1
Machine Learning for Trading – From Idea to Execution

Algorithmic trading relies on computer programs that execute algorithms to automate some or all elements of a trading strategy. **Algorithms** are a sequence of steps or rules designed to achieve a goal. They can take many forms and facilitate optimization throughout the investment process, from idea generation to asset allocation, trade execution, and risk management.

Machine learning (**ML**) involves algorithms that learn rules or patterns from data to achieve a goal such as minimizing a prediction error. The examples in this book will illustrate how ML algorithms can extract information from data to support or automate key investment activities. These activities include observing the market and analyzing data to form expectations about the future and decide on placing buy or sell orders, as well as managing the resulting portfolio to produce attractive returns relative to the risk.

Ultimately, the goal of active investment management is to generate alpha, defined as portfolio returns in excess of the benchmark used for evaluation. The **fundamental law of active management** postulates that the key to generating alpha is having accurate return forecasts combined with the ability to act on these forecasts (Grinold 1989; Grinold and Kahn 2000).

This law defines the **information ratio** (**IR**) to express the value of active management as the ratio of the return difference between the

portfolio and a benchmark to the volatility of those returns. It further approximates the IR as the product of the following:

- The **information coefficient** (**IC**), which measures the quality of forecasts as their rank correlation with outcomes
- The square root of the **breadth of a strategy** expressed as the number of independent bets on these forecasts

The competition of sophisticated investors in financial markets implies that making precise predictions to generate alpha requires superior information, either through access to better data, a superior ability to process it, or both.

This is where ML comes in: applications of **ML for trading** (**ML4T**) typically aim to make more efficient use of a rapidly diversifying range of data to produce both better and more actionable forecasts, thus improving the quality of investment decisions and results.

Historically, algorithmic trading used to be more narrowly defined as the automation of trade execution to minimize the costs offered by the sell-side. This book takes a more comprehensive perspective since the use of algorithms in general and ML in particular has come to impact a broader range of activities, from generating ideas and extracting signals from data to asset allocation, position-sizing, and testing and evaluating strategies.

This chapter looks at industry trends that have led to the emergence of ML as a source of competitive advantage in the investment industry. We will also look at where ML fits into the investment process to enable algorithmic trading strategies. More specifically, we will be covering the following topics:

- Key trends behind the rise of ML in the investment industry
- The design and execution of a trading strategy that leverages ML
- Popular use cases for ML in trading

> You can find links to additional resources and references in the README file for this chapter in the GitHub repository (https://github.com/PacktPublishing/Machine-Learning-for-Algorithmic-Trading-Second-Edition).

The rise of ML in the investment industry

The investment industry has evolved dramatically over the last several decades and continues to do so amid increased competition, technological advances, and a challenging economic environment. This section reviews key trends that have shaped the overall investment environment and the context for algorithmic trading and the use of ML more specifically.

The trends that have propelled algorithmic trading and ML to their current prominence include:

- Changes in the **market microstructure**, such as the spread of electronic trading and the integration of markets across asset classes and geographies
- The development of investment strategies framed in terms of **risk-factor exposure**, as opposed to asset classes
- The revolutions in **computing power**, **data generation and management**, and **statistical methods**, including breakthroughs in deep learning
- The **outperformance of the pioneers** in algorithmic trading relative to human, discretionary investors

In addition, the financial crises of 2001 and 2008 have affected how investors approach diversification and risk management. One outcome is the rise in low-cost **passive investment vehicles** in the form of **exchange-traded funds** (**ETFs**).

Amid low yields and low volatility following the 2008 crisis, which triggered large-scale asset purchases by leading central banks, cost-conscious investors shifted over $3.5 trillion from actively managed mutual funds into passively managed ETFs.

Competitive pressure is also reflected in **lower hedge fund fees**, which dropped from the traditional 2 percent annual management fee and 20 percent take of profits to an average of 1.48 percent and 17.4 percent, respectively, in 2017.

From electronic to high-frequency trading

Electronic trading has advanced dramatically in terms of capabilities, volume, coverage of asset classes, and geographies since networks started routing prices to computer terminals in the 1960s. Equity markets have been at the forefront of this trend worldwide. See Harris (2003) and Strumeyer (2017) for comprehensive coverage of relevant changes in financial markets; we will return to this topic when we cover how to work with market and fundamental data in the next chapter.

The 1997 order-handling rules by the SEC introduced competition to exchanges through **electronic communication networks** (**ECNs**). ECNs are automated **alternative trading systems** (**ATS**) that match buy-and-sell orders at specified prices, primarily for equities and currencies, and are registered as broker-dealers. It allows significant brokerages and individual traders in different geographic locations to trade directly without intermediaries, both on exchanges and after hours.

Dark pools are another type of private ATS that allows institutional investors to trade large orders without publicly revealing their information, contrary to how exchanges managed their order books prior to competition from ECNs. Dark pools do not publish pre-trade bids and offers, and trade prices only become public some time after exe-

cution. They have grown substantially since the mid-2000s to account for 40 percent of equities traded in the US due to concerns about adverse price movements of large orders and order front-running by high-frequency traders. They are often housed within large banks and are subject to SEC regulation.

