"From the top-down aspects of macroeconomics and asset allocation to the bottom-up building blocks of markets, securities, factors, and alphas, Robbins covers all bases, artfully combining theory and practice in one complete tome."

—**MARK BAUMGARTNER, PhD, CFA**, Chief Investment Officer at the Carnegie Corporation of New York and Former Chief Investment Officer at the Institute for Advanced Study

QUANTITATIVE ASSET MANAGEMENT

Factor Investing and Machine Learning for Institutional Investing

MICHAEL ROBBINS

"From the top-down aspects of macroeconomics and asset allocation to the bottom-up building blocks of markets, securities, factors, and alphas, Robbins covers all bases, artfully combining theory and practice in one complete tome."

—**MARK BAUMGARTNER, PhD, CFA**, Chief Investment Officer at the Carnegie Corporation of New York and Former Chief Investment Officer at the Institute for Advanced Study

QUANTITATIVE ASSET MANAGEMENT

Factor Investing and Machine Learning for Institutional Investing

MICHAEL ROBBINS

Praise for

QUANTITATIVE ASSET MANAGEMENT

Quantitative Asset Management is a comprehensive practical handbook for designing, building, and operating a quantitative investment business. It is a useful reference for investors and quants grounded in academic research yet wishing to thrive in the unforgiving investing world by exploiting structural advantages, such as multi-period rebalancing and modern machine learning methods.

—Frank J. Fabozzi
Editor, *Journal of Portfolio Management* and coeditor, *Journal of Financial Data Science*

Michael bridges comprehensive risk concepts and its underpinnings to the new generation of artificial intelligence (machine learning) tools to generate investment returns in a volatile and interconnected global market and economy. The book builds on his experience of building and applying new generation artificial intelligence (machine learning) tools across the full risk lifecycle and spectrum of risk sources. The book offers insights into what works and what to watch out for with a practical roadmap for seasoned investment professionals looking to update their knowledge and students looking to bring cutting-edge approaches to an investment career.

—Deven Sharma
Past President, Standard & Poor's

With impressive precision, Michael Robbins has written a remarkable book on the critical questions and answers required when making investment decisions. We need to recognize our limitations and apply the right techniques or tools to operate successfully in the markets. A great and readable book!

—Dr. Ulrich Stephan
Chief Investment Officer, Deutsche Bank

Quantitative methods are a black box for most institutional investors. Michael pierces the historical opacity of this discipline with a comprehensive and practical guide that should help pave the way for better understanding of quantitative approaches by asset owners and improved monitoring of portfolio characteristics and performance by those responsible for governance. From the top-down aspects of macroeconomics and asset allocation to the bottom-up building blocks of markets, securities, factors, and alphas, Michael covers all bases, artfully combining theory and practice in one complete tome.

—Mark Baumgartner, PhD, CFA
Chief Investment Officer at the Carnegie Corporation of New York and Former Chief Investment Officer at the Institute for Advanced Study

The biggest determining factor for successful investing is whether or not there is a system in place. Michael Robbins shares a career's worth of systematic thinking in this excellent book. A must-read for all professionals, regardless of their experience level or style.

—Joshua M. Brown
CEO of Ritholtz Wealth Management and Star of CNBC's *The Halftime Report*

This is a true "how to" book for the investment professional. In addition to providing practical and detailed directions on quantitative investing—covering data analysis, factor investing, asset allocation strategies, and much more—it also provides a framework for managing the process and business of investment management. The book lays out the importance of culture, investment policy statements, and risk management, and even discusses how institutional factors can affect your investment process. Unlike most books in the field, this one truly is geared toward developing total and complete investors.

—Bob Browne
Emeritus Chief Investment Officer at Northern Trust

Michael Robbins and *Quantitative Asset Management* deliver a thought-provoking, revealing, and insightful series of perspectives for professional investors, investment committee stewards, and others. It is a “must read” for those interested in quantitative investing and committed to continuous learning.

—Steven L. Fradkin
President of Wealth Management at Northern Trust

Quantitative Asset Management covers the entire investment process with foundational and practical state-of-the-art quantitative, ML, and AI methods and strategies. Written from an investor’s perspective, Michael provides valuable and innovative insights on the effective use of big data and technology to analyze investment options and enhance performance—while addressing important risk management issues. It’s an excellent text for both seasoned finance professionals and students seeking to efficiently learn the latest investment tools and strategies.

—Andy Naranjo
Susan Cameron Professor of Finance, Chairman, Eugene F. Brigham Finance, Insurance & Real Estate Department, and Director, International Business Center, University of Florida, Warrington College of Business

The challenge with quantitative investing is largely rendering the apparently complex simple. Michael not only manages this feat comfortably, but then goes on to engage the reader by bridging the gap between academia and practice, by offering insights that will make sense to both the experienced investor and interested lay person.

—Eoin Murray
Head of Investment at Federated Hermes Limited

If you care about investing and have an interest in AI and machine learning, then add this book to your basket immediately! Michael neatly weaves all of the above into a practical "do's and don'ts" of investment management.

—Dean Berry
Group Head of Trading & Banking Solutions at London Stock Exchange Group

The creation and preservation of wealth are long-term pursuits and inherently involve the proper management of a wide set of risks. Factor modeling now includes machine learning, algorithms, and new inputs from a very deep lake of alternative data. Perspective and experience will always matter, and *Quantitative Asset Management* is an excellent source for today's practitioner facing new challenges and opportunities as part of their investment due diligence.

—William J. Kelly, CAIA
Chief Executive Officer at Chartered Alternative Investment Analyst (CAIA) Association

QUANTITATIVE ASSET MANAGEMENT

Factor Investing and Machine Learning for Institutional Investing

MICHAEL ROBBINS

Mc Graw Hill

New York Chicago San Francisco Athens London Madrid
Mexico City Milan New Delhi Singapore Sydney Toronto

ISBN: 978-1-26-425845-1
MHID: 1-26-425845-3

The material in this eBook also appears in the print version of this title: ISBN: 978-1-26-425844-4, MHID: 1-26-425844-5.

eBook conversion by codeMantra
Version 1.0

Library of Congress Cataloging-in-Publication Data

Names: Robbins, Michael (Professor of finance), author.
Title: Quantitative asset management : factor investing and machine learning for institutional investing / Michael Robbins.
Description: New York : McGraw Hill Education, [2023] | Includes bibliographical references and index.
Identifiers: LCCN 2023003736 (print) | LCCN 2023003737 (ebook) | ISBN 9781264258444 (hardback) | ISBN 9781264258451 (ebook)
Subjects: LCSH: Institutional investments—Mathematical models. | Portfolio management—Mathematical models.
Classification: LCC HG4515.2 .R63 2023 (print) | LCC HG4515.2 (ebook) | DDC 332.6—dc23/eng/20230127
LC record available at https://lccn.loc.gov/2023003736
LC ebook record available at https://lccn.loc.gov/2023003737

CONTENTS

PART I

PLANNING OUR WORK

The Equation of Investing

Not long ago, an intelligent and well-educated bank executive took me aside to a room with financial news blaring and a stock ticker scrolling across a large television screen. He drew an economics equation on a whiteboard and said, "I understand economics; this is the equation for inflation. What's the equation for investing?"

This book is written for professional investors. Despite excellent educations, most investors have learned a highly specialized skill on the job, and many lack perspective outside of their niche. These investors may focus on a particular type of investment and sometimes neglect the nearly limitless opportunities in other markets. Other investors may see generalities and can miss practical nuances. Academically minded researchers who hone in on the more common factors may forgo the benefits of exploiting "edge effects," frictions, and inefficiencies.

Regardless of how much we know, we all have gaps that we might not recognize. As we expand our ambitions, we can expose crucial deficiencies in our otherwise expert skill set. Like all of us, the chief economist I wrote about underestimated the extent of his ignorance. Without exposure to topics that we think we understand, these errors can cause us to chase down the correct answers to the wrong questions.

Some readers of this book will merely fill in gaps or learn a few tricks. Others may extend their technically formidable investing skills and learn to develop more comprehensive strategies, work effectively in a different environment or culture, weather new challenges, or build a business using skills they had not yet developed fully.

These themes—awareness of other markets, thoughtful planning to ask the right questions and answer them effectively, opportunity through diversity and specialization, and the struggle against uncertainty—are among many crucial lessons woven throughout this book.

■ ■ ■

I, too, began my career as a specialist, focusing exclusively on producing risk-adjusted returns using quantitative techniques without regard for more nuanced business goals. As I worked for different organizations and in more corporate roles, I understood how the motivations and drivers of these purported investment businesses deviated from, and sometimes conflicted with, the pursuit of investment excellence. Navigating these obstacles can mean the difference between a successful investment effort and one that is frustrated and stymied by influences unrelated to investment performance.

The business lessons that I learned as I developed from investor to portfolio manager to chief investment officer form Part I of the book. To be effective, we must reframe the straightforward concept of earning investment returns into a suitable product for the business's goals, culture, resources, and clients (Chapter 1).

As investors, we must secure the resources we need with conviction and purpose. We need to inspire confidence when others are nervous and doubting. Defining and aligning our strategy with a business is discussed in Chapter 2. To make this argument persuasively, we need to be explicit about our business case and investment policy (Chapter 3). We need to outline what we offer, what we are going to do, and how we will do it. We do this because it is often required and almost always a good idea, but mainly because methodically thinking through the plan helps us build firm footing for uncertain times.

Part II begins our discussion about quantitative financial modeling. Like any similar text, we discuss important topics like overfitting, mitigating unrealistic assumptions, managing substitutions, enhancing minority classes, and missing data imputation. We do not dawdle on the basics or emphasize technology or theory. We focus on practical investment solutions and do not shy away from detail and complexity. Our scientific and academic discussions are often motivated by fundamental and economic rationales, but pragmatism dominates. Technical drivers like investment vehicle mechanics or order flow can be more reliable and pertinent reasons for investment decisions than compelling theories.

Investing requires far more than merely predicting which investments will rise and fall. In Part III, we develop our strategy into an investment product, including the alpha models, risk models, implementation, backtesting, and cost optimization. Parts II and III are not merely about quantitative methods. Several important themes are pervasive:

- **The planning discussed in Part I is crucial.** Quantitative investment management is largely project management. Many experienced managers fail for a lack of planning, especially those who ask the wrong questions and solve the wrong problems. There is solace in trusting a well-designed plan during difficult times.
- **Identifying risk is easier than forecasting returns.** A good investor can make money with a bad investment, but a poor trader may lose with likely bets. "Wrong way" risk occurs at the worst times and can be catastrophic, while "right way" risk can be more easily tolerated.
- **Interactions are unavoidable.** Factors span asset classes and do a better job isolating the essence of investment performance than traditional methods. Asset classes, sectors, and other classifications interact and can muddle and confuse hidden drivers and risks. Because of interactions, the combination of assets in a portfolio can be more important than selecting which investments to include.
- **Complexity may be worth the effort.** Complexity does not have to be paralyzing or even confusing. Taken step by step, layer by layer, tractable elements can be combined with significant effect.
- **Systems mitigate biases.** One of the significant advantages of quantitative methods is that they can avoid dangerous biases. If used properly, the tendencies of others can be used for advantage in systematic trading and feature design.

Part IV describes the management of an ongoing investment business: measuring performance, learning from mistakes, managing risk, and surviving tragedy. We wrote our battle plan and built our fortifications in Part I, and in Part IV, we discuss how to soldier on when things do not go as planned. We balance the need for objectivity with the limitations of an inherently flawed system and the political realities of multiple stakeholders.

Box 0-1 Endnotes and Other Resources Are Online!

There are extensive endnotes for this text. We would have liked to have included inline references to all the brilliant research that made this book possible, but this would have made the book twice as long!

Endnotes, deeper details, examples, and other resources including computer code, videos, lesson plans, and quizzes are available at www.QuantitativeAssetManagement.com. We strongly encourage that you visit and use these resources frequently. They will be augmented and updated regularly.

1

Choosing Our Product

Fit for Purpose

It is common to think that the technical skills of investing—predicting profitable investments and managing money—are transferable. Without experience, it may seem that investing one's own capital, working for a hedge fund, working for an advisor, or managing a treasury are essentially the same task. They are not. The business goals, available resources, client needs, constraints, and incentives differ and may be more important than good performance. An investor needs to understand and make the most of his situation.

This chapter will examine a few typical businesses and how to manage investments effectively within them. To start our plan on the right foot, we need to assess some essential concepts:

- **Firm type.** The ultimate goal of the firm may not be returns alone. It is critical for our investment efforts to support the business's goals. These goals are often identified by firm type. It may be surprising that many investment firms care far more about narratives, sales, or maximizing fees than returns. The firm's source of capital may have particular needs or preferences. It may not be possible to dictate the character of our returns; we may have to invest to suit our benefactors.
- **Skill.** Our resources may be limited. We may direct a SEAL team of extremely capable experts or an army of advisors trained as salesmen. Our goals must reflect our capabilities. We must pick a fight we can win.
- **Fund structure.** We may inherit a fund structure or design a new fund. Either way, it is important to know the various strengths and

weaknesses in different jurisdictions. Legal and tax technicalities often dictate investment style.

- **Incentives.** Incentives define success for those around you. Returns may not be the primary goal of the firm or for those above you. Similarly, you should align the incentives you provide to your employees with your goals to avoid unintended outcomes.

This chapter discusses these concepts, which frame our efforts. We will build our strategy in Chapter 2 and develop our plan in Chapter 3.

Firm Type

The business model of the firm will dictate the firm's resources, client bases, and motivations. The goals and purpose of an investment effort should be aligned with these needs and capabilities. Following are some examples of company types and how they reward investment managers:

Proprietary trading firms. These are often considered the most capitalistic form of investment company, with the purest alignment between the firm, staff, and clients. Proprietary trading firms have "skin in the game" because they provide leverage. To help align traders with that risk, proprietary trading firms, or *prop houses*, often demand a symmetrical payout scheme in which managers participate in both losses and gains. Perverse incentives are also common, such as encouraging overtrading so that commissions flow from the investment team to the firm. The investment team often pays for resources.

Hedge funds and private equity. On the surface, the *incentive fees*[1] of these funds seem well aligned with investment goals. As a fund grows, a more significant share of revenue is often derived from the *management fee*, which is proportional to money under management, and less from their performance or risk management. An emphasis on the management fee can cause the business to focus more on marketing than performance. Sometimes funds exceed capacity constraints and accept more clients than can be accommodated by the strategy. Resources at hedge funds may be generous.

Wealth management advisories. Advisors are not immune to conflicts and may be enticed to recommend products that produce higher commissions or prioritize capital accumulation (*assets under management,* or AUM). To do so, they may emphasize services like life planning or tax management, client experience, differentiation, speed to market, “democratization,” or some other noninvestment characteristic at the expense of robust investment returns.

Clients may have conflicting goals, as in the case of a trust that may reward one beneficiary with income and another with appreciation. Or a current client may want to deplete a fund to the detriment of future generations or against the wishes of the original benefactor. Business models emphasizing noninvestment goals may view investment resources as burdensome, and investment teams may lack the skill necessary for challenging tasks.

Family offices. Family offices are diverse, and so are their goals and investments. Their purpose and operation depend significantly on the offices’ size and management. Family offices may behave like advisories, proprietary trading firms, or endowments. They may be more institutional and accept niche investment strategies, co-investments, and club deals. Conflicts may include kowtowing to the family or incompatible goals for different members and generations.

Investment banks, private banks, and insurance companies. These firms often produce sophisticated products and services, including bespoke structures. Financial engineering is often employed to design and tailor solutions to client needs. These firms are usually well resourced with data, technology, and skill.

Pensions, governments, endowments, charities, sovereign wealth funds, and corporate plans. These tend to be highly institutionalized. Some are well resourced and highly sophisticated, while others are financially constrained and managed by boards composed of laypeople representing the beneficiaries or sponsoring institutions. Some large public institutions are loathe to pay high salaries but have lavish budgets for outside consultants and services. The conflicts of outsourcing critical management skills are challenging.

