).

ii. The same threshold τ is uniformly applied regardless of the prevailing volatility. It is often beneficial to compute daily volatility at intraday estimation points by applying an exponentially weighted moving standard deviation over a span of n days.

Method 1.4.20. *Labelling with Triple-Barrier Method*
This method assigns a label to an observation based on the first barrier touched among three defined barriers.

i. Establish two horizontal barriers and one vertical barrier. Horizontal barriers are defined by profit-taking and stop-loss limits, dynamically determined as a function of estimated volatility (either realised or implied). The vertical barrier represents the expiration limit in terms of the number of bars elapsed since the position was taken.

ii. If the upper barrier is touched first, label the observation as 1. Conversely, if the lower barrier is reached first, label it as -1. If the vertical barrier is touched first, label the observation either by the sign of the return or with 0.

Note that this method is path-dependent. To label an observation, one must consider the entire price path over the interval $[t_{i,0}, t_{i,0} + h]$, where h defines the vertical barrier (expiration limit). Let $t_{i,1}$ be the time at which first barrier is touched, with the associated return given by $r_{t_{i,0}, t_{i,1}}$. horizontal barriers may not be symmetric.

Remark 1.4.21. *Triple-Barrier Method Configurations*
A barrier configuration is denoted by the triplet $[pt, sl, t1]$, which represents the upper barrier, lower barrier, and the physical (vertical) barrier, respectively. A value of 0 indicates that the barrier is inactive, while a value of 1 signifies an active barrier. The three useful configurations are:

i. $[1, 1, 1]$: To capture profit while imposing a maximum tolerance for losses and a defined holding period.

ii. $[0, 1, 1]$: To exit after a set number of bars unless a stop-loss is triggered.

iii. $[1, 1, 0]$: To secure profit as long as the stop-loss is not hit.

The three less realistic configurations are:

i. $[0, 0, 1]$: Essentially equivalent to the fixed-time horizon method.

ii. $[1, 0, 1]$: Position held until a profit is achieved or the maximum holding period is exceeded, ignoring any immediate unrealised losses.

iii. $[1, 0, 0]$: Position maintained until a profit is made, potentially resulting in a very prolonged holding period.

The two illogical configurations are:

i. $[0, 1, 0]$: This configuration is ambiguous, as it holds the position solely until a stop-loss occurs.

ii. $[0, 0, 0]$: With no barriers in place, the position remains open indefinitely, and no label can be generated.

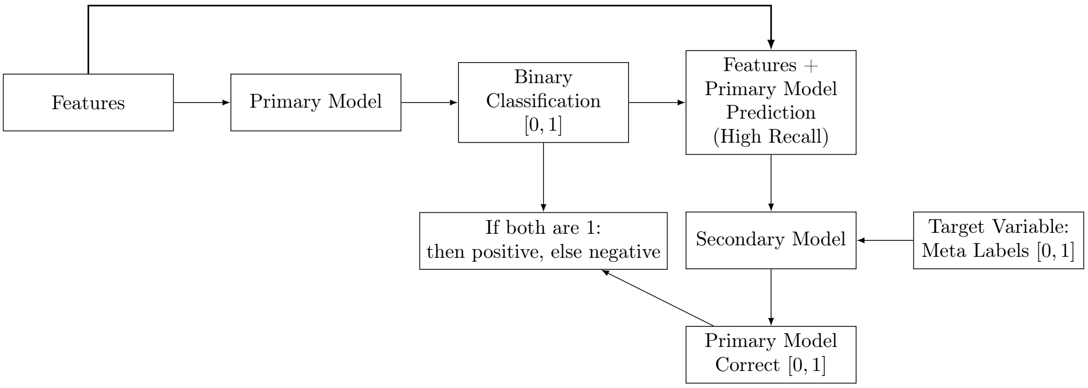
Features
Primary Model
Binary Classification [0, 1]
Features + Primary Model Prediction (High Recall)
If both are 1: then positive, else negative
Secondary Model
Target Variable: Meta Labels [0, 1]
Primary Model Correct [0, 1]

Figure 3: Meta-Labelling Process

Method 1.4.22. *Meta-Labelling*
Meta-labelling is a technique that is particularly effective for achieving higher F1-scores.
The process begins by constructing a primary model that prioritizes high recall, even if its precision is not optimal. Subsequently, meta-labelling is applied to the positive predictions made by the primary model in order to filter out false positives. In essence, the secondary model is designed to discern whether a positive prediction from the primary model is truly valid.

i. Train a primary binary classification model.

ii. Determine a threshold at which the primary model attains high recall; ROC curves can be employed to aid in selecting an appropriate threshold.

iii. Construct a secondary model using features that typically include:

   i. The primary model's features concatenated with its predictions.
   ii. Indicators of the current market state.
   iii. Features that signal potential false positives.
   iv. Distribution-related characteristics.
   v. Recent performance metrics of the primary model.

   The meta labels serve as the target variable for this secondary model.

iv. The final prediction is made by combining the outputs of both models—only when both the primary and secondary models predict a positive does the observation receive a true positive label.

Remark 1.4.23. *Limitations of Meta-Labelling*

i. If the primary model overfits the data, the addition of meta-labelling may provide little to no benefit.

ii. When trades are not considered independent observations, the meta-model may inadvertently be forced to capture day-to-day exposures, which is not the intended application of the technique.

iii. This technique involves trading recall for precision; it requires a large number of trades for effective training, while accepting a reduced frequency of trades.

References

Narang, R. K. (2013). *Inside the Black Box: The Simple Truth About Quantitative Trading.* Wiley Finance Series. Wiley.

Prado, M. L. D. (2018). *Advances in Financial Machine Learning.* Wiley Finance Series. Wiley.

Thierry, A. and G. Helyette (2000, October). Order flow, transaction clock, and normality of returns. *The Journal of Finance 55*(5), 2259–2284.

Velu, R. (2020). *Algorithmic Trading and Quantitative Strategies.* CRC Press.