With the rise of electronic trading, **algorithms for cost-effective execution** developed rapidly and adoption spread quickly from the sell-side to the buy-side and across asset classes. Automated trading emerged around 2000 as a sell-side tool aimed at cost-effective execution that broke down orders into smaller, sequenced chunks to limit their market impact. These tools spread to the buy side and became increasingly sophisticated by taking into account, for example, transaction costs and liquidity, as well as short-term price and volume forecasts.

Direct market access (DMA) gives a trader greater control over execution by allowing them to send orders directly to the exchange using the infrastructure and market participant identification of a broker who is a member of an exchange. Sponsored access removes pre-trade risk controls by the brokers and forms the basis for **high-frequency trading (HFT)**.

HFT refers to automated trades in financial instruments that are executed with extremely low latency in the microsecond range and where participants hold positions for very short periods. The goal is to detect and exploit **inefficiencies in the market microstructure**, the institutional infrastructure of trading venues.

HFT has grown substantially over the past 10 years and is estimated to make up roughly 55 percent of trading volume in US equity markets and about 40 percent in European equity markets. HFT has also grown in futures markets to roughly 80 percent of foreign-exchange futures volumes and two-thirds of both interest rate and Treasury 10-year futures volumes (Miller 2016).

HFT strategies aim to earn small profits per trade using **passive or aggressive strategies**. Passive strategies include arbitrage trading to profit from very small price differentials for the same asset, or its derivatives, traded on different venues. Aggressive strategies include order anticipation or momentum ignition. Order anticipation, also known as liquidity detection, involves algorithms that submit small exploratory orders to detect hidden liquidity from large institutional investors and trade ahead of a large order to benefit from subsequent price movements. Momentum ignition implies an algorithm executing and canceling a series of orders to spoof other HFT algorithms into buying (or selling) more aggressively and benefit from the resulting price changes.

Regulators have expressed concern over the potential link between certain aggressive HFT strategies and **increased market fragility and volatility**, such as that experienced during the May 2010 Flash Crash, the October 2014 Treasury market volatility, and the sudden crash by over 1,000 points of the Dow Jones Industrial Average on August 24, 2015. At the same time, market liquidity has increased with trading volumes due to the presence of HFT, which has lowered overall transaction costs.

The combination of reduced trading volumes amid lower volatility and rising costs of technology and access to both data and trading venues has led to financial pressure. Aggregate HFT revenues from US stocks were estimated to have dropped beneath $1 billion in 2017 for the first time since 2008, down from $7.9 billion in 2009. This trend has led to **industry consolidation**, with various acquisitions by, for example, the largest listed proprietary trading firm, Virtu Financial, and shared infrastructure investments, such as the new Go West ultra-low latency route between Chicago and Tokyo. Simultaneously, start-ups such as Alpha Trading Labs are making HFT trading infrastructure and data available to democratize HFT by crowdsourcing algorithms in return for a share of the profits.

Factor investing and smart beta funds

The return provided by an asset is a function of the uncertainty or risk associated with the investment. An equity investment implies, for example, assuming a company's business risk, and a bond investment entails default risk. To the extent that **specific risk characteristics predict returns**, identifying and forecasting the behavior of these risk factors becomes a primary focus when designing an investment strategy. It yields valuable trading signals and is the key to superior active-management results. The industry's understanding of risk factors has evolved very substantially over time and has impacted how ML is used for trading. *Chapter 4*, *Financial Feature Engineering – How to Research Alpha Factors*, and *Chapter 5*, *Portfolio Optimization and Performance Evaluation*, will dive deeper into the practical applications of the concepts outlined here; see Ang (2014) for comprehensive coverage.

Modern portfolio theory (**MPT**) introduced the distinction between idiosyncratic and systematic sources of risk for a given asset. Idiosyncratic risk can be eliminated through diversification, but systematic risk cannot. In the early 1960s, the **capital asset pricing model** (**CAPM**) identified a single factor driving all asset returns: the return on the market portfolio in excess of T-bills. The market portfolio consisted of all tradable securities, weighted by their market value. The systematic exposure of an asset to the market is measured by **beta**, which is the correlation between the returns of the asset and the market portfolio.

The recognition that the risk of an asset does not depend on the asset in isolation, but rather how it moves relative to other assets and the market as a whole, was a major conceptual breakthrough. In other words, assets earn a **risk premium** based on their exposure to underlying, **common risks** experienced by all assets, not due to their specific, idiosyncratic characteristics.

Subsequently, academic research and industry experience have raised numerous critical questions regarding the CAPM prediction that an asset's risk premium depends only on its exposure to a single factor measured by the asset's beta. Instead, **numerous additional risk factors** have since been discovered. A factor is a quantifiable signal, attribute, or any variable that has historically correlated with future stock returns and is expected to remain correlated in the future.

These risk factors were labeled anomalies since they contradicted the **efficient market hypothesis** (**EMH**). The EMH maintains that market equilibrium would always price securities according to the CAPM so that no other factors should have predictive power (Malkiel 2003). The economic theory behind factors can be either rational, where factor risk premiums compensate for low returns during bad times, or behavioral, where agents fail to arbitrage away excess returns.