Investment Skill

A man who built an enviable investment company based on low-cost passive investing once asked me how to reconcile respect for active management with his success. Conflicting styles and ideas can coexist under different conditions and constraints. A firm with less access to talent and opportunity should pursue a more passive, lower cost, and more diversified strategy, and a firm with greater access to talent should be more active, pay for access and speed, and concentrate on high-conviction ideas.

Passive investors can try to minimize the impact of many decisions, but some choices, like the timing of transactions and rebalancing, are unavoidable. Even mechanical, arbitrary timing rules are choices with consequences. Skillful investors spend risk as a valuable and effective resource rather than reduce and avoid risk as if it were a hazard.[2]

Fund Structure

Some readers who focus exclusively on managing investments may not yet have been exposed to fund structures. They may manage proprietary money or aggregated accounts like a treasury or a pension. When these readers work in a more complex arrangement, like a Registered Investment Advisor (RIA), or start a new fund, choosing the proper legal and taxation structure will be critical.

Some clients demand separately managed accounts (SMA), mandates, segregated accounts (“segs”), or side pockets. These accounts provide the client with customized treatment and contain actual securities rather than shares in a commingled fund. The accounts can add or exclude investments to suit. When they do, they begin to “fork” from the model portfolio or fund as soon as they are created and diverge more as time passes. Because the accounts are not shared, they are not held to the same regulations as the fund they were modeled after.

Funds combine investments in a structure shared by more than one investor. Funds enjoy economies of scale, liquidity, and reduced costs. Complicated or expensive treatments that may be much easier in a fund might be impossible in segs. Tax wrappers, derivative use, and other features may be better employed in a fund.

Nearly every jurisdiction has its form of funds with accompanying legal and tax treatments and commensurate restrictions. Funds come in nuanced varieties and have names including managed funds, investment pools, and collective investment schemes (CIS). They may be publicly traded, as in the case of exchange-traded funds (ETF), mutual funds, unit investment trusts (UIT), real estate investment trusts (REIT), and other closed-end funds, or special-purpose acquisition companies (SPAC). They may be private, as in limited partnerships that are popular with hedge funds or private equity.

Subtle differences can be crucial. For instance, mutual funds are incentivized to sell their most liquid assets to pay large redemptions, leaving less attractive assets in the fund, while ETFs can remove the least liquid assets, retaining better holdings.

Box 1-1 American Funds

American funds that invest in securities need to be registered as RIAs. Even foreign managers can register as an American RIA. All managers that transact securities must be registered someplace to do business with US citizens, regardless of their home country's regulations. It is common for US funds to create an offshore blocker fund to shelter tax-exempt investors (like foundations and endowments) from unrelated business taxable income (UBTI) liability. Master-feeder structures are popular in which investors are often serviced by onshore or offshore feeder funds that invest in a master fund. A significant drawback is that US dividends are taxed at 30 percent in offshore funds.

Products sold by firms are often either *open architecture*, allowing clients to choose products managed outside the firm, or *fettered*, limiting clients to in-house products. Open firms offer greater variety and may seem more impartial. They can be plagued by preferences and arrangements that incentivize them to provide subpar or expensive products and burden clients with multiple layers of fees.

In-house offerings allow a firm to exercise more control and influence, potentially providing better and more aligned results. They may enjoy more transparency for better performance attribution and risk analysis and accommodate more flexible transfers, like *payment-in-kind* (PIK) and segregated accounts.

Subadvisors manage funds on behalf of another asset manager and may include *white-labeling* (renaming the fund in the asset manager's name) for broader access to clients. Sometimes a retail product is marketed as "institutional quality" when the subadvisor caters to institutional clients.

Box 1-2 ETFs

ETFs generally have lower fees, are more tax-efficient than mutual funds, and have attractive characteristics relative to vehicles like UITs (including dividend reinvestment schemes that favor downtrends). There are many ETFs, making it easy to express views. For instance, equally weighted ETFs hold more small-cap stocks than cap-weighted funds and reduce performance chasing.

Many of these specialized ETFs do not deliver what's "on the tin"—notably, commodity ETFs, leveraged funds, and inverse funds. Contrary factor funds, like value and growth funds, often share holdings. Geographically constrained funds may receive income and have expenses from locations other than those advertised. Fixed-income ETFs often have a stable duration, while fixed-income assets generally have decreasing duration. Real estate ETFs generally do not contain residential property exposure, even though residential mortgage products are a large part of that market. Commodities ETFs often invest in a mix of futures rather than spot assets. They may have complex tax structures and may be taxed on unrealized gains or as collectibles. They may even require inconvenient K-1 filings.[3]

Incentives

Incentives are critical in product design, the investment process, and effective team management. They are a primary driver of performance and risk (including operational, legal, and regulatory risk) and are the important part of Part I and Chapter 15.

A fund manager's most compelling task is keeping his job and being compensated. We must meticulously analyze and impose incentives that motivate desired behavior in ourselves and those we hire. We can apply game theory, Monte Carlo simulation, and stochastic calculus to understand and negotiate fairer fee structures, reduce moral hazards and agency problems, and encourage healthy risk-taking.

We can simulate how a rational investment manager would behave by modeling our product. We combine the fund's performance with the various fee and cost schedules, rational behavior, and heuristic biases to help us evaluate product configurations, provide income for the manager, and provide risk-adjusted performance for the investors. Improperly motivated managers may maximize fees, take excessive risk, or be too cautious (e.g., waiting for a compensation reset date, tracking, hugging, or closet indexing).

There are many variations in structures and parameters. Most funds extract a periodic management and incentive fee. As the fund gathers assets, a manager may be tempted to concentrate on marketing to raise capital and earn a sizable management fee rather than provide outstanding performance. Hurdles, high-water marks, and clawbacks may limit performance fees, but resets may negate these limits.[4]

Box 1-3 Case Study: An Incentives Alignment Model

Modeling the effect of incentives is helpful when negotiating with fund managers or designing products to align incentives with desired results. The compensation date (e.g., the end of a bonus period) is an essential feature of the incentive model. Figure 1-1A shows how a manager might be more inclined to decrease his risk as he gains or loses money. As the bonus date approaches, he may be more inclined to "double down" when losing (Figure 1-1B). The manager's fear of redemptions or job loss moderates his risk-taking as reflected by the plateau at the extreme left of Figure 1-1B. Compensation incentives can exacerbate poor performance rather than improve it by encouraging greedy risk aversion. The profit curve (dotted line) skews to the left when a bonus is near, and losses accumulate.

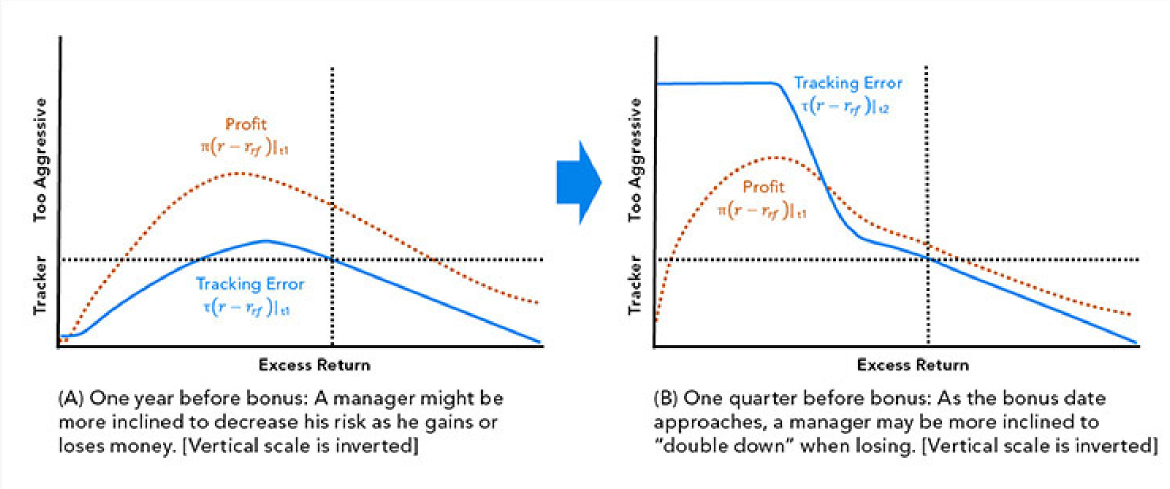


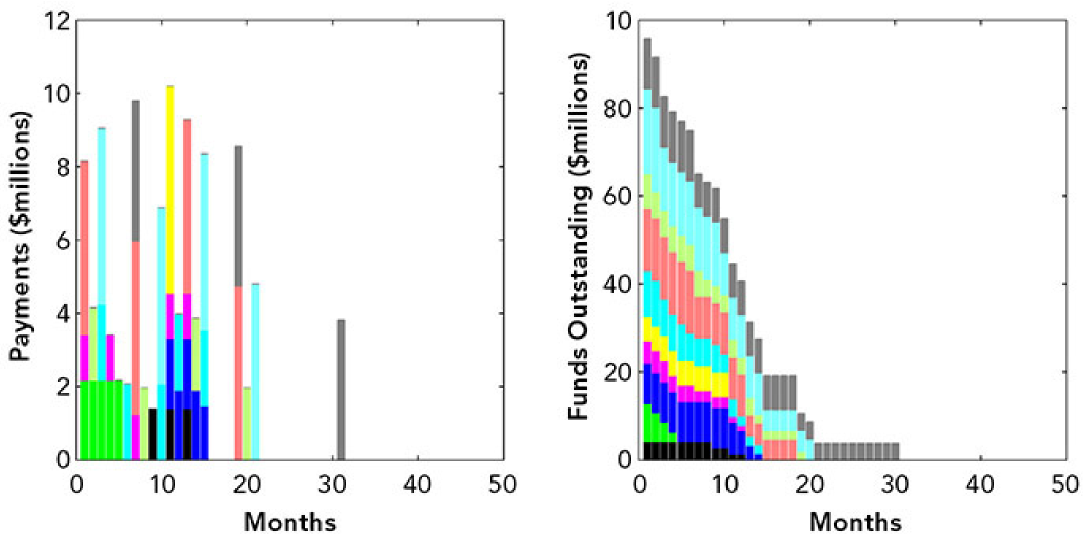
FIGURE 1-1 The results of a simulation of investor behavior in response to incentives. Greedy, risk-averse investment managers are too meek at the beginning of an investment period (A) they may be too aggressive when performing poorly at the end of an investment period when the managers gamble in the hopes of a reward as crystallization time approaches (B). The vertical dotted line represents breakeven.

Restrictions, like frequencies, *notice periods*, *gates*, *lock-ups*, and penalties,[5] are designed to restrain investors from recovering their assets. They exist to delay or moderate the forced transaction of illiquid holdings and protect the "sticky" investors who remain from "fast money" or "hot money," and to provide breathing room for managers to recover performance. The downside of these restrictions is that investors may be trapped in a depreciating fund with little recourse. One way to minimize or remove this burden is by using side letters.

Box 1-4 Case Study: What Are the Hidden Costs of Gates, Lock-ups, and Settlement Periods?

A large family office asked us to quantify the effect of their managers' incentives so the funds could be compared and the terms negotiated. We found restrictions can reduce performance by as much as 7 percent annually. We calculated the *liquidity value adjustment* (LVA) by answering, “How much can the fund lose after we decide to redeem?”

We examined a portfolio of 10 funds with multiple redemption requests resulting in 21 tranches. Each tranche represented an individual investment in a fund. The recovery of cash flows was spread out over time based on each fund’s restrictions. The left panel of Figure 1-2 shows how the redemption payments from a portfolio can be staggered because of these numerous gates, lock-ups, and settlement periods. Each shade represents a different fund. The right panel in Figure 1-2 illustrates how the funds are returned to the investor over a period of years, reducing exposure.



(A) Future cash flows (payments) are estimated. The different colors (shades) represent different investments.

(B) Each of the payments from panel (A) reduce the exposure (called funds outstanding).

FIGURE 1-2 Future cash flows and called funds outstanding

We simulated many possible realizations for each tranche and redemption and determined how much would be gained or lost due to the restrictions. Initially, we modeled the investor’s liabilities to determine when the funds were likely to be called for reasons other than performance. The benefits of this analysis are twofold. First, it is imperative to differentiate *right-way risk* (a tolerable accident) from *wrong-way risk* (a problem that manifests itself at the worst possible time, such as when pension disbursements increase). Secondly, knowing when the firm needs to raise capital for payments and how that might coincide with losses is helpful.

The funds' valuations were modeled using forward-looking factor analyses, performance histories (if available), or proxy portfolios. Scenario and stress tests generated return distributions for statistics like *maximum shortfall* and *value-at-risk* (VaR).

Redemption rules and liquidity restrictions were imposed on the dates determined by the liability model to compute when the investor would receive the investment cash flows over time, and in what quantities.

We then calculated *potential future value* (PFV) and *potential future exposure* (PFE) by combining the redemption rules with the simulated paths and averaging the results, as shown in Figure 1-3.

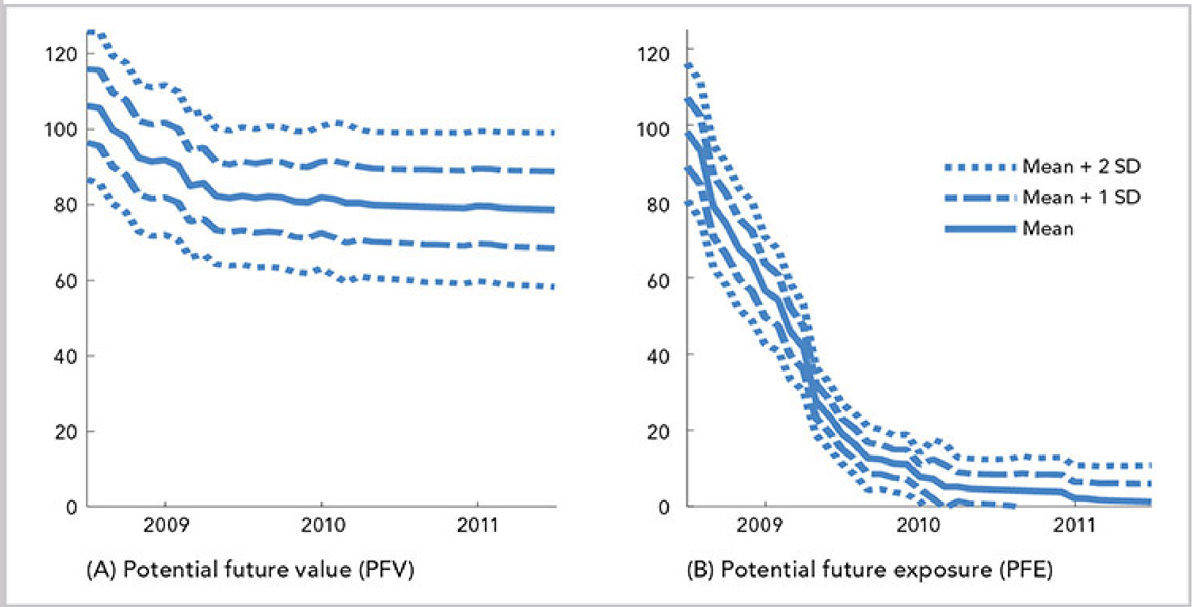


FIGURE 1-3 Potential future value (PFV) and potential future exposure (PFE)

We then calculated the loss distributions, *maximum likelihood estimator* (MLE), and VaR (Figure 1-4). We determined the monetary impact of those restrictions by comparing the simulated profits and losses with the restrictions applied to those simulated without restrictions.

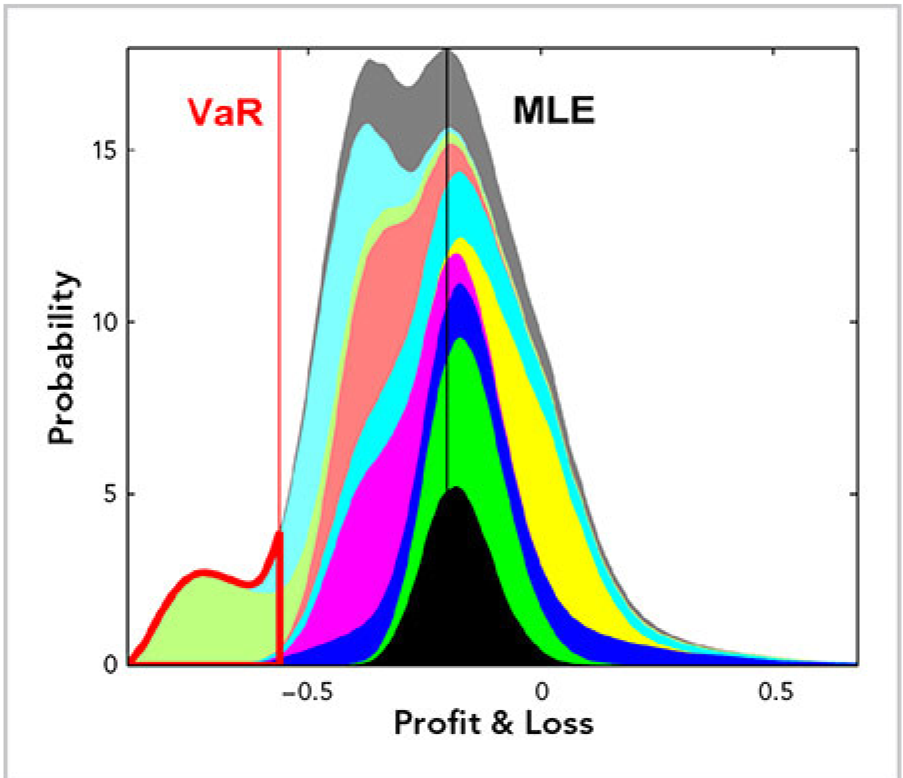


FIGURE 1-4 Distribution of losses for all portfolios, tranches, and simulations

By multiplying the expected loss, VaR, or other risk statistics by the likelihoods of the events, we determined a charge or hurdle necessary to compensate for the risk of loss due to the redemption restrictions. Different scenarios and allocations were analyzed.

We combined assets and liabilities dynamically in a single portfolio to manage the asset/liability ratio volatility and applied probabilistic and actuarial models to estimate the needs and wants of our clients' constituents.

By performing this exercise, we were able to price these risks and incorporate them into fund valuations. This made valuations more fair and allowed us to compare funds with different structures.

■ ■ ■

Investment managers rarely have the luxury of working in a vacuum. They must consider the environment they work in, including their firm, resources, and culture, as well as the products that deliver the returns to investors.

---

1. Incentive fees (also called *performance fees*) are paid by investors and are proportional to the profit generated by the manager, subject to conditions. Management fees are paid in proportion to the amount invested and may not be reduced by poor performance. The canonical “two-and-twenty” refers to a management fee of 2 percent of assets per year and a performance fee of 20 percent of profits.

2. Investment skill relates to both the ability of the investor and the potential for skill to affect the investment outcome. Many prefer passive investments because they believe no level of skill can produce risk-adjusted performance after fees. Grinold and Kahn’s Fundamental Law of Active Management, $E(R) = TC \times IC \times \sqrt{BR} \times \sigma$ roughly translates to “expected return is proportional to efficiency, skill, and the number of poorly correlated opportunities.” If an investor does not believe he has sufficient skill, he can still improve his efficiency and his universe with diligence, technology, and access to opportunities.

3. United States Internal Revenue Service Form 1065, Schedule K-1 is used to report partnership financial information.

4. A manager must first earn a minimum *hurdle rate* before accruing a performance fee. The hurdle may be an absolute number or may be measured relative to a *reference rate*. *High-water marks* prevent a manager from accruing a performance fee while the fund underperforms. *Clawbacks* allow investors to recover fees. Periodically, a high-water mark is reset to the current portfolio value and prior losses are excluded from future calculations.

5. There are many and varied provisions to protect a fund from *outflows*. Redemptions may be restricted to certain periods, e.g., 20 percent quarterly, and may require prior notice, e.g., 30 days. Redemptions may be limited, or *gated*, e.g., a 20 percent per redemption period. Redemptions may be excluded for periods of time (*hard* lock-up) or subject to redemption penalties (*soft* lock-up).

2

The Investment Process

How to Invest

Now that we have determined our purpose within the business and how to further its objectives in Chapter 1, we can turn our eyes toward how we plan to achieve those goals. Once we have a plan, we will crystalize and communicate it in Chapter 3. A well-thought-out process provides structure, reliability, and guidance that we can lean on in uncertain times.

Strategic Asset Allocation (SAA)

Some readers will not require a plan as comprehensive as the one presented in this chapter. Those who specialize in a narrower skill, like stock picking, can still benefit from understanding where their efforts fit in the scope of a larger agenda to better fulfill the needs and goals of their company, manager, and clients.

Factor exposures. The strategic or long-term allocation, also called the *policy portfolio*, is usually defined as target allocations for different categories of investments with a band or range to allow the manager flexibility to adjust their risk to the current situation. While most institutions still define the bulk of their allocations by specifying asset classes and monetary weights, both policies are archaic and misguided. Risks span assets, and the risks of different categories frequently vary. Defining types of investments by factors and quantities in terms of *risk budget*[1] is far more helpful and practical.

The benefits of diversification. Some investors resist asset allocation strategy and overweight risky assets. This myopic view ignores long periods of *drawdowns* in nearly every investment type. While “you can’t eat risk-adjusted returns,” drawdowns devastate investors with liabilities that they need to fund and those near retirement.

One of the most common risky asset classes, American large-cap equities, experienced an extended drawdown after the Great Depression of the 1930s and during several periods since then. It is fair to say that large drawdowns are not anomalous and that the 49 years that followed the Great Depression misled many people into thinking stocks reliably appreciate.[2] If the reader concentrates on the shaded drawdowns of Figure 2-1 rather than the long-term trend and imagines being “under water” for 5 or 10 years before retirement, risky assets may appear less attractive. The chart would look significantly more dire if it included World War I and the Spanish Flu.

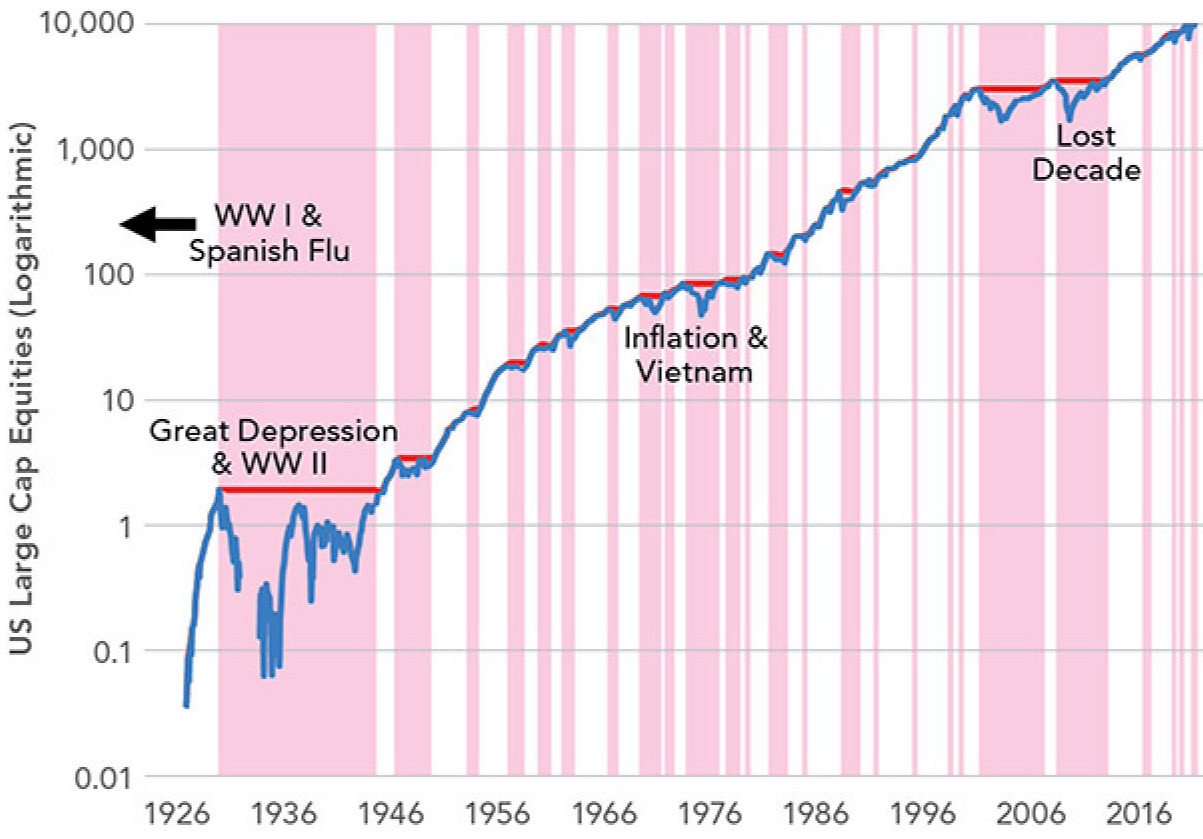


FIGURE 2-1 Equities spent a great deal of time in drawdown. Do not get stuck!

Institutions own about 80 percent of US equities. Many, like underfunded state pensions, would find it challenging to endure extended periods without income. Even if a portfolio can produce a higher level of *terminal wealth*, ensuring more time “above water” may be worth the effort. Figure 2-2 shows that most state pension systems are underfunded, and many of them significantly so.

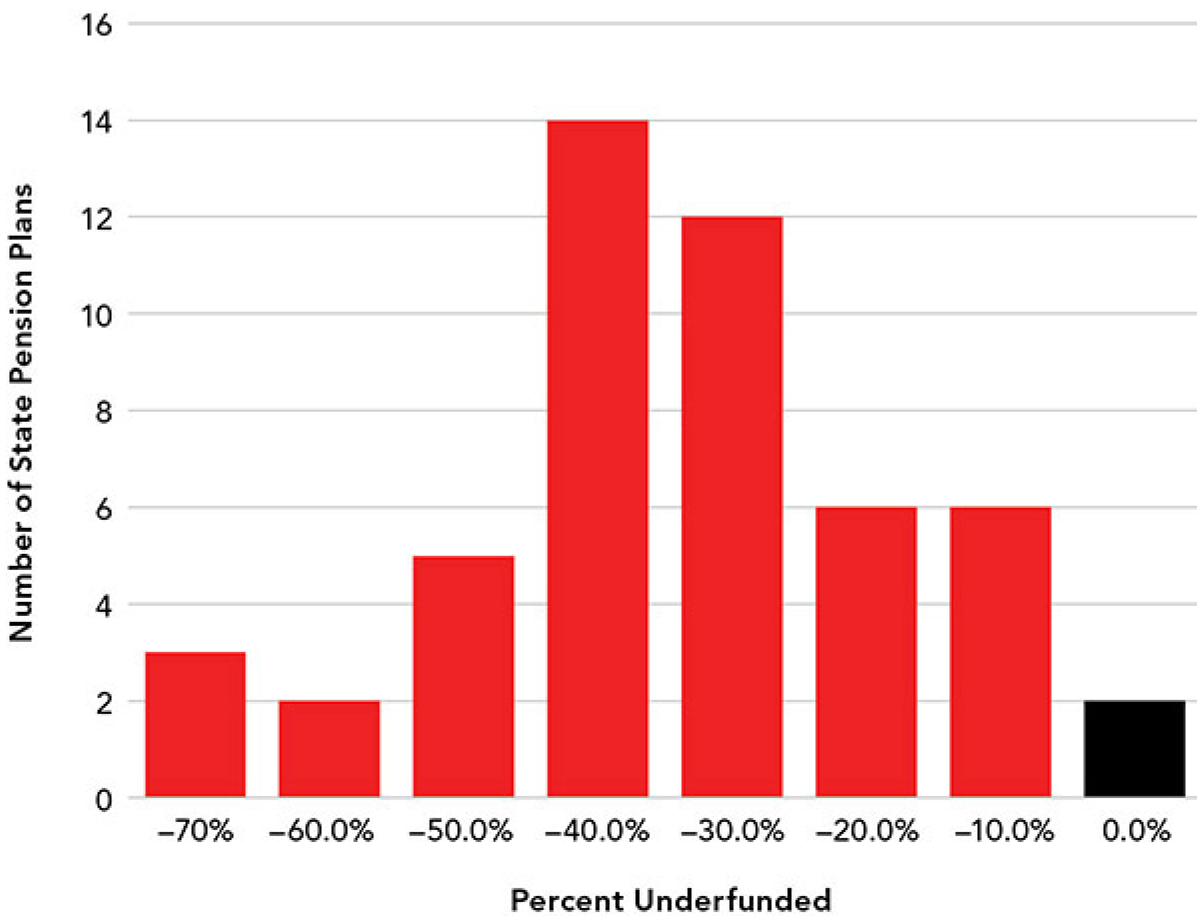


FIGURE 2-2 US state pensions, the funded ratio minus 100%

Data: “The State Pension Funding Gap: 2016: Investment shortfalls, insufficient contributions reduced funded levels for public worker retirement plans,” Pew Charitable Trust, April 12, 2018

Asset allocation vs. security selection. Asset allocation may benefit less from skill than security selection does and is more likely to be the significant driver in performance for multi-asset class investors. Many institutions expend a disproportionate effort choosing managers and individual investments rather than allocating to factors. As Paul Samuelson suggests,[3] security selection often demands more skill because markets tend to be more efficient in selecting securities than in choosing asset classes. One reason large category asset classes are more inefficient than narrower security categories is that asset classes may combine a wider variety of agents, time frames, and motivations, making mispricings more difficult to arbitrage.

Strategic policies inherently enjoy long time horizons. However, shorter periods may pressure investment selection and management decisions.

Corrosive costs and fees can diminish the benefit of increasing trading frequency and reignite the ongoing discussion of active versus passive investment during challenging times. Trading is required to maintain diversification and risk premia. It is critical to address difficult choices head-on rather than demur to a less relevant but more convenient goal.

Tactical Asset Allocation (TAA)

Tactical opportunities are found in trends and dislocations. Expectations implied by economic data can be compared to those implied by *capital markets assumptions*. Inconsistencies and mispricing exist in term structures, volatility surfaces, and other frameworks. Time-varying factors like risk and covariance, exogenous shocks and timely views, cyclicality seasonality, and supply and demand disruptions can all be exploited.

If we compare our forecasts, or *views*, of asset valuation, factor evolution, and market forces (sentiment) to those implied, or *priced into the market*, we can determine *overweights* and *underweights*, or *tilts*. These tilts temporarily skew the investment portfolio from the strategic goal.

Many firms schedule investment meetings to discuss sweeping and slow global macroeconomic trends periodically (e.g., monthly or quarterly) and cross-asset strategy and positioning more frequently (e.g., weekly). Actionable decisions require managers to identify the precise assets to be traded and in what proportions. This *best expression* of the trade may be a straightforward purchase, or it may be some complex synthetic asset constructed from many derivative transactions and spanning tax havens.

Entry and exit levels, trade duration, risk limits, conviction, and other details must be evaluated and estimated as well. An *opportunity matrix* or *threat matrix* (Figure 2-3A) can represent this discussion with quadrants of varying returns per unit of risk versus conviction. Another typical output is a tactical *scenario matrix* (Figure 2-3B) that displays opportunities for various economic realizations. Instead of four quadrants, many cells may be included to accommodate a more detailed set of threats and scenarios.