Well-known anomalies include the value, size, and momentum effects that help predict returns while controlling for the CAPM market factor. The **size effect** rests on small firms systematically outperforming large firms (Banz 1981; Reinganum 1981). The **value effect** (Basu et. al. 1981) states that firms with low valuation metrics outperform their counterparts with the opposite characteristics. It suggests that firms with low price multiples, such as the price-to-earnings or the price-to-book ratios, perform better than their more expensive peers (as suggested by the inventors of value investing, Benjamin Graham and David Dodd, and popularized by Warren Buffet).

The **momentum effect**, discovered in the late 1980s by, among others, Clifford Asness, the founding partner of AQR, states that stocks with good momentum, in terms of recent 6-12 month returns, have higher returns going forward than poor momentum stocks with similar market risk. Researchers also found that value and momentum factors explain returns for stocks outside the US, as well as for other asset classes, such as bonds, currencies, and commodities, and additional

risk factors (Jegadeesh and Titman 1993; Asness, Moskowitz, and Pedersen 2013).

In fixed income, the value strategy is called **riding the yield curve** and is a form of the duration premium. In commodities, it is called the **roll return**, with a positive return for an upward-sloping futures curve and a negative return otherwise. In foreign exchange, the value strategy is called **carry**.

There is also an **illiquidity premium**. Securities that are more illiquid trade at low prices and have high average excess returns, relative to their more liquid counterparts. Bonds with a higher default risk tend to have higher returns on average, reflecting a credit risk premium. Since investors are willing to pay for insurance against high volatility when returns tend to crash, sellers of volatility protection in options markets tend to earn high returns.

Multifactor models define risks in broader and more diverse terms than just the market portfolio. In 1976, Stephen Ross proposed the **arbitrage pricing theory**, which asserted that investors are compensated for multiple systematic sources of risk that cannot be diversified away (Roll and Ross 1984). The three most important macro factors are growth, inflation, and volatility, in addition to productivity, demographic, and political risk. In 1993, Eugene Fama and Kenneth French combined the equity risk factors' size and value with a market factor into a single three-factor model that better explained cross-sectional stock returns. They later added a model that also included bond risk factors to simultaneously explain returns for both asset classes (Fama and French 1993; 2015).

A particularly attractive aspect of risk factors is their **low or negative correlation**. Value and momentum risk factors, for instance, are negatively correlated, reducing the risk and increasing risk-adjusted returns above and beyond the benefit implied by the risk factors. Furthermore, using leverage and long-short strategies, factor strate-

gies can be combined into **market-neutral approaches**. The combination of long positions in securities exposed to positive risks with underweight or short positions in the securities exposed to negative risks allows for the collection of dynamic risk premiums.

As a result, the factors that explained returns above and beyond the CAPM were incorporated into investment styles that tilt portfolios in favor of one or more factors, and assets began to migrate into factor-based portfolios. The 2008 financial crisis underlined how asset-class labels could be highly misleading and create a false sense of diversification when investors do not look at the underlying factor risks, as asset classes came crashing down together.

Over the past several decades, quantitative factor investing has evolved from a simple approach based on two or three styles to **multi-factor smart or exotic beta products**. Smart beta funds have crossed $1 trillion AUM in 2017, testifying to the popularity of the hybrid investment strategy that combines active and passive management. **Smart beta funds** take a passive strategy but modify it according to one or more factors, such as cheaper stocks or screening them according to dividend payouts, to generate better returns. This growth has coincided with increasing criticism of the high fees charged by traditional active managers as well as heightened scrutiny of their performance.

The ongoing discovery and successful forecasting of risk factors that, either individually or in combination with other risk factors, significantly impact future asset returns across asset classes is a key driver of the surge in ML in the investment industry and will be a key theme throughout this book.

Algorithmic pioneers outperform humans

The track record and growth of **assets under management** (**AUM**) of firms that spearheaded algorithmic trading has played a key role in

generating investor interest and subsequent industry efforts to replicate their success. **Systematic funds differ from HFT** in that trades may be held significantly longer while seeking to exploit arbitrage opportunities as opposed to advantages from sheer speed.

Systematic strategies that mostly or exclusively rely on algorithmic decision-making were most famously introduced by mathematician James Simons, who founded **Renaissance Technologies** in 1982 and built it into the premier quant firm. Its secretive Medallion Fund, which is closed to outsiders, has earned an estimated annualized return of 35 percent since 1982.

D. E. Shaw, Citadel, and Two Sigma, three of the most prominent quantitative hedge funds that use systematic strategies based on algorithms, rose to the all-time top-20 performers for the first time in 2017, in terms of total dollars earned for investors, after fees, and since inception.

D. E. Shaw, founded in 1988 and with $50 billion in AUM in 2019, joined the list at number 3. Citadel, started in 1990 by Kenneth Griffin, manages $32 billion, and ranked 5. Two Sigma, started only in 2001 by D. E. Shaw alumni John Overdeck and David Siegel, has grown from $8 billion in AUM in 2011 to $60 billion in 2019. **Bridgewater**, started by Ray Dalio in 1975, had over $160 billion in AUM in 2019 and continues to lead due to its Pure Alpha fund, which also incorporates systematic strategies.