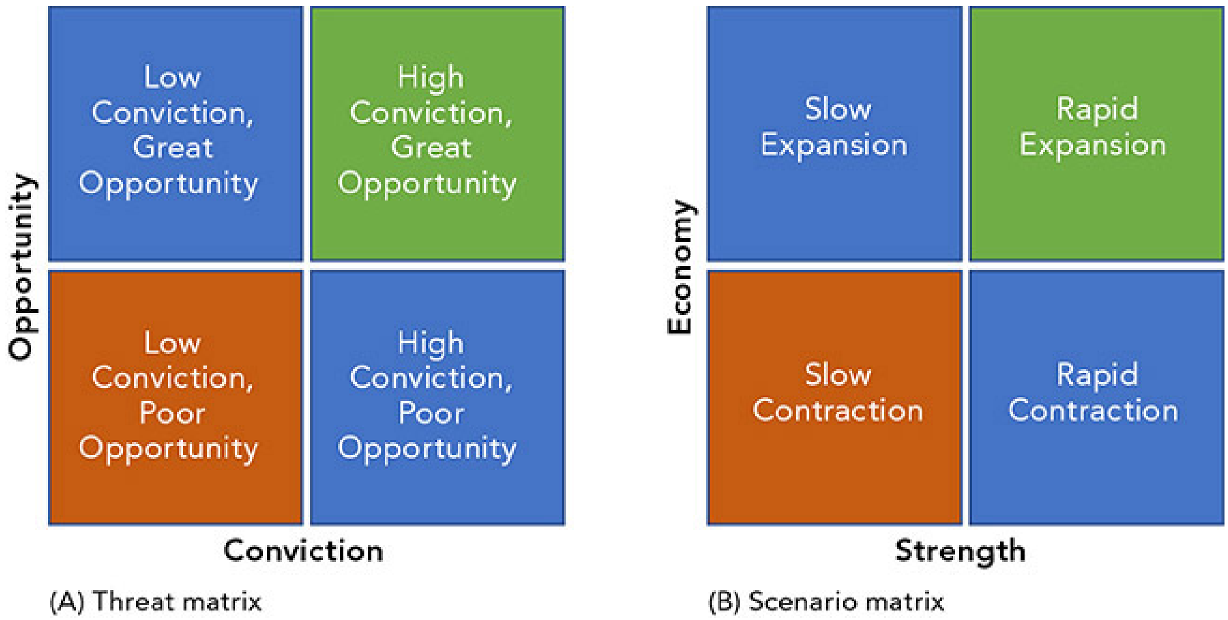


FIGURE 2-3 Matrix representations of threats/scenarios

It is almost irresistible to meddle with the portfolio. The hazards of overtrading are trite, but it is also notable that tactical allocations reduce the long-term diversification of an optimized portfolio unless the short-term views are spot-on. The opportunity cost of less diversification should discount the expected but uncertain benefit of the tilt. Some investors tilt risk allocations in ways that maintain a target risk level for the portfolio or categories of investments.

Factor Investing

There are a great many ways to define and use factors. A key outcome of quantitative methods, including *risk premia* attribution and optimization, is the reverse process of exposing implied assumptions and psychological biases such as short-termism, home bias, and procyclicality (herding and momentum). Antti Ilmanen categorizes investments into asset classes (stocks, sovereigns, credits, alternatives), strategy styles (value, trend, volatility, carry), and risk factors (growth, inflation, illiquidity, tail risk).[4]

A common strategy for identifying and quantifying factors is reverse engineering the prevailing assumptions priced in the market to determine *fair values*. It is even more insightful to reverse engineer our investment decisions or those of managers we employ. Many investors do not objectively analyze their performance and behavior and invest in portfolios that do not support their views and logic. Narrative, persuasion, and confidence may dominate scrutiny, performance, and skill.

Risk premia. A risk premium is merely the marginal reward of an asset beyond that of less risky investment, especially when the risk is procyclical or "wrong way."

Behavioral heuristics show that investors are not entirely rational and that cognitive biases can sustain persistent or *structural* premia. Supply and demand can also create a premium that never "cleans up," such as for a bond with a maturity that is popular for hedging purposes. As *smart beta*[5] products evolve, some investment expertise is reclassified from alpha to premium—"there is no alpha, just undiscovered beta."

Predicting premium levels is notoriously tricky. A forecast's highest purpose is not in the prediction itself but in the perspective and insight gained from the process. Even when forecasted accurately, premia are time-varying and subject to tail risk, which is challenging to capture.

Structural risk premia can be estimated in many ways, including with theoretical models using social, supply, and demand data. Forward-looking estimates, such as the yield curve shape, bootstrapping, and surveys, can be complicated by conflicting prediction horizons and the tendency to focus much more on near-term detail. These forecasts are often asymptotic on longer horizons. Two standard methods of separating premia include stacking and demand-based orthogonal decomposition.

Stacking premia. Different categories are added or layered when stacking premia, like the capital structure distribution waterfall. For example, the equity premium can be separated into the risk-free rate, inflation, bond premia, and equity premia. Each can be estimated separately; added together, they make up the equity risk premium. The bond risk premium reuses some premia and consists of the risk-free rate, inflation, and the bond horizon premium.

Multifactor and orthogonal premia are sometimes estimated mathematically with little regard to economic intuition. This has many advantages, such as flexibility, ease of automation, scalability, and mathematical rigor. Orthogonal premia's lack of interpretability can be a severe drawback, however.

Classic factors, like Fama-French factors,[6] are generally stacked (subadditive) and intuitive. They are well researched, and there is great competition in advancing those methods. Shorter-term structural premia and premia that are more localized are more likely to be exploitable. Differing objectives, time frames, and other goals generate inefficiencies but are not persistent. Some can be anticipated, like the cash-futures basis, while others may be too ephemeral or require too much data to be derived from economic models, like some transaction features. More automated methods, including spectral methods, favor the latter. New and promising structural causal models have the potential to automate the design and operation of interpretable economic models.

"Quantamental" techniques combine economic intuition with more quantitative methods, usually by marrying humans with computer models rather than focusing on one or the other. These approaches often involve thematic factors that have efficacy, interpretability, and narrative appeal for product marketing. This appeal is not cosmetic; it is far more helpful than simply making the strategy easy to explain and sell.

Themes can also better align components of a multi-asset strategy. If the overall portfolio is risk budgeted, allocating to a fund that uses the same methodology produces a more coherent process. Likewise, an economically minded chief investment officer (CIO) may be interested in a fund weighted by gross domestic product (GDP) rather than the market cap. Highly specialized funds, like those that use large quantities of alternative data, can use these methods. For instance, some funds specialize in analyzing regulatory filings or legal documents, and they can find patterns in language that help identify opportunities.

Computers do not work like people, and forcing them to mimic investment styles—even quantitative ones—may not be the most effective way to design algorithms. Nonetheless, effectiveness is often a luxury compared to explainability and interpretability, which are essential to build trust, raise assets, and keep those assets during difficult periods.

Details. Minute details accumulate into serious impediments and can taint factors or make them impotent. A myriad of insidious considerations like fees, overdiversification, sampling error, and cash drag amplify ever-present and immediate concerns like crowding, contagion, and style drift. A good example of style drift can be seen by examining how a seemingly singular fund, Fidelity Magellan, had a tracking error that increased and decreased during alternating management regimes (Figure 2-4). Magellan famously employed Peter Lynch's "buy what you know" strategy, but they subsequently varied their strategy as managers changed, including the "closet indexer" Bob Stansky.

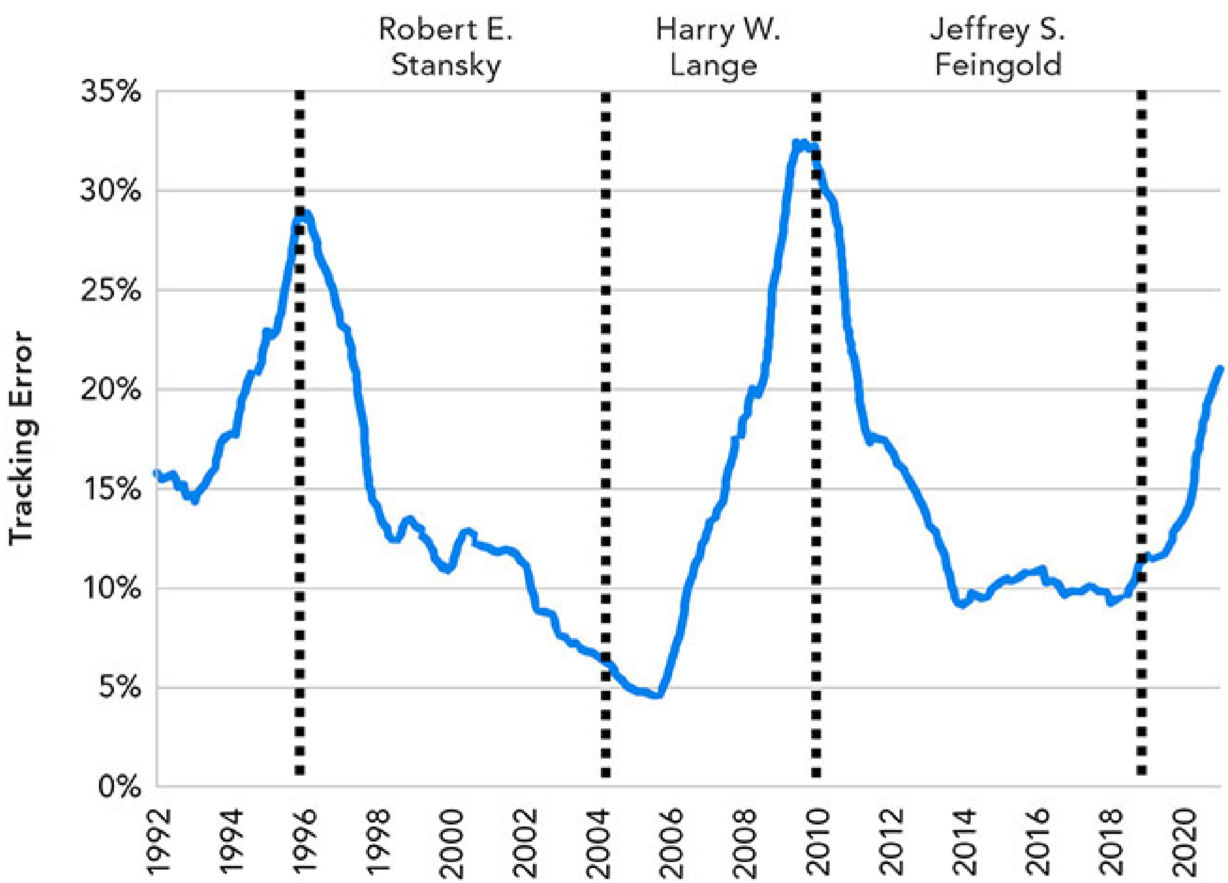


FIGURE 2-3 Two year rolling tracking error of the Fidelity Magellan Fund

Data: Bloomberg, LP

Hubris and oversimplification are the siren song of unwise investors. When choosing factors, building a process, or designing an algorithm, starting with "bare metal" in order to fully understand the assumptions and nuances of the process is often better than modifying existing techniques.

- **Clarity and control.** Nearly every step in the complex development of a quantitative scheme involves expert *qualitative* judgments. We disagree with many software design decisions, if only because so many arbitrary decisions are made. For instance, Bloomberg's PORT function reports daily returns for funds that only provide quarterly 13F[7] reports, assuming (incorrectly) that the portfolios do not trade between reporting periods and merely drift between filings. Although not entirely accurate, this method is a sensible solution, but it is misleading if it tricks us into assuming that the daily returns reported by PORT represent actual returns.
- **Differentiation.** It is challenging to differentiate a quantitative process from a competitor's process in a meaningful way if the competitor used the same third-party tools and data.

Factors that may seem simple are rarely straightforward to execute. For example, a high-level decision tree may start with a choice to overweight stocks or bonds using factors like liquidity, money supply, leading economic indicators, etc., and may become more specific and scrutinize details like the debt-to-cash flow of individual stocks. Factors often interact and compound each other to result in complex formulations like "the 5-year Z-score of forward enterprise value over sales."

The Role of Theory

In many fields, researchers seek a universal "black box" that demands no domain knowledge. Many naive or unseasoned financial researchers fall into this trap. We should always guard against spurious relationships and overfitting. In doing so, it is natural to look to theory (high-bias models) before statistics (high-variance models) for a rational, explainable, and interpretable basis.

Economic models. There are many ingenious economic models for studying supply and demand, business cycles, monetary and fiscal policy, trade and flows, and much more. It is tempting to start with these models and let the capital market assumptions flow from them. A great deal of this is valid, for instance, when identifying the phase of the business cycle. Economics serves a different purpose than quantitative trading models. Economic models are not for predicting asset values but for understanding economies' complex and dynamic interactions. In this sense, economic models can be more important than quantitative trading models that fit the data better. Economic models make Herculean and unrealistic simplifications to try to determine fundamental relationships. Simple policy rules are better than clouding tenuous intuition with dubious detail.

It is easy for the rigor and precision of a theory to lull us into trusting its results as primary signals, but the world is far too complicated for that, and economics has a long way to go before we can trade based on these theories. It is best to use economic theories to condition our views rather than to drive them. There is a lively debate regarding the value of complex mathematics at this stage of economic theory development. Economic models can be intractable due to conditioning state variables and multiple conflicting cycles. Supply models can describe asset prices, and demand models can estimate individuals' and companies' consumption through credit growth and the willingness to exploit it. For all this effort, asset prices are not well correlated with the business cycle and can only explain, at most, two-thirds of the variation.

Econometric models are far less ambitious and more statistical than theoretical. They often impose a high-bias academic relationship and help predict capital markets assumptions. They can be used to produce the scenario matrix quadrants when it is "better to be roughly right than precisely wrong."[8]

Quantitative and market models are used to describe the behavior of asset prices and market participants. They include mechanical models (such as futures arbitrage), statistical models (such as pairs trades), and agent models (such as market impact models). These models are distinctive because they do not seek to identify asset values but rather the optimal transient state or the end state of a deterministic event, like the convergence of a futures expiration or market equilibrium on a concise time scale. These models often resemble engineering methods more than scientific or statistical formulations.

■ ■ ■

Economic, econometric, factor, and quantitative research is highly competitive. Looking to relatively untapped sources of information using alternative data and cutting-edge techniques is rewarding. We are in a golden age where these data and techniques are accessible and powerful yet rapidly improving. Older statistical methods are prone to spurious results, but newer techniques, including machine learning methods, aggressively and explicitly address overfitting and other dangers. Techniques like structural causal models may soon overcome the most critical objections of modern, complex financial machine learning methods by producing dynamic, robust, interpretable, self-healing, and fault-tolerant predictions.

Security Selection

Security selection is the most domain-specific and diverse step in the investment process. Some investors focus entirely on asset allocation and invest in efficient ETFs; others find particular and nuanced opportunities in esoteric instruments. Each method deserves an entire book of its own. A simple high-level example of how factors can translate into investment choices appears in Figure 2-5, but this only scratches the surface.

Fixed Income	Cash	Inflation	Duration	Credit
Cash	✓			
TIPS	✓	✓	✓	
Treasuries	✓		✓	
Corporates	✓		✓	✓
HY	✓		✓	✓
EM	✓		✓	✓
IG	✓		✓	✓
Infrastructure	✓	✓	✓	✓

Equity	Dividend Yields	Earnings Growth	P/E	FX
Futures		✓	✓	
Indexes	✓	✓	✓	
Sector Indexes	✓	✓	✓	
DM	✓	✓	✓	✓
EM	✓	✓	✓	✓

FIGURE 2-5 How factors translate into investment choices

Most funds focus on security selection for many reasons, including the breadth of opportunity available in various instruments and expressions, such as the capital structure, derivatives, and risk proxies. *Plain vanilla* options combinations offer many methods to gain exposure to a particular investment, including conversions, reversals, boxes, and put-call parity. Synthetic constructions and arbitrages are commonplace. Nuances like tax treatments may produce more profit than the original thesis and may be more structural and reliable.