Similarly, on the Institutional Investors 2018 Hedge Fund 100 list, the four largest firms, and five of the top six firms, rely largely or completely on computers and trading algorithms to make investment decisions—and all of them have been growing their assets in an otherwise challenging environment. Several quantitatively focused firms climbed the ranks and, in some cases, grew their assets by double-digit percentages. Number 2-ranked **Applied Quantitative Research (AQR)**

grew its hedge fund assets by 48 percent in 2017 and by 29 percent in 2018 to nearly $90 billion.

ML-driven funds attract $1 trillion in AUM

The familiar three revolutions in computing power, data availability, and statistical methods have made the adoption of systematic, data-driven strategies not only more compelling and cost-effective but a key source of competitive advantage.

As a result, algorithmic approaches are not only finding **wider application** in the hedge-fund industry that pioneered these strategies but across a broader range of asset managers and even passively managed vehicles such as ETFs. In particular, **predictive analytics** using ML and algorithmic automation play an increasingly prominent role in all steps of the investment process across asset classes, from idea generation and research to strategy formulation and portfolio construction, trade execution, and risk management.

Estimates of **industry size** vary because there is no objective definition of a quantitative or algorithmic fund. Many traditional hedge funds or even mutual funds and ETFs are introducing computer-driven strategies or integrating them into a discretionary environment in a human-plus-machine approach.

According to the *Economist*, in 2016, systematic funds became the largest driver of institutional trading in the US stock market (ignoring HFT, which mainly acts as a middleman). In 2019, they accounted for over 35 percent of institutional volume, up from just 18 percent in 2010; just 10% of trading is still due to traditional equity funds. Measured by the Russell 3000 index, the **value of US stocks** is around $31 trillion. The three types of **computer-managed funds**—index funds, ETFs, and quant funds—**run around 35 percent**, whereas human managers at traditional hedge funds and other mutual funds manage just 24 percent.

The market research firm Preqin estimates that almost 1,500 hedge funds make a majority of their trades with help from computer models. Quantitative hedge funds are now responsible for 27 percent of all US stock trades by investors, up from 14 percent in 2013. But many use data scientists—or quants—who, in turn, use machines to build large statistical models.

In recent years, however, funds have moved toward true ML, where artificially intelligent systems can analyze large amounts of data at speed and improve themselves through such analyses. Recent examples include Rebellion Research, Sentient, and Aidyia, which rely on evolutionary algorithms and deep learning to devise fully automatic **artificial intelligence** (**AI**)-driven investment platforms.

From the core hedge fund industry, the adoption of algorithmic strategies has spread to mutual funds and even passively managed EFTs in the form of smart beta funds, and to discretionary funds in the form of quantamental approaches.

The emergence of quantamental funds

Two distinct approaches have evolved in active investment management: **systematic** (**or quant**) and **discretionary investing**. Systematic approaches rely on algorithms for a repeatable and data-driven approach to identify investment opportunities across many securities. In contrast, a discretionary approach involves an in-depth analysis of the fundamentals of a smaller number of securities. These two approaches are becoming more similar as fundamental managers take more data science-driven approaches.

Even **fundamental traders** now arm themselves with quantitative techniques, accounting for $55 billion of systematic assets, according to Barclays. Agnostic to specific companies, quantitative funds trade based on patterns and dynamics across a wide swath of securities.

Such quants accounted for about 17 percent of total hedge fund assets, as data compiled by Barclays in 2018 showed.

Point72, with $14 billion in assets, has been shifting about half of its portfolio managers to a human-plus-machine approach. Point72 is also investing tens of millions of dollars into a group that analyzes large amounts of alternative data and passes the results on to traders.

Investments in strategic capabilities

Three trends have boosted the use of data in algorithmic trading strategies and may further shift the investment industry from discretionary to quantitative styles:

- The exponential increase in the availability of digital data
- The increase in computing power and data storage capacity at a lower cost
- The advances in statistical methods for analyzing complex datasets

Rising investments in related capabilities—technology, data, and, most importantly, skilled humans—highlight how significant algorithmic trading using ML has become for competitive advantage, especially in light of the rising popularity of passive, indexed investment vehicles, such as ETFs, since the 2008 financial crisis.

Morgan Stanley noted that only 23 percent of its quant clients say they are not considering using or not already using ML, down from 44 percent in 2016. **Guggenheim Partners** built what it calls a supercomputing cluster for $1 million at the Lawrence Berkeley National Laboratory in California to help crunch numbers for Guggenheim's quant investment funds. Electricity for computers costs another $1 million per year.

AQR is a quantitative investment group that relies on academic research to identify and systematically trade factors that have, over time, proven to beat the broader market. The firm used to eschew the

purely computer-powered strategies of quant peers such as Renaissance Technologies or DE Shaw. More recently, however, AQR has begun to seek profitable patterns in markets using ML to parse through novel datasets, such as satellite pictures of shadows cast by oil wells and tankers.

The leading firm **BlackRock**, with over $5 trillion in AUM, also bets on algorithms to beat discretionary fund managers by heavily investing in SAE, a systematic trading firm it acquired during the financial crisis. Franklin Templeton bought Random Forest Capital, a debt-focused, data-led investment company, for an undisclosed amount, hoping that its technology can support the wider asset manager.

ML and alternative data

Hedge funds have long looked for alpha through **informational advantage** and the ability to uncover new uncorrelated signals. Historically, this included things such as proprietary surveys of shoppers, or of voters ahead of elections or referendums.

Occasionally, the use of company insiders, doctors, and expert networks to expand knowledge of industry trends or companies crosses legal lines: a series of prosecutions of traders, portfolio managers, and analysts for using **insider information** after 2010 has shaken the industry.

In contrast, the informational advantage from exploiting conventional and alternative data sources using ML is not related to expert and industry networks or access to corporate management, but rather the ability to collect large quantities of very diverse data sources and analyze them in real time.

Conventional data includes economic statistics, trading data, or corporate reports. **Alternative data** is much broader and includes sources such as satellite images, credit card sales, sentiment analysis, mobile

geolocation data, and website scraping, as well as the conversion of data generated in the ordinary course of business into valuable intelligence. It includes, in principle, **any data source containing (potential) trading signals**.

For instance, data from an insurance company on the sales of new car insurance policies captures not only the volumes of new car sales but can be broken down into brands or geographies. Many vendors scrape websites for valuable data, ranging from app downloads and user reviews to airline and hotel bookings. Social media sites can also be scraped for hints on consumer views and trends.

Typically, the datasets are large and require storage, access, and analysis using **scalable data solutions** for parallel processing, such as Hadoop and Spark. There are more than 1 billion websites with more than 10 trillion individual web pages, with 500 exabytes (or 500 billion gigabytes) of data, according to Deutsche Bank. And more than 100 million websites are added to the internet every year.

Real-time insights into a company's prospects, long before their results are released, can be gleaned from a decline in job listings on its website, the internal rating of its chief executive by employees on the recruitment site Glassdoor, or a dip in the average price of clothes on its website. Such information can be combined with satellite images of car parks and geolocation data from mobile phones that indicate how many people are visiting stores. On the other hand, strategic moves can be learned from a jump in job postings for specific functional areas or in certain geographies.

Among the most valuable sources is data that directly reveals consumer expenditures, with **credit card information** as a primary source. This data offers only a partial view of sales trends, but it can offer vital insights when combined with other data. Point72, for instance, at some point analyzed 80 million credit card transactions every day. We will explore the various sources, their use cases, and how

to evaluate them in detail in *Chapter 3, Alternative Data for Finance – Categories and Use Cases*.

Investment groups have more than doubled their **spending on alternative sets** and data scientists in the past two years, as the asset management industry has tried to reinvigorate its fading fortunes. In December 2018, there were 375 alternative data providers listed on `alternativedata.org` (sponsored by provider Yipit).

Asset managers spent a total of $373 million on datasets and hiring new employees to parse them in 2017, up 60 percent from 2016, and will probably spend a total of $616 million this year, according to a survey of investors by `alternativedata.org`. It forecast that overall expenditures will climb to over $1 billion by 2020. Some estimates are even higher: Optimus, a consultancy, estimates that investors are spending about $5 billion per year on alternative data, and expects the industry to grow 30 percent per year over the coming years.

As competition for valuable data sources intensifies, exclusivity arrangements are a key feature of data-source contracts, to maintain an informational advantage. At the same time, privacy concerns are mounting, and regulators have begun to start looking at the currently largely unregulated data-provider industry.

Crowdsourcing trading algorithms

More recently, several algorithmic trading firms have begun to offer investment platforms that provide access to data and a programming environment to crowdsource risk factors that become part of an investment strategy or entire trading algorithms. Key examples include WorldQuant, Quantopian, and, most recently, Alpha Trading Labs (launched in 2018).

WorldQuant was spun out of Millennium Management (AUM: $41 billion) in 2007, for whom it manages around $5 billion. It employs hun-

dreds of scientists and many more part-time workers around the world in its alpha factory, which organizes the investment process as a quantitative assembly line. This factory claims to have produced 4 million successfully tested alpha factors for inclusion in more complex trading strategies and is aiming for 100 million. Each alpha factor is an algorithm that seeks to predict a future asset price change. Other teams then combine alpha factors into strategies and strategies into portfolios, allocate funds between portfolios, and manage risk while avoiding strategies that cannibalize each other. See the *Appendix*, *Alpha Factor Library*, for dozens of examples of quantitative factors used at WorldQuant.

Designing and executing an ML-driven strategy

In this book, we demonstrate **how ML fits into the overall process of designing, executing, and evaluating a trading strategy**. To this end, we'll assume that an ML-based strategy is driven by data sources that contain predictive signals for the target universe and strategy, which, after suitable preprocessing and feature engineering, permit an ML model to predict asset returns or other strategy inputs. The model predictions, in turn, translate into buy or sell orders based on human discretion or automated rules, which in turn may be manually encoded or learned by another ML algorithm in an end-to-end approach.

Figure 1.1 depicts the key steps in this workflow, which also shapes the organization of this book:

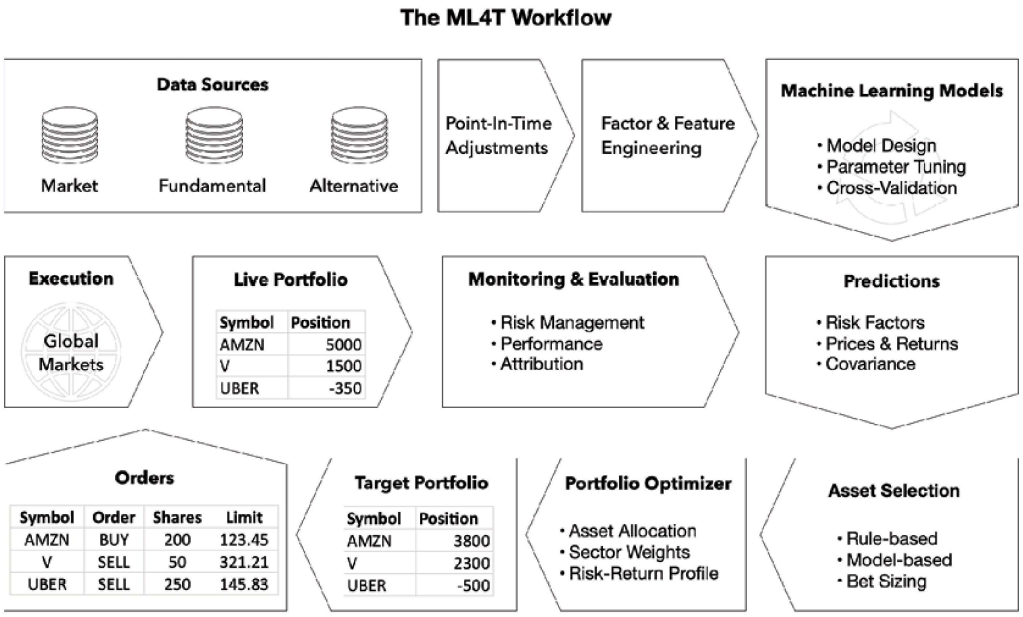

Figure 1.1: The ML4T workflow

Part 1 introduces important skills and techniques that apply across different strategies and ML use cases. These include the following:

- How to source and manage important data sources
- How to engineer informative features or alpha factors that extract signal content
- How to manage a portfolio and track strategy performance

Moreover, *Chapter 8*, *The ML4T Workflow – From Model to Strategy Backtesting*, in Part 2, covers strategy backtesting. We will briefly outline each of these areas before turning to relevant ML use cases, which make up the bulk of the book in Parts 2, 3, and 4.

Sourcing and managing data

The dramatic evolution of data availability in terms of volume, variety, and velocity is a key complement to the application of ML to trading, which in turn has boosted industry spending on the acquisition of new data sources. However, the proliferating supply of data requires care-

ful selection and management to uncover the potential value, including the following steps:

1. Identify and evaluate market, fundamental, and alternative data sources containing alpha signals that do not decay too quickly.
2. Deploy or access a cloud-based scalable data infrastructure and analytical tools like Hadoop or Spark to facilitate fast, flexible data access.
3. Carefully manage and curate data to avoid look-ahead bias by adjusting it to the desired frequency on a point-in-time basis. This means that data should reflect only information available and known at the given time. ML algorithms trained on distorted historical data will almost certainly fail during live trading.

We will cover these aspects in practical detail in *Chapter 2*, *Market and Fundamental Data – Sources and Techniques*, and *Chapter 3*, *Alternative Data for Finance – Categories and Use Cases*.

From alpha factor research to portfolio management

Alpha factors are designed to extract signals from data to predict returns for a given investment universe over the trading horizon. A typical factor takes on a single value for each asset when evaluated at a given point in time, but it may combine one or several input variables or time periods. If you are already familiar with the ML workflow (see *Chapter 6*, *The Machine Learning Process*), you may view alpha factors as domain-specific features designed for a specific strategy. Working with alpha factors entails a research phase and an execution phase as outlined in *Figure 1.2*:

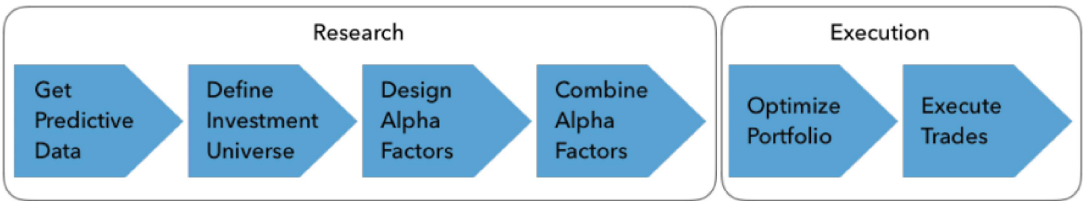

Figure 1.2: The alpha factor research process

The research phase

The **research phase** includes the design and evaluation of alpha factors. A **predictive factor** captures some aspect of a systematic relationship between a data source and an important strategy input like asset returns. Optimizing the predictive power requires creative feature engineering in the form of effective data transformations.

False discoveries due to **data mining** are a key risk that requires careful management. One way of reducing the risk is to focus the search process by following the guidance of decades of academic research that has produced several Nobel prizes. Many investors still prefer factors that align with theories about financial markets and investor behavior. Laying out these theories is beyond the scope of this book, but the references highlight avenues to dive deeper into this important framing aspect.

Validating the signal content of an alpha factor requires a **robust estimate of its predictive power** in a representative context. There are numerous methodological and practical pitfalls that undermine a reliable estimate. In addition to data mining and the failure to correct for multiple testing bias, these pitfalls include the use of data contaminated by survivorship or look-ahead bias, not reflecting realistic **Principal, Interest and Taxes** (**PIT**) information. *Chapter 4*, *Financial Feature Engineering – How to Research Alpha Factors*, discusses how to successfully manage this process.

The execution phase

During the **execution phase**, alpha factors emit signals that lead to buy or sell orders. The resulting portfolio holdings, in turn, have specific risk profiles that interact and contribute to the aggregate portfolio risk. Portfolio management involves optimizing position sizes to

achieve a balance of return and risk of the portfolio that aligns with the investment objectives.

Chapter 5, *Portfolio Optimization and Performance Evaluation*, introduces key techniques and tools applicable to this phase of the trading strategy workflow, from portfolio optimization to performance measurement.

Strategy backtesting

Incorporating an investment idea into a real-life algorithmic strategy implies a significant risk that requires a **scientific approach**. Such an approach involves extensive empirical tests with the goal of rejecting the idea based on its performance in alternative out-of-sample market scenarios. Testing may involve simulated data to capture scenarios deemed possible but not reflected in historic data.

To obtain unbiased performance estimates for a candidate strategy, we need a **backtesting engine** that simulates its execution in a realistic manner. In addition to the potential biases introduced by the data or a flawed use of statistics, the backtesting engine needs to accurately represent the practical aspects of trade-signal evaluation, order placement, and execution in line with market conditions.

Chapter 8, *The ML4T Workflow – From Model to Strategy Backtesting*, shows how to use backtrader and Zipline and navigate the multiple methodological challenges and completes the introduction to the end-to-end ML4T workflow.

ML for trading – strategies and use cases

In practice, we apply ML to trading in the context of a specific strategy to meet a certain business goal. In this section, we briefly describe

how trading strategies have evolved and diversified, and outline real-world examples of ML applications, highlighting how they relate to the content covered in this book.

The evolution of algorithmic strategies

Quantitative strategies have evolved and become more sophisticated in three waves:

1. In the 1980s and 1990s, signals often emerged from **academic research** and used a single or very few inputs derived from market and fundamental data. AQR, one of the largest quantitative hedge funds today, was founded in 1998 to implement such strategies at scale. These signals are now largely commoditized and available as ETF, such as basic mean-reversion strategies.
2. In the 2000s, **factor-based investing** proliferated based on the pioneering work by Eugene Fama and Kenneth French and others. Funds used algorithms to identify assets exposed to risk factors like value or momentum to seek arbitrage opportunities. Redemptions during the early days of the financial crisis triggered the quant quake of August 2007, which cascaded through the factor-based fund industry. These strategies are now also available as long-only smart beta funds that tilt portfolios according to a given set of risk factors.
3. The third era is driven by investments in **ML capabilities and alternative** data to generate profitable signals for repeatable trading strategies. Factor decay is a major challenge: the excess returns from new anomalies have been shown to drop by a quarter from discovery to publication, and by over 50 percent after publication due to competition and crowding.

Today, traders pursue a range of different objectives when using algorithms to execute rules:

- Trade execution algorithms that aim to achieve favorable pricing

- Short-term trades that aim to profit from small price movements, for example, due to arbitrage
- Behavioral strategies that aim to anticipate the behavior of other market participants
- Trading strategies based on absolute and relative price and return predictions

Trade-execution programs aim to limit the market impact of trades and range from the simple slicing of trades to match time-weighted or volume-weighted average pricing. Simple algorithms leverage historical patterns, whereas more sophisticated versions take into account transaction costs, implementation shortfall, or predicted price movements.

HFT funds most prominently rely on very short holding periods to benefit from minor price movements based on bid-ask or statistical arbitrage. **Behavioral algorithms** usually operate in lower-liquidity environments and aim to anticipate moves by a larger player with significant price impact, based, for example, on sniffing algorithms that generate insights into other market participants' strategies.

In this book, we will focus on strategies that trade based on expectations of relative price changes over various time horizons beyond the very short term, dominated by latency advantages, because they are both widely used and very suitable for the application of ML.

Use cases of ML for trading

ML is capable of extracting tradable signals from a wide range of market, fundamental, and alternative data and is thus applicable to strategies targeting a range of asset classes and investment horizons. More generally, however, it is a flexible tool to support or automate decisions with quantifiable goals and digital data relevant to achieving these goals. Therefore, it can be applied at several steps of the trading

process. There are numerous use cases in different categories, including:

- Data mining to identify patterns, extract features, and generate insights
- Supervised learning to generate risk factors or alphas and create trade ideas
- The aggregation of individual signals into a strategy
- The allocation of assets according to risk profiles learned by an algorithm
- The testing and evaluation of strategies, including through the use of synthetic data
- The interactive, automated refinement of a strategy using reinforcement learning

We briefly highlight some of these applications and identify where we will demonstrate their use in later chapters.

Data mining for feature extraction and insights

The cost-effective evaluation of large, complex datasets requires the detection of signals at scale. There are several examples throughout the book:

- **Information theory** helps estimate a signal content of candidate features and is thus useful for extracting the most valuable inputs for an ML model. In *Chapter 4, Financial Feature Engineering – How to Research Alpha Factors*, we use mutual information to compare the potential values of individual features for a supervised learning algorithm to predict asset returns. Chapter 18 in De Prado (2018) estimates the information content of a price series as a basis for deciding between alternative trading strategies.
- **Unsupervised learning** provides a broad range of methods to identify structure in data to gain insights or help solve a downstream task. We provide several examples:

    - In *Chapter 13*, *Data-Driven Risk Factors and Asset Allocation with Unsupervised Learning*, we introduce clustering and dimensionality reduction to generate features from high-dimensional datasets.
    - In *Chapter 15*, *Topic Modeling – Summarizing Financial News*, we apply Bayesian probability models to summarize financial text data.
    - In *Chapter 20*, *Autoencoders for Conditional Risk Factors and Asset Pricing*, we use deep learning to extract nonlinear risk factors conditioned on asset characteristics and predict stock returns based on Kelly et al. (2020).
- **Model transparency** emphasizes model-specific ways to gain insights into the predictive power of individual variables and introduce a novel game-theoretic approach called **SHapley Additive exPlanations** (**SHAP**). We apply it to gradient boosting machines with a large number of input variables in *Chapter 12*, *Boosting Your Trading Strategy*, and the *Appendix*, *Alpha Factor Library*.

Supervised learning for alpha factor creation

The most familiar rationale for applying ML to trading is to obtain predictions of asset fundamentals, price movements, or market conditions. A strategy can leverage multiple ML algorithms that build on each other:

- **Downstream models** can generate signals at the portfolio level by integrating predictions about the prospects of individual assets, capital market expectations, and the correlation among securities.
- Alternatively, ML predictions can inform **discretionary trades** as in the quantamental approach outlined previously.

ML predictions can also **target specific risk factors**, such as value or volatility, or implement technical approaches, such as trend-following or mean reversion:

- In *Chapter 3*, *Alternative Data for Finance – Categories and Use Cases*, we illustrate how to work with fundamental data to create inputs to ML-driven valuation models.
- In *Chapter 14*, *Text Data for Trading – Sentiment Analysis*, *Chapter 15*, *Topic Modeling – Summarizing Financial News*, and *Chapter 16*, *Word Embeddings for Earnings Calls and SEC Filings*, we use alternative data on business reviews that can be used to project revenues for a company as an input for a valuation exercise.
- In *Chapter 9*, *Time-Series Models for Volatility Forecasts and Statistical Arbitrage*, we demonstrate how to forecast macro variables as inputs to market expectations and how to forecast risk factors such as volatility.
- In *Chapter 19*, *RNNs for Multivariate Time Series and Sentiment Analysis*, we introduce recurrent neural networks that achieve superior performance with nonlinear time series data.

Asset allocation

ML has been used to allocate portfolios based on decision-tree models that compute a hierarchical form of risk parity. As a result, risk characteristics are driven by patterns in asset prices rather than by asset classes and achieve superior risk-return characteristics.

In *Chapter 5*, *Portfolio Optimization and Performance Evaluation*, and *Chapter 13*, *Data-Driven Risk Factors and Asset Allocation with Unsupervised Learning*, we illustrate how hierarchical clustering extracts data-driven risk classes that better reflect correlation patterns than conventional asset class definition (see *Chapter 16* in De Prado 2018).

Testing trade ideas

Backtesting is a critical step to select successful algorithmic trading strategies. Cross-validation using synthetic data is a key ML technique to generate reliable out-of-sample results when combined with appro-

priate methods to correct for multiple testing. The time-series nature of financial data requires modifications to the standard approach to avoid look-ahead bias or otherwise contaminating the data used for training, validation, and testing. In addition, the limited availability of historical data has given rise to alternative approaches that use synthetic data.

We will demonstrate various methods to test ML models using market, fundamental, and alternative data sources that obtain sound estimates of out-of-sample errors.

In *Chapter 21*, *Generative Adversarial Networks for Synthetic Time-Series Data*, we present **generative adversarial networks (GANs)**, which are capable of producing high-quality synthetic data.

Reinforcement learning

Trading takes place in a competitive, interactive marketplace. Reinforcement learning aims to train agents to learn a policy function based on rewards; it is often considered as one of the most promising areas in financial ML. See, for example, Hendricks and Wilcox (2014) and Nevmyvaka, Feng, and Kearns (2006) for applications to trade execution.

In *Chapter 22*, *Deep Reinforcement Learning – Building a Trading Agent*, we present key reinforcement algorithms like Q-learning to demonstrate the training of reinforcement learning algorithms for trading using OpenAI's Gym environment.

Summary

In this chapter, we reviewed key industry trends around algorithmic trading strategies, the emergence of alternative data, and the use of ML to exploit these new sources of informational advantage. Furthermore, we introduced key elements of the ML4T workflow and

outlined important use cases of ML for trading in the context of different strategies.

In the next two chapters, we will take a closer look at the oil that fuels any algorithmic trading strategy—the market, fundamental, and alternative data sources—using ML.