Portfolio Construction

Portfolio construction determines the proportions of the investments, the timing, and the strategy involved in buying and selling investments, the ongoing upkeep of those proportions (*rebalancing*), and hedging techniques. Engineering-style methods of portfolio construction have been mainstream for a long time. Nonetheless, *the real promise of scientific allocation has never been completely fulfilled.* Most managers describe their portfolio construction in the context of these quantitative methods, although the actual execution is sometimes handled less rigorously, depending on the manager's temperament.

For the most part, modern weighting techniques implicitly or explicitly involve some optimization method. The objectives and constraints vary dramatically and can be forecasted or calculated parametrically, historically, or by Monte Carlo simulation. Specific implementations like multi-armed bandits and Kelly criterion approaches can be intriguing, but no particular system is dominant, and all have their flaws. Stochastic optimization and dynamic, robust, and multi-period methods have all shown promise.

Timing. Many investors decry the futility of timing, but timing is unavoidable. Every investment decision involves timing, and guilt by omission is not a solution. As with overfitting and interpretability, it is better to face the imperfect and challenging task than ignore it. Even systematic solutions like opportunistic rebalancing are implicit bets on timing. As crucial as entry and exit targets are, their value is more apparent in the discipline they engender than the actual levels they enforce. A good investor is likely to profit from a poor trade, but an undisciplined investor often loses despite good ideas. Transition management, shortfall management, and control over other efficiencies can account for most structural profits in many investment businesses, except for high-margin endeavors like mergers and acquisitions.

Rebalancing techniques are an essential aspect of portfolio management, complicated by the trade-offs between skill, costs, and path dependency. At the edges, it is evident that high-frequency traders with strong information coefficients benefit from extremely fast rebalancing, while low-skill indexers would benefit from more passive approaches.

As a rule, turnover becomes constrained as fund size increases, as shown in Figure 2-6. When a fund is small, few pay attention to its transactions, but as it accumulates significant shares in the assets it trades, it moves markets, making rebalancing expensive. Judicial use of flows can move the portfolio closer to the target allocations, reducing tracking error and alpha decay.

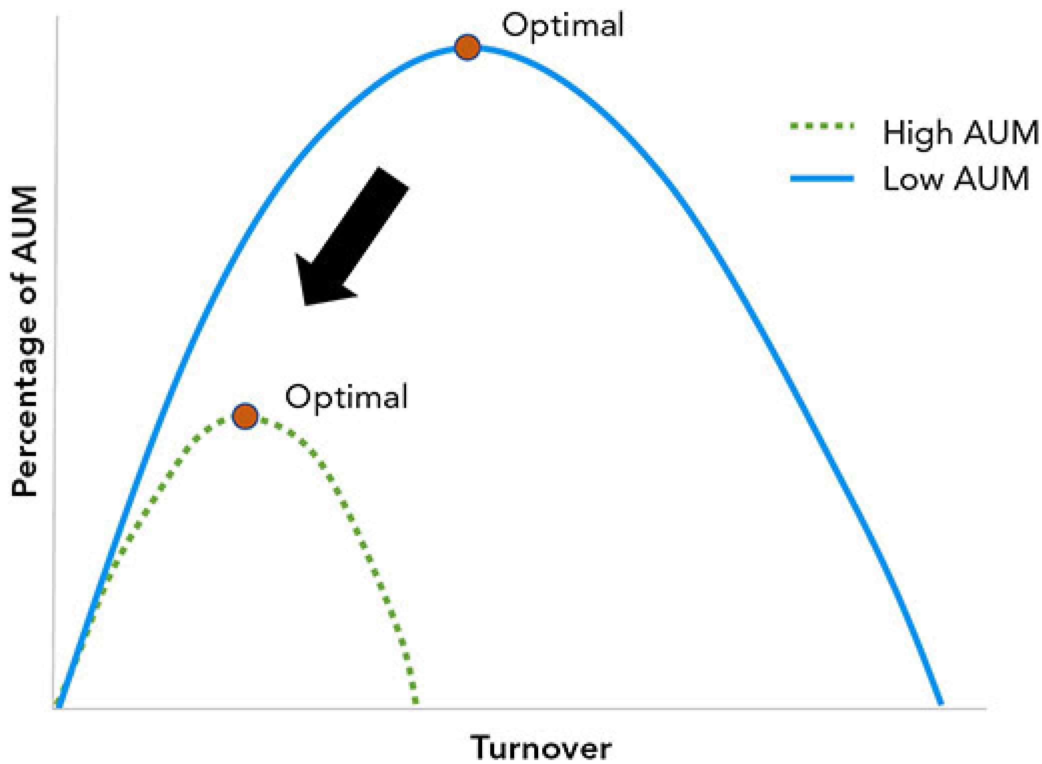


FIGURE 2-6 Optimal turnover decreases as firm size increases.

Adaptive asset allocation sidesteps the market timing aspects of rebalancing. It avoids frequent significant contrarian adjustments (selling high and buying low) by rebalancing relative to the market portfolio as it drifts. The logic is that the market portfolio reflects the changes in expected returns as it drifts. These expectations can be exposed through reverse optimization.

A criticism of rebalancing is that it produces a structural "drag on returns" due to a tendency to react too slowly or miss a trend. While there may be some truth to this, a falling market with increasing volatility tends to be *mean-reverting*, as do sudden movements in equities. Multi-asset portfolios are less likely to cause regret, and both trend-following and mean-reversion strategies can be helpful. Focusing on cross-sectional returns may be more fruitful than regime exploits.

Systematic hedging strategies, like stop-losses and overlays, may also be designed to provide adequate risk management. Risk control is essential to the longevity of most strategies. Most investors are wrong most of the time but benefit from their good decisions much more than they suffer from the bad ones. Ongoing management is critical, but systematic hedging is a tailwind.

Like many strategies involving mortgages, some hedges are subject to systematic losses in a kind of adverse hedge, while others risk significant tail losses ("picking up dimes in front of bulldozers"). Many businesses are built on hard work and operational risk ("sweat equity") rather than skill and luck. They are compensated for accepting and managing structural inconveniences like negative convexity.

Negative biases may be an unwelcome feature of a strategy, but often protection is available at a cost. As a rule, the convenience of insurance incurs a cost that exceeds the economic value of the protection. It is advisable to purchase protection when the cost of loss is unbearable, even if it is cheaper to self-insure. An overpriced "lottery ticket" has value if the price is minimal and the payoff is meaningful. Dealers diversify risks by matching trades and earning nearly riskless income by providing insurance—though it does not mean they will offer it cheaply.

Ongoing Management

The ongoing monitoring and management of performance and risk and the general cash management of a fund are in many ways more important than its thesis. Managing involves probing and testing for flawed or invalidated assumptions, new relationships, and sensitivities, including exogenous risks, tail events, operational issues, and myriad seemingly innocuous details that could doom a fund. Adherence to policy and governance can serve as guardrails against these risks, and often failures are clustered around minor deviations that may become an Achilles' heel, causing a cascade of losses.

We explore many ways of decomposing risk and return, attributing drivers and sources, and viewing these metrics from different perspectives to identify threats while they are still actionable. Perturbation, stress testing, scenario analysis, counterfactuals, and other methods are all elements of the risk management toolkit. With all these tools, many think risk measurement is a science or a form of reporting. Risk management is an art that requires significant experience and insight into many conflicting clues and the dangerous assumptions needed to make analysis tractable.

■ ■ ■

We have outlined a general investment process as an overview for a deep dive in Parts II, III, and IV. It is easy to lose track of the process when focused on the intricate details, and this chapter was designed to frame how it all fits together. Beginning with a long-term view—the strategic goal—we deviate tactically to take advantage of opportunities. We use factor investing rather than asset class categories, and risk budgeting rather than monetary weights, because they are more closely aligned and true to our purpose. We explore many kinds of models and quantitative strategies that can add insight. Security selection and portfolio construction translate our ideas into purchases and sales, executed with care.

Finally, investing is rarely a "set it and forget it" activity. It requires constant management and diligence—often in direct opposition to our instincts and biases.

---

1. A risk budget measures allocations in terms of risk rather than a percentage of *assets under management* (AUM).

2. Edward F. McQuarrie, "Stock Market Charts You Never Saw," September 2017.

3. Paul Samuelson, "Summing up on Business Cycles: Opening Address," in Jeffrey C. Fuhrer and Scott Schuh, eds., *Beyond Shocks: What Causes Business Cycles* (Federal Reserve Bank of Boston, 1998).

4. Antti Ilmanen, *Expected Returns on Major Asset Classes*, John Wiley & Sons, 2011.

5. Willis Towers Watson, "Smart beta: sometimes smart, sometimes not," 2006.

6. Eugene F. Fama and Kenneth R. French, "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics* 33 (1993) 3–56; Eugene F. Fama and Kenneth R. French, "A Five-Factor Asset Pricing Model," *Journal of Financial Economics* 116, no. 1 (April 2015): 1–22.

7. United States Securities and Exchange Commission Form 13F, "Reports Filed by Institutional Investment Managers."

8. Often attributed to John Maynard Keynes but more likely originated in Carveth Read, *Logic: Deductive and Inductive*, 1898, or earlier.

3

Leadership and Governance

Be Accountable

Three common lethal investment errors are hubris, ignorance, and flexible morals. Nearly all the thousands of professional and academic investment schemes that I have supervised would have failed due to these blunders had I not intervened.

These selfsame vices are also a well-worn path to success, often to the detriment of clients and creditors. For every dramatic lucky success, there are many silent failures. Consequently, *availability bias* leads to *confirmation bias*. I preach the virtue of discipline, probability, and skill over impetuousness, speculation, and luck—grind over flash. These virtues increase the likelihood of success and the targeted and deliberate impact of that success.

Safeguarding good behavior, alignment with investors, guidance for ubiquitous moral hazards, and legal compliance are not the only reasons for prudent governance and effective leadership. Even the best-intentioned plans frequently lack a path to sustainable success or inadvertently create organizational difficulties without a convenient remedy.

We begin this chapter by discussing good investment *culture*. These traits are essential to individual investors, small funds, and large organizations. They are frequently insincerely professed as core beliefs, and their absence corrodes the investment process from the inside out.

I have watched many struggle and fail unnecessarily for lack of planning. A *business case* aids awareness and prepares for the uncertain and challenging process of building an investment business. The benefit of writing a case is in the insight it brings, not in the plan itself, not as an oracle, but in the ability that it affords its creator to act knowledgeably, confidently, decisively, and wisely to challenges. The benefit as a document for others is a distant second. Tailor the message for specific audiences when elements of the case are used in other documents.

Most enterprises do not require a formal and extensive *Investment Policy Statement* (IPS) to define the relationship between an investment manager and his investors. Readers unfamiliar with these documents may benefit from the many details the documents consider. Just like the business case, an IPS can help us better understand our target so we can aim directly for it rather than adjust painfully midway.

The *Investment Due Diligence* (IDD) and *Operational Due Diligence* (ODD) documents ensure that the business is built well, is running properly, is suitable, and is being applied earnestly and without deviation. These documents are often required by larger investors and can force us to hold ourselves accountable to willfully ignorant investors. Accountability is a good thing; it keeps us on the path.

Finally, we discuss some considerations and constraints, including the size and culture of an organization and its investment appetite.

Culture

Good governance of investment decisions—training employees to feel, know, and understand the proper course of action—often seems burdensome, especially in small companies formed to circumvent the layers of dissonance that plague so many big firms.

In small teams, it is easy to forget that the separation of governance and execution is fundamental to accountability, performance, and risk management. In a business infused with emotion, uncertainty, stress, and temptation, leaders are responsible for their teams' investment decisions. Good governance is most effective in an environment that nurtures and supports responsibility and accountability. A leader may outsource tasks but not the responsibility of leadership and governance.

Paradoxically, cultures that encourage experimentation often lead to careless decision-making when more competence is required. With the trust to fail and learn comes the increased responsibility to use risk wisely and with the consideration of consequences.

Responsibility, character, skepticism, delayed gratification, self-efficacy, and self-honesty are vital in quantitative investing. Each of us may have differing beliefs on how effectively we control our circumstances, a concept called the “locus of control.”[1] In an environment requiring strength and action, many abdicate responsibility by shying away from the activity, paralyzed by the magnitude of the task or for fear of mistakes. Others shrug off responsibility and simply rely on luck—“it wasn’t my fault” or “it was meant to be.” See Figure 3-1.

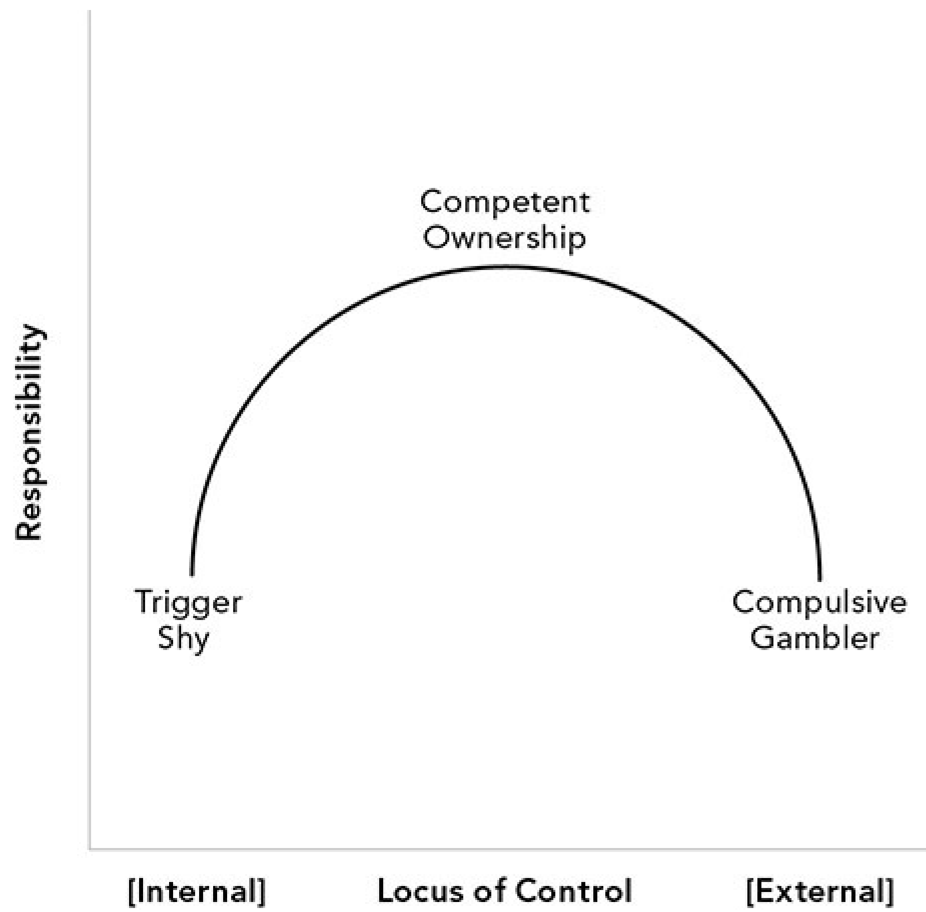


FIGURE 3-1 The Arc of Responsibility

Reliable accomplishment requires ownership and resilience. Not everyone is 100 percent present, resilient, and self-efficacious all the time. Good culture guards against inevitable weakness and the many mental frailties and biases we all possess by forcing us to at least go through the motions of a reasoned decision. What is required is the fortitude to grind *away at a problem and then execute an imperfect solution within the window of opportunity while taking responsibility for the results.*

Most people do not scientifically scrutinize their beliefs but instead substitute the feeling of belief as a proxy for truth. When confronted with the possibility that their view has not been vetted with the proper skepticism, they often react rather than analyze. Investing is much too difficult a task for such luxuries. Behavioral biases can be extremely detrimental to critical decision-making. Correct processes help enforce the scientific method and the necessary checks and balances when it is tempting to cut corners and let the noise distort the signal.

Nihilism ("trigger shy" on the left of the arc in Figure 3-1) is as powerful and destructive of an organizational force as overconfidence ("compulsive gambler" on the right). It is popular to write about how poorly professional investors and analysts perform. It is true that, after generating and extracting hefty compounded fees, the average professional investor does not beat their benchmark. While investing is difficult, underperformance is often largely a result of inadequate self-mastery and discipline.

If we do not think we can beat our competitors, we should not be competing. Even if our strategy is a "race to zero [fees],"[2] we can be better and more innovative than our competition. A sense of purpose and efficacy are essential and need to be communicated and reinforced.

There are many ways to outperform, and many of them hinge on small advantages like low costs, solid infrastructure, or access to information. A sales pitch is expected to ignore weaknesses and obscure competitors' strengths, but self-deception is dangerous. Scientific scrutiny and psychological resilience are essential in quantitative asset management, and cultural self-deception is commonplace.

Rather than seek a goal, it is critical to plan and follow a process, separate functions for good governance to manage misalignments and decentralization, and work diligently to stay true to best intentions. Creating these documents is the first step toward good planning and management.

Business Case

Often a new investment effort will elicit questions from stakeholders:

- Why is this effort important to the business?
- What will the execution be like?
- Is this course of action the best way to get the result? What are the alternatives?
- What do you need from me? What will be my share of the task?

A business case differs from a business plan. It does not include an analysis of the competition or the competitive environment, nor is it intended to provide long-range projections of revenues, expenses, business strategies, and so on.

Worked examples can make an abstract discussion clear. I strongly recommend visiting www.QuantitativeAssetManagement.com for an example of a business case for an investment business.

Preparing an argument. A business case is an excellent preparation for meetings with boards, committees, and managers who approve funding. Regardless of whether a formal business case is required, writing one will result in a well-thought-out and compelling rationale and a concrete plan. Depending on the organization and culture, the business case may ultimately be a short presentation, memo, or a thorough report complete with venerable planning techniques like program evaluation review technique (PERT) diagrams, Gantt charts, and the PRINCE2 (PRojects IN Controlled Environments) or critical path methods. The document may be permitted to gloss over knowledge commonly known within the firm. The case is designed for internal consumption though outside experts may review it.

Solve problems. The goal of a business case is always to solve problems. As a general rule, focus on outcomes, not deliverables. Start with the vision and goals, draw a flow chart, define the scope, determine the necessary capabilities and features, build a road map, and create the team.

Be deliberate and thorough. The following highly structured guidelines represent just one of many ways to write a business case. The structure ensures that all elements are addressed methodically. The fundamental goal is to use a format that helps us think through our plan to prepare for probing questions.

Fortify for conflict. Frequently, questions asked about the case are not just critical but adversarial. It is likely that funding will be taken from some other project and that another project leader will defend his purpose against yours. The key to building a solid case is avoiding the speculation and biases that confound our thinking. Should criticism occur without a chance to rebut, the document should stand independently.

Address the audience. Some organizations prefer free-form suggestions, while other institutions require a strict, rigid, and detailed format, such as for military-industrial projects. Figure 3-2 illustrates the main sections of this case: the Executive Summary, Introduction, Opportunity Statement, Solution Analysis, Program Definition, and Implementation Increments.

Audience: EXECUTIVE COMMITTEE

EXECUTIVE SUMMARY
Opportunity Statement
Vision
Rough Cost Estimate

INTRODUCTION
Purpose
Background

Audience: INVESTMENT COMMITTEE

SOLUTION ANALYSIS
Analysis of Alternatives
Preferred Solution
Program Outcomes
Risks and Risk Mitigation
Critical Success Factors

- High-Level Outcomes
- Business Outcomes
- Increment Capability Delivery Outcomes

Assumptions

Audience: OPERATING COMMITTEE

PROGRAM DEFINITION
Concept of Operations

- The Investment Process
- Implementation

Financial Analysis

- Financial Benefit Summary
- Economic Analysis and Life-Cycle Cost
- Sensitivity Analysis
- Funding Profile

FIGURE 3-2 Elements of a highly structured business case

In preparing our plan, it is always crucial to consider our audience. After all, if no one sees or hears it, our work has limited impact. Too much information is worse than not enough and may overwhelm readers and cause them to avoid a discussion entirely. It is better to leave them hungry for more detail than to overwhelm them with too much.

Preparation guards against regret. Once we have socialized our plan, we can guard against being put on the spot by preparing supplementary material. It demonstrates competence and diligence and prevents floundering when a presenter is nervous or pressured. Overpreparation can build confidence but can also reduce on-the-spot creativity.

For the Executive Committee

The following documents are of value to the executive committee:

Executive Summary. An executive summary describes the opportunity, the vision, and a rough estimate of the costs we expect to incur throughout our project or business. The executive summary is merely an “elevator speech” for a potential sponsor or champion.

Introduction. The introduction should clearly state the purpose of the effort and the background necessary to understand the context. While intended readers may be fully aware and understand the necessity and evolution of our plan, they may share the document with others, such as consulting firms, to get an outside assessment of the plan.

Opportunity Statement. The opportunity statement consists of two scenarios. The first part, the “as-is” analysis, explains the difficulties and limitations of thriving within the current environment. The second, the “to-be” analysis, explains how to unlock the firm’s full potential to achieve the goal.

"As-is" analysis. The "as-is" analysis explains the firm's current capabilities, the opportunities available if those resources and assets are exploited, and the constraints preventing the company from achieving its potential. Remember that factors and processes that may be obvious to us may be unknown to the reader, even those within the firm. The reasons behind these obstacles may be more nuanced and delicate than we realize. Be sensitive and write more to help and elevate than to raze and rebuild.

"To-be" analysis. The "to-be" analysis should describe several critical aspects of the plan:

- The services and products that can be added as a result of the new effort
- A high-level description of the expected outcomes, including criteria that measure and monitor progress
- How the current business process needs to be reengineered, the beneficial and detrimental consequences of the new methods, and how results will be measured
- The recommended course of action and a rough estimate of the costs involved.

For the Investment Committee

The following analysis is directed to the investment committee:

Solution Analysis. Lay out the plan, beginning with a realistic analysis of alternative solutions. Delineate the preferred solution, including program outcomes, risks, and mitigation measures. Include critical success factors, high-level results, business outcomes, and increment capability delivery outcomes. Be sure to conclude the solution with a list of assumptions. Assumptions and alternatives add credibility and increase the chance that you will have considered criticism thoroughly.

For the Operating Committee

The following documents are intended for the operating committee and others:

Program Definition. Describe how the plan works for the operating committee. They are interested in the costs involved and how they evolve. The program definition has two parts: the Concept of Operations and the Financial Analysis.

Concept of Operations. Delineates a functional description of the process and the implementation, how to build from the current capability through maturity.

The Investment Process. The investment process may include:

- Creating a *thesis* through modeling and research, including the analysis of market, macroeconomic, and fundamental data, and outside research.
- Confirming the *predictive power and repeatability* of the methods identified in the thesis.
- *Forecasting* future conditions as probability distributions.
- Constructing a *model portfolio* by combining the expectations and confidence of the forecasts with the risk and return associated with those outcomes as a multidimensional probability distribution. When combined with simulations, including stress tests, these estimates inform asset allocation choices and security selection. Existing products may be enhanced or new products created.
- *Implementing* the portfolio by transitioning from the current investments to the desired ones efficiently and cost-effectively.
- *Managing the investment risk*, conducting stress tests, analyzing risk and return potential, and monitoring and improving the research process.

Implementation. Explain the basics of implementation in a paragraph for high-level decision-makers, not the operating committee. A more thorough implementation discussion follows.

Financial Analysis. The financial analysis of the project build-out and future maintenance should be comprehensive and conservative. The reader should not expect any surprises due to an incomplete or inaccurate assessment.

Because of the difficulty of setting an accurate budget, we should compare our forecast to a similar project through *reference class forecasting*.[3] Do not forget to account for survivorship bias. The financial analysis may have four parts:

- **Financial Benefit Summary.** Make the financial benefit summary of the plan clear and compelling. Do not use this document to sell the project so much as show the outcomes under various scenarios.
- **Economic Analysis and Life-Cycle Cost.** This section should include initial and ongoing costs throughout the life cycle of the endeavor.
- **Sensitivity Analysis.** Evaluate and quantify the inputs of our economic analysis and the effect of changes in those inputs. For instance, if the cost of borrowing capital to fund the project increases by 2 percent, what effect would that have on the funding profile and expense schedule? Would there be ways to mitigate the impact? Funding rates can change drastically in short periods and alter a project's viability.
- **Funding Profile.** The funding profile lays out the project's financial requirements through time and is often tied to contingent milestones. We discuss the milestones in detail in the following section and provide dates and related expenses. If different funding sources are expected, such as loans and capital raises, or if currency or date mismatches may come into play, they should also be discussed here.

Implementation Increments. Sometimes project components may be built concurrently, but generally, projects are implemented in stages. This section spells out how each increment increases the capacity of the project and the firm.

Increment Capability Delivery Options. Each milestone produces a significant new capability in a *capability roadway*. This capabilities-based approach assesses the needs of each increment and makes it easier to determine and allocate required resources, monitor progress accurately, and realize outcomes predictably and reliably.

Investment Policy Statement

The Investment Policy Statement (IPS) is a contract between the investor and the investment management team that describes the goals, restrictions, and methods employed in managing the portfolio. It is a road map with guardrails that also provides continuity when staff changes. Frequently, an IPS is short, especially in advisories—sometimes less than a single typed page. For social portfolios, restrictions can be extensive and include many pools of assets (operating and perpetual, restricted and unrestricted, donor intent, etc.).

It is instructive to explore the framework of a complete institutional IPS that is more comprehensive than most (Box 3-1). We can choose what we include from this thorough example and discard what is unnecessary. As with the business case, the framework helps us contemplate essential details, even if we do not communicate the details.

Box 3-1 Key Elements of an Institutional Investment Policy Statement

The key elements of a well-conceived institutional plan include the following:

Executive Summary. Begin with the Overview and Executive Summary. Few people read an IPS like a novel. It is a reference and an agreement between those managing the portfolio and the people they are managing it for.

Start by giving a high-level overview of the document, describing the major sections. Next comes the Executive Summary, which should include:

- Introduction and background
- Objectives
- Authority and key restrictions
- Asset allocation
- Measurement and reporting
- Tracking error

Each item should be limited to no more than a sentence or two. The asset allocation should include a summary table of the significant investment categories and their weightings and ranges. Often, goals may be complex and conflicting.

The objective may read something like this:

> The Fund's objectives are to (a) control risk and (b) achieve a long-term rate of return that exceeds (i) the assumed rate of return adopted by the Investment Committee (IC), (ii) inflation plus a stated percentage, as required by each investment portfolio, individually, and (ii) the Fund's Policy Benchmark. The Fund is subject to applicable law including the local and sovereign law.

Fund and Portfolio Design. Fund and Portfolio Design follows with an Introduction, Purpose and Design, Role of the Investment Committee, staff, advisors and consultants, and a brief description of the various investment committees.

In most cases, trustees retain responsibility and can delegate specific functions. This section exists to introduce the governance structure. Sometimes many related funds are included in the same IPS. Larger organizations may have separate committees for each fund's investments, governance, management, risk, and implementation. There may also be an advisory committee for special funds, such as an Islamic finance fund that requires an expert advisor. An example of text from an investment committee description may read:

> *The Investment Committee (IC) reviews, considers, and authorizes proposed investments and external manager engagements as required by this Policy. Additionally, the IC manages the currency hedge ratios and reviews as needed. A letter is required if the IC intends a purchase or sale outside of stated limits or an allocation is intended to an external manager for the first time . . .*

Objectives and Investment Standard. The Investment Standard provides the level of care, and the Objectives detail the risk and return goals and limits. Targets are often related to the organization’s spend rate (with or without inflation and fees), a relative return to a benchmark, a cash plus or inflation-plus bogey,[4] a peer ranking, or a combination.

Depending on the sophistication and concern placed on the document, details like risk-adjusted returns, uncorrelated returns, and asymmetric returns might be discussed, but that is not common. Risk limits may include preservation of capital, maximum loss or drawdown, volatility, and liquidity, among other items.

Asset Mix and Benchmark. The Asset Mix and Benchmark is a well-read and often referenced page on which the strategic asset allocation and ranges are displayed and updated as needed. It may also include the permissions and procedures for stretching the ranges temporarily without amending the document. Calculation methods for items such as illiquid investments, derivatives, leverage, and currency adjustments may also be detailed. The design, weighting, and flexibility to modify benchmarks should also be discussed. That would include rebalancing frequency and whether the benchmark reflects an investable target. Custom benchmarks are not uncommon. A distinction may be made between strategic and tactical weights.

Measurement and Reporting Criteria. Measurement and Reporting Criteria can enforce best practices and consistency across teams and assign responsibility and a timetable for the process. The details are more interesting around the fringes, e.g., statistically significant samples, acceptable sources of data and reliability, distribution assumptions, mixed periodicity, appraisals, liquidity, leverage, and so on. We do not have to be precise, but mentioning these things ensures that they do not get left out of the analysis.

Highly systematic funds or contracts may require specific contingencies for precise and extensive details like data feed delays and failures, exchange closings and limits, and the like.

Private Investments and Overlays. Private Investments and Overlays are different from exchange-traded instruments. They should be discussed explicitly, including how the portfolios are constructed, objectives, authorizations (authority for the initiation, additions, commitments, termination, etc.), how diversification is viewed, restrictions, and conflicts.

Authorized Use of Derivatives. If the fund is allowed to use derivatives, describe the policy and scope of the program, what is permitted and what is not permitted, who is authorized to use derivatives, authorized use by external managers and private funds, documentation and controls, and limitations (counterparty risks, global risks, position limits, risk management, compliance, etc.).

Risk Management and Oversight. Risk Management and Oversight should be extensive, covering such areas as market risk, foreign exchange, credit, liquidity, external funding and its limitations, operations (including overdrafts, custody, and settlement), legal, compliance (including passive and active violations, cure periods, and remedies), and leverage.

In turn, market risk may discuss allocation and risk limits, proxy securities and indices, private holdings, and active risk limits. Credit may address counterparty exposure for over-the-counter (OTC) derivatives, repurchase agreements, and securities lending.

Liquidity. This section may detail when funds may be dispersed voluntarily, use of derivatives, external managers, and unfunded capital commitments. And the leverage section may provide guidance on short sales, foreign exchange hedging, risk parity, embedded leverage, collateralized funding, and leverage in general.

Appendices. An extensive appendix may include such components as an investment integrity program restricting employee investments (an important thing!), external manager selection framework, technical details (such as hedge ratios), general authority resolutions (who can do what), definitions, and data sources.

Review. The document should also recommend how often the policy should be reviewed and by whom.

Firm-Specific Considerations

Institutions. Some institutions have explicit liabilities. In the United States, defined benefit (DB)[5] pension plans must fund their pension benefit obligation (PBO), as calculated by actuaries. Actuarial estimates are subject to assumptions that can be manipulated. For instance, a high expected return overestimates the probability of meeting funding requirements, a high interest rate assumption underestimates the present value of obligations, and demographics estimates (mortality and illness) may extend the funding horizon by overestimating how many participants will be paying in and underestimating how many will be drawing benefits. An insolvent sponsor, restructuring, or risk transfer may dramatically shorten the investment horizon.

Endowments and foundations. Endowments are much like foundations and may use intricate formulas that depend on trailing averages and inflation to determine their spending rate. In the United States, private foundations must spend at least 5 percent of their assets. Similar entities often use this as a spending guideline—even if they are not required to do so. Donor constraints may be severe and complicated. Taxes are usually not a significant concern, greatly expanding the assets considered for investment.

Insurance. Life insurance companies are highly regulated and must maintain adequate capital to pay their obligations, especially their death benefit. A minimum required rate of return may be targeted with a *liability-driven investment* (LDI) scheme to ensure adequate capital with the excess capital invested in more aggressive assets. These differing goals may be described separately in the IPS for clarity.

Credit risk, reinvestment risk, and liquidity for unexpected payments or loans are concerns for cash flow management. They may be explicitly discussed in this section of the document or, more appropriately, with the discussion regarding constraints.

Private clients. While institutions have well-defined (and often regulated) objectives, the objectives of private clients are usually determined through consultation. The goals and preferences of individual investors can vary widely, but they all must consider *longevity* risk and *mortality* risk (the risk of outliving their income or underspending), which can be solved with a defined benefit pension plan or annuities (but at a cost). Private clients frequently overestimate their need for liquidity (incurring cash drag) and tend to be overly concerned with taxes, causing the clients to miss income opportunities in the pursuit of minimizing costs. *Mental accounting*, a common behavioral bias among private clients, leads to *compartmentalization* ("playing with house money") and can be exacerbated by *goal-based investing*.[6] All of these concerns belong in this section of the IPS.

Advisors often deviate from model portfolios when adjusting them for minimization or avoidance of taxes, encumbrances like large illiquid positions, and other constraints. Without a disciplined scientific process, most deviations cause unrewarded tracking error. Common problems include:

- Trading too frequently through a misplaced perception of skill
- Trading too infrequently due to a fear of taxes or fees
- Holding too many assets, or layers of funds
- Employing more risk than intended, such as keeping the stock/bond allocation constant while investing in higher risk stocks and bonds
- Selling assets to raise capital for purchases without considering the timing of the sale[7]

Thousands of references addressing these and other issues in detail are available at this book's accompanying website (www.QuantitativeAssetManagement.com).

Investment Due Diligence

An outside party conducts an *Investment Due Diligence* (IDD) to verify that a business is safe to invest in and that investment returns may continue under varying circumstances, including a management change. Managers should prepare to fill out a *due diligence questionnaire* (DDQ) and answer a wide range of probing questions. Acceptable responses may require years of preparation, such as a history of management working together, or are expensive to restate, like Global Investment Performance Standards (GIPS) compliance. Additionally, these questions identify best practices and may guide you to build a better business. The following description is not comprehensive but is an example of one way to perform an IDD.

Here is a breakdown of some of the critical elements of an IDD:

Business structure. Due diligence may begin with an in-depth background check before interviewing a manager. A manager should not be surprised if interviewers know details about his office that he thought were private. It is also essential to understand that analysts need to understand details to evaluate a business. They interview managers frequently and have deep knowledge of all investment styles. They do not need to know a manager's "secret sauce," but they need to know how he operates daily and in crises.

Overview. The overview is a cursory outline of some elements of an IDD.

Initial description. A fund analyst should start with a short description of the fund, including attributes like its style, risk and return goals, asset class exposures, and diversification. Indicate what type of investor should be interested in a fund like this based on its goals and risk profile. The analyst should discuss what he likes and does not like about the fund and, of course, the fund's performance, process, and team. The fund's revenue base, philosophy, structure, alignment of interests, liquidity, and reporting practices may be relevant if appropriate.

Strategy and investment process. After the basics, the analyst might want to delve into the strategy and investment process. Provide an overview, describe the portfolio, the investment opportunity, and the research process, and discuss how deals are sourced and decisions made.

Management and team. The management and team often make or break a firm. *Key man risk* and volatile workplaces make entertaining stories but are not attractive for investment. When discussing the management and team, include experience, additions and subtractions, compensation levels and structure, and the amount of money the team has invested in the fund ("skin in the game"), including vesting periods, clawbacks, and dilution.

Portfolio characteristics. Portfolio characteristics may change frequently, so keeping track of their specific holdings at any particular time may be challenging. Describe the portfolio's alpha, beta, and risk profile, including systemic vs. idiosyncratic risk, risk processes and mitigation, leverage and liquidity, valuation methods and reliability, and past performance.

Terms. The terms of the investment vary by share class and are sometimes negotiable. These can be complex and may require specialized legal expertise to interpret. The terms of a share class can make all the difference between a good and a bad investment. An overview should discuss the structure, currency, various classes, and alternative vehicles. The structure should include the general partner and limited partner terms and commitments, the fund's planned life, options to extend, investment period, commitments, expenses, placement fees, final closing, penalties for delayed entry, and side letters.

The continuing costs include drawdown notices, initiation fees, tail-down fees, transaction fees, and operating expenses. The waterfall may have distribution types, the flow, hurdles, catch-ups, carried interest splits, and clawbacks. The legal elements may be involved and nuanced. The real purpose of the language may not be obvious and may be disputed when activated. Some items of legal interest include key person, for-fault divorce, advisory board, limited partner transfers, in-kind distribution, co-investment, reporting, meetings, restrictions on raising and capacity, successor funds, reinvestment, and various covenants. There may also be other items of interest, including leverage and varying diversification limits.

Fundraising. Fundraising deserves a discussion that includes handling capacity constraints and the effect of large shareholders on liquidity.

Appendix. The appendix should include manager biographies, sample investments, quarterly newsletters, and updates, including periodic interviews with the fund's principals and operators.

Special Considerations

There are many details and considerations involved in building an investment process. Some firms have a meta-process, while others rely on experts like lawyers and other service providers. The management team's opinion on basic philosophies like active management and diversification should be clear, especially with some trickier applications like environmental, social, and governance (ESG) investing. The type of firm managing the portfolio is a primary consideration.

ESG Mandates

ESG and impact investing mandates have become increasingly popular. Like many good things, however, there are unintended consequences to ESG strategies that may include less accountability for corporate management due to the vagueness of a dual mandate of stakeholder capitalism rather than the comparatively simple responsibility to shareholder value. Rankings, screenings, exclusions, engagement, and full integrations have not been decisively effective.[8] Most successes are difficult to disentangle from the sector and asset class biases inherent in the investable universe.

Firm Type

As with the product decision, the investment plan requires many adjustments based on firm type, client base, and product. Hedge funds and proprietary trading firms are relatively straightforward and strategy-focused, but wealth management is nuanced.

Advisor goals, philosophies, and narratives are diverse. Many advisers—even large ones—recommend a short list of stock picks, while others lean toward low-cost index funds and exchange-traded funds (ETF), or complex structures and private investments. Many advisers focus on goal-based investing. Wealth clients tend to be more concerned with benchmarks, especially those presented by the media. Tax avoidance, turnover aversion, and other goals like ESG investing, may take on outsized importance beyond their financial benefit and may impose irrational and detrimental constraints. Overemphasis on rebalancing frequency and transaction costs can depress returns.

Clients may demand highly customized solutions at the higher end of the wealth spectrum. Limited portfolio size and an overemphasis on liquidity often limit clients' ability to diversify into esoteric and private investments. Clients may also request separately managed accounts (SMA), restricting their product choices, diversification, and liquidity. Wealthier clients often pay bundled or *wrap fees* or similar fees that may cover a wide variety of services, including total fees, transaction fees, principal and dealing fees, spreads, management, and performance fees, nested fees (fees on fees and funds-of-funds), and more.

Investment banks and insurance companies typically offer a range of tools, skills, access, solutions, and customization. In return for their high fees, they must maintain reserves of regulatory capital. They can easily adjust risk, return, shortfall, liquidity income, appreciation, tax efficiency, horizon, bias, and cost.

Cost is an excellent example of the difficulty faced by the unassuming consumer. Cost is different from price; an expensive product can cost nothing up front but still cut into consumers' returns over the long run; hidden costs may be "costless" but not free. The production and use of these products can play a significant role in how thoughtful governance documents are written.

Pensions, governments, sovereign wealth funds, charities, and corporate plans are largely unconstrained from an asset class perspective but may face constraints from complex legal, regulatory, donor-prescribed, or other areas. They may limit liquidity or credit quality, maximum income or minimum distribution or spending rate, social constraints, and be unable to invest in classes like real estate or commodities (for UCITS—Undertakings for the Collective Investment in Transferable Securities—funds). They may need complex procedures, restrictions, and structures to maintain their tax-exempt or not-for-profit treatment. Some entities are perpetual, while others need to maintain separate income and appreciation flows, finance actuarially determined funding ratios, or maintain asset/liability balances. The larger institutions can usually afford expensive expert help from consultants and often have access to investments unavailable to most investors.

■ ■ ■

Good governance and planning are often overlooked, frequently contributing to the failure of otherwise promising investment firms and programs. Writing a thorough business case and an Investment Policy Statement and preparing for a comprehensive Investment Due Diligence and Operational Due Diligence is far more than a paperwork exercise. Thinking through and writing these documents ensures our plan is well-conceived and formidable.

---

2. Julian B. Rotter, "Generalized Expectancies for Internal Versus External Control of Reinforcement," *Psychological Monographs: General and Applied* 80, no. 1 (1966):1–28.

2. It is common for some financial companies to market low-cost investment strategies. While this may be an appropriate business model for some firms, cutting costs to offer lower-fee services "shrinks the pie" and can result in an industry-wide reduction in margins until weaker companies are forced into mergers or failure.

3. In "How (In)accurate Are Demand Forecasts in Public Works Projects? The Case of Transportation" (2005), Flyvbjerg, Holm, and Buhl describe how to avoid Kahneman and Tversky's *planning fallacy* by using an "outside view" and distributional information.

4. These fixed-income style targets are sometimes applied to equity or *balanced portfolios* because they are attractive to investors seeking to offset liabilities such as *cost of living* (COLA) increases. They are often inappropriate and impractical objectives for the asset mix used and lead to frequent and protracted underperformance.

5. A defined benefit plan agrees to pay the beneficiaries a predetermined amount (which may change in accordance with a set formula), while a *defined contribution* (DC) plan agrees to invest an agreed amount but does not promise any specific outcome.

6. Goal-based investing seeks to fund a future expense or target rather than generating a high risk-adjusted return.

7. A particularly interesting work describes the tendency for advisors to neglect the value of exit decisions in favor of shiny new opportunities. This has been confirmed empirically by the author. See Klakow Akepanidtaworn, Rick Di Mascio, Alex Imas, and Lawrence Schmidt, "Selling Fast and Buying Slow: Heuristics and Trading Performance of Institutional Investors," December 2018.

8. Vitaly Orlov, Stefano Ramelli, and Alexander F. Wagner, "Revealed Beliefs About Responsible Investing: Evidence from Mutual Fund Managers," Swiss Finance Institute Research Paper No. 22-98, February 6, 2023.

PART II

DATA, FEATURES, AND RESPONSE

Laying the Groundwork

In Part I, we specified our product, wrote a business plan outlining the necessary resources and a timeline for our project, produced an Investment Policy Statement (IPS) with clear goals and constraints, and saw what our result would look like if a sophisticated investor were to do due diligence on our final product.

In Part II, "Data, Features, and Response," we will begin laying out our quantitative process, which we will complete in Part III, "Building Our Process."

In Part II, we will learn how to:

- Store, explore, and manipulate the available data
- Assess the wide range of investments available, as sources of feature forecasts and also as ways to express the portfolio of factors
- Analyze and select the optimal features and factors to build our models
- Forecast factors for forward-looking models that reflect views
- Build realistic objective functions to represent our strategies in a tractable way

After managing hundreds of quantitative investment projects, we have seen a few administrative problems recur often, and their solutions are simple. In their enthusiasm, researchers can overlook obvious but essential considerations. The resulting errors may be so fundamental that an expensive and taxing effort may need to be restarted from the beginning. To avoid such problems, it will help to follow these five steps:

- **Plan.** Break the task up into manageable parts and complete each one.
- **Specify the problem correctly.** Many projects fail because they are not thoroughly considered. Once the project has begun, it may become evident that the question was not appropriately posed. A thorough literature search can be helpful to identify the problems and solutions that have been found by others and to offer potential ways forward.
- **Ensure that resources are available.** Projects are frequently started without a thorough survey of available data only to find that there is not enough high-quality data for the analysis or that computing resources are lacking.
- **Keep the calculations as tangible as possible.** Quants frequently get lost in the details causing practitioners to marvel at how such highly educated and intelligent people cannot grasp intuitive facts.
- **Work the plan.** Grind it out, hit milestones, and take ownership. Even if the task is not completed, we will still build something impressive.

4

Asset Types

A (Not So) Quick Tour

In this chapter, we will examine some of the data we can use to make decisions; however, the data can be abstract. Investments don't always behave exactly like our data. We may analyze pure interest rates but purchase bonds, loans, debentures, or another instrument replete with incongruous attributes that may offer opportunities or hazards not considered in the original analysis.

This chapter provides an overview of various investments used to create products. Many investors prefer the most common investment vehicles and may fail to exploit the diversity and depth of well-developed markets. By considering all possibilities, investors can express their views on risk and return more precisely and are more likely to uncover mispricings and dislocations, identify less competitive counterparties, or discover some other edge that adds value to the original thesis—or reduce the frictions, challenges, and complications inherent in implementing that thesis with actual investments.

Factors, or *explanatory variables*, are the data we use to help us understand an *expected effect*, or *response*. While factors can be a key element of quantitative asset management, they may also be seen as an abstraction that could potentially strain our ability to implement our investment thesis if we lose sight of the actual instruments available to execute our plan.

The logical time to evaluate potential investments is when we are defining our factors, to make them relevant to our response, when we select securities (Chapter 13), and when we define our objective and calculate costs (Chapters 14 and 15). There are many ways to build a factor, and some "pure" ways to invest in factors without polluting them, such as by using total return swaps (TRS). But these investments can be inefficient, expensive, and complex.

Generally, we build our portfolios through basic assets like stocks and bonds. Our mandate might force us to focus on assets by specifying asset class targets, bands, and other constraints. By selecting the most appropriate investment combinations, we can choose the most efficient expression of our concept and we can exploit inefficiencies before they are filtered out and diversified away.

The cliché "a stock bought right is half sold" emphasizes that the right choice of instrument, price, and timing are as important as the reason motivating the transaction (as are fees, costs, and taxes). A myriad of different investments can be bought and sold to create the same factor exposure, but it can be challenging to be an expert in all of them.

Categorizing investments into asset classes can be rife with complications. For instance, many actively managed assets are poorly correlated with each other but are often grouped by necessity. Even emerging markets equities or debt issues make for poor groupings since they typically encompass a wide range of geographies that may be different from one another.

Contagion is common among asset classes,[1] and the linkage between them is constantly varying. Categorizing investments is further complicated by the vehicle used, such as common stock, exchange-traded funds (ETFs), swaps, or options. The structure of the product offering is often difficult to disentangle from the asset itself. Two identical bonds with different coupons, such as fixed and floating, are exposed to different risks even though their defining exposure, such as company risk, may be the same. For example, even two floating-rate notes, one with a typical reset period that incentivizes refinancing and another that frequently resets (familiar in Australia) have vastly different exposures to interest rates. Their prepayment risk varies greatly as well. Because of the timely resets, Australian mortgages are relatively insensitive to interest rates, while interest rates are a primary driver of more traditional mortgages elsewhere around the world.

Factors. Although factors and products may get muddled, we can be more precise when we focus strictly on factors because there tends to be less overlap between factor exposures. The left panel of Figure 4-1 demonstrates how investments, such as convertible bonds or *preferreds*, span classes. In this example, preferreds combine both equity and fixed income, currency, and commodity (FICC) exposure. Some products span all classes, such as options.

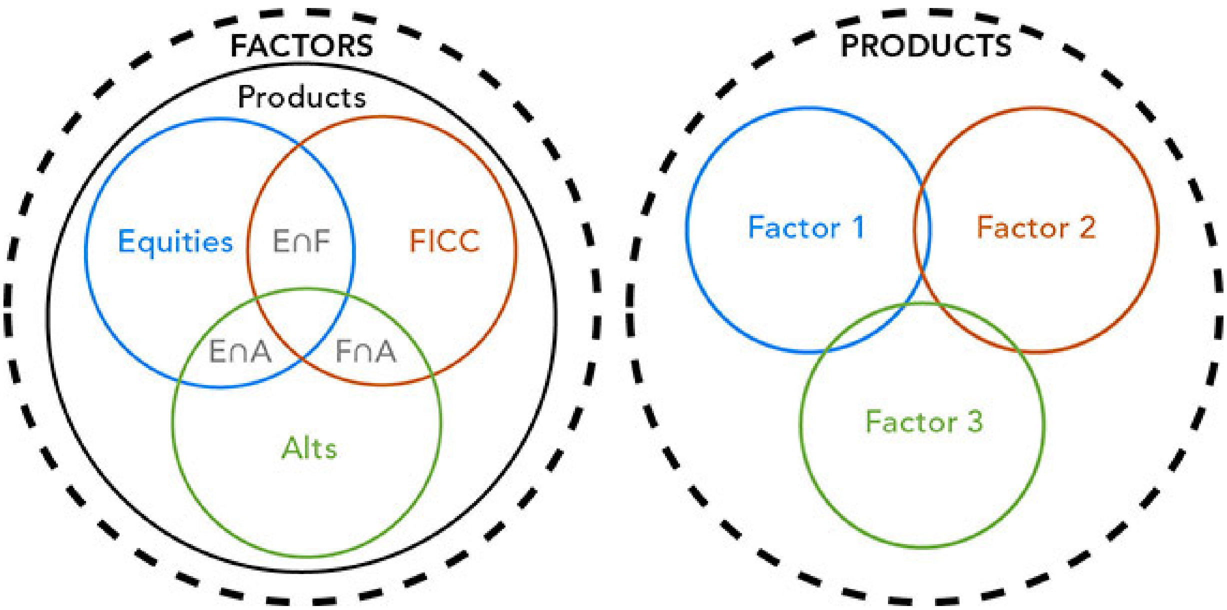


FIGURE 4-1 Asset classes overlap; factors are more orthogonal.

Products often fall short of their purpose; their construction often produces unintended results. The mechanics behind these instruments are often more complex than they might seem. For instance, Treasury futures are often used as a proxy for a *constant maturity Treasury* (CMT). Most futures track one of a *basket of deliverables*, the *cheapest to deliver* (CTD), based on a conversion formula. There are also five embedded options in a Treasury futures contract that must be considered beyond the underlying CTD and its carry. Nuances like these can make an investment behave unexpectedly and cause investment performance to diverge from factor performance (Box 4-1).

Company Exposure

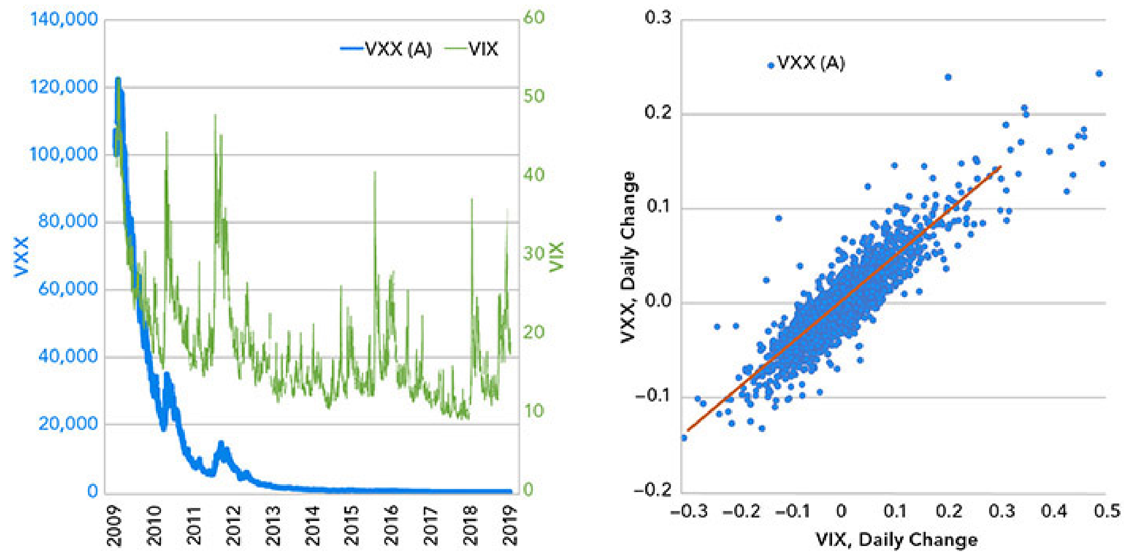
Equities have traditionally been popular investments due to the attractive returns during the past century. Equities include secondary trading of listed stocks on exchanges like the New York Stock Exchange (NYSE), alternative trading systems (ATSs) like electronic communications networks (ECNs) and crossing networks (both lit and dark), initial public offerings (IPOs), and unregistered securities.

Financing

Lending, borrowing, and carry are considerations in almost any transaction, and equities are no exception. Different vehicles involve different risks; ignoring these nuances can be disastrous, such as being caught short in a squeeze. Stock loans, repurchase agreements (repos and reverse repos), buy-sellbacks, and other schemes are similar in purpose although the mechanics, risks, and tax treatments differ. Forwards, to-be-announced (TBA), when-issued (WI), and delayed delivery involve the payment and receipt of a security on a future date, possibly subject to conditions and events.

Box 4-1 WYSINWYG (What We See Is Not What You Get)

An excellent example of a product that offers exposure to a pure factor (volatility) but can deliver something different is the popular iPath S&P 500 VIX Short-Term Futures ETN (VXX). While volatility is a relatively simple concept, the mechanics of VXX are complex. VXX does not mimic VIX as many assume but instead combines the first two CBOE Volatility Index (VIX) futures contracts in dynamic proportions. Figure 4-2A shows the dramatic difference in performance between VXX and VIX due to the roll between the two futures. Figure 4-2B shows how VXX can track well between daily rebalancings despite deviations over more extended periods.



(A) The dramatic difference in performance between VXX and VIX due to the roll between the two futures.

(B) How VXX can track well between daily rebalancings despite deviations over more extended periods.

FIGURE 4-2 VXX is not VIX.

Counterparty risk, often expressed as a *credit value adjustment* (CVA), exists because the seller may fail to deliver an appreciated security on the agreed date. Operational risk is also a concern; *rogue traders* often purchase securities for delayed settlement without reporting the trades (“stuffing tickets in their desks”). A buy-sellback, unlike a repo, involves the transfer of legal ownership and does not require a written document, making income payments and mark-to-market adjustments legally unenforceable. A lender may be difficult to find or replace and may call the security, leaving the borrower in jeopardy of a forced close-out. The borrower may have to pay an unexpectedly high price to return the security or a “buy-in” cost.

Synthetic and Commingled Exposure

It is often more efficient to get company exposure synthetically, rather than by buying shares. Total return swaps or total rate of return swaps (TRORS) may also be less expensive and contain no hidden fees, costs, custodial fees, market impact, withholding tax, tracking error, or basis risk. Because they are *off balance sheet*, there is a reduced burden for both parties that may further decrease the implementation costs. There is no legal ownership of the asset. Payments are netted, so transactions are inherently levered for the receiver. The term may be shorter than the life of the asset, providing financing for the payer. Strengths are often weaknesses. The off-balance-sheet nature may obscure exposure. If a counterparty's risk is unknown, spreading counterparty risk among many partners may decrease the expected severity of default, but increase the probability of default. Contract for difference (CFD) is another synthetic net payment contract that is highly leveraged and is often marketed to smaller investors. It generates no bid-offered spread, market impact, legal ownership of the underlier, or shareholder rights. Tax-averse investors, such as a founder of a private company that becomes public producing a low *cost basis*,[2] can hedge exposure without creating a tax event using synthetic assets such as TRORSs.

Equity exposure is often purchased through investment company securities, including open-end and closed-end investment companies, unit investment trusts (UITs), and exchange-traded funds (ETFs). Though fees are incurred by investing in these instruments, there are potential efficiencies, including the ability to invest in international securities that are restricted from particular domiciles. These investments may be subject to interruptions in trading by the exchange, and not all the instruments in the reference basket may be held by the fund.

Investment managers who focus on security selection can get company exposure in many other more esoteric forms like business development companies (BDCs), special-purpose acquisition companies (SPACs, or "blank check" companies), warrants, master limited partnerships (MLPs), and passive foreign investment companies (PFICs). Each of these carry specific risks. The online bibliography for this book contains many references to the investments discussed in this chapter; please look for it at www.QuantitativeAssetManagement.com.

Categories and Geography

Typical ways to categorize equities include geography, capitalization, style, and so on. Definitions vary even for the most common categories and change with time. Accounting rules and *day-count bases*[3] differ significantly, as do many other characteristics. International securities can be accessed through depository receipts (American depository receipts or ADRs, European depository receipts or EDRs, global depository receipts or GDRs, etc., generally referred to as international depository receipts or IDRs) and non-voting depository receipts (NVDRs). International stocks have some characteristics not applicable to domestic stocks, including currencies and capital controls, political risks, and tax treatments. Other concerns include illiquidity, information scarcity, and varying accounting standards.

Indices are dynamic and historical relationships are nuanced. Benchmarking or categorizing equities can be complicated by frequent changes to index constituents, weights, and other characteristics. Since the composition of indices and benchmarks changes frequently, analyses based on them can be unpredictable and unreliable. Geographic proportions in a primary index frequently change. Even indices designed to be weighted formulaically, such as by market capitalization, are not entirely mechanical in their weightings and inclusions. Standard and Poor's' decision to include Tesla in the S&P 500 was an example of the technical and sometimes arbitrary nature of index construction that can cause uncertainty and imprecision in analyses that rely on these benchmarks. In Tesla's case, many investors assumed that Tesla would be included in the index quickly because of Tesla's high market capitalization. More nuanced rules and human judgment delayed its inclusion until it was larger than any previous addition.

Inflation adjustments can significantly impact performance, especially in emerging markets. High inflation in the 1970s can explain negative real returns during that period. Even though raw returns do not make it evident, VIX deciles and S&P 500 deciles are strongly negatively correlated. Indices often include or exclude constituents in undesirable ways; the Japanese stock and bonds markets are good examples of inflation's influence. The Bloomberg US Aggregate Bond Index ("Agg") has a large Japanese component, and Japanese performance has been lacking, making it easy for an active fixed-income manager to outperform the Agg by merely underweighting Japan. Japanese stocks are uncorrelated with other developed markets, depressing their performance in equities. Emerging markets (EM) and frontier markets equities have also underperformed developed markets for over a decade.

EM countries are idiosyncratic, and grouping them is of dubious value. There are some similarities when EM investments are segregated by commodity focus (such as importers versus exporters and types of commodities involved), current accounts, etc. Inflation can be high in some countries, and tax treatments are uneven. Some countries have treaties, while others withhold taxes, leading to double taxation and making total returns net of nonresident taxes an appropriate metric for those countries.

Fixed Income, Currencies, and Commodities (FICC)

FICC is a broad catchall for investments that are not equities and are not alternatives. Many equity-focused investors may think of equities, particularly US large-cap stocks, as “the market” and all else of less importance. Compared to equities, FICC investments tend to be more macroeconomically influenced, more quantitative and precise, and more technically multidimensional (such as in credit, term structure, and volatility). While FICC investments have uncertainty and volatility like equities, many aspects of their relationships are calculated in more precise detail than for many other asset classes. Notably, FICC investments are often marketed to sophisticated institutional clients and thus enjoy fewer trading restrictions than are imposed on equities to protect retail investors. Unlike common stock, which is marketed to retail clients, FICC products often contain significant leverage, asymmetrical and nonlinear risk, and complex nuances like specific conventions for various markets and localities. Due to leverage, picayune technical errors can create huge losses. FICC nonlinearities often defy extrapolation and complicate multi-asset portfolio analysis.

Cash, Governments, Munis

The fixed-income (FI) market is more extensive and diverse than equities, and is somewhat of a misnomer, implying investments where payments are unconditionally obligatory, unlike the dividends and appreciation of equity securities. FI investments do not necessarily make payments of a specified amount; payments may be defined in terms of formula, and there may only be a single payment at the end of the life of an investment (*bullets*, *balloons*, or *zeros*).

Variable floating-rate notes (FRNs, or *floaters*) make periodic payments in excess of a reference index, specified by a schedule of payment formulas. Coupons may be combinations of fixed and floating (hybrids), and the borrower may be able to choose to switch between fixed and floating periodically (such as with Option ARMs).

Amortizing and sinking-fund instruments may reduce the principal amount over the investment's life, and there may be options for prepayment, conversion, puts, calls, and more. Though these instruments trade over the counter, some transaction data is available through databases like the Trade Reporting and Compliance Engine (TRACE) and swap data repositories (SDRs).

Cash or money markets are generally considered obligations with term of one-quarter year or less and may include many kinds of investments. The money market has no exchange, and dealers transact as principal.

US government bills, notes, and bonds ("govies") are considered to possess the highest credit quality, though political events have caused concern, including a credit rating downgrade on August 5, 2011. Govies also include federal agency obligations (agencies) like US Federal National Mortgage Association notes (FNMA or Fannie-Mae), even though the federal government neither guarantees Agency notes nor is obligated to pay their debt.

Global developed market sovereigns like German Bundesobligationen (Bunds) and UK Gilts can be riskier than US Treasuries. Currency fluctuations can dominate returns which may be small in comparison. Inexpensive currency hedging using forwards, futures, and the like are usually recommended. As mentioned above, the Agg has a large proportion of low-yield Japanese government bonds (JGBs) and European bonds, making deviations from the index enticing to many portfolio managers. Non-sovereigns, including Europe, issue bonds in currencies other than that of the bond's origin for sale outside that region. Supranational agencies like the World Bank and the European Investment Bank (EIB) issue bonds and rely on their members' support, which is not guaranteed.

Emerging markets (EM) sovereign debt can be issued in hard currency (USD or EUR) or soft (local) currency and is sometimes securitized. While this debt enjoys a low correlation to Treasuries and a high Sharpe ratio, it can periodically suffer huge losses. Opportunities exist when investing in EM debt but require experience, skill, care, and research. EM debt is not homogeneous, and investors must pick and choose carefully in order to excel.

Taxes are an essential consideration for many investors, and a wide variety of tax-advantaged products are available. In the United States, non-federal municipalities issue debt obligations and revenue bonds, private-activity bonds, residual interest bonds, industrial development bonds, and Build America Bonds (collectively, AMT bonds). They usually pay lower yields but compare favorably to other obligations on an after-tax basis.

General obligation bonds (GOs) may not have an explicit funding source or use but are often the least risky of all municipal bonds or *munis*. Munis carry default risk, and some, like housing revenue bonds, have prepayment risk. Bonds may be collateralized by assets or revenue rather than by the issuer. There is also a risk that the tax treatment may be altered by legislation.

Corporates

Credit spans all classes but is most explicitly discussed in relation to loans and bonds. Even US Treasury debt has some credit risk element, usually related to delayed payment or default, and many forms of debt have other risk dimensions, such as prepayments. Corporate fixed income includes loans, bonds, preferred stock, and convertible stock.

Bank loans, which are generally senior, lever balance sheets and often pay 3 percent or more above their reference rate (London Interbank Offered Rate [LIBOR], Secured Overnight Financing Rate [SOFR], etc.) periodically, e.g., quarterly with monthly or quarterly resets. They are often near the top of the capital stack (beneath protected claims like taxes and pensions). Bank loans are usually backed by assets or cash flows, like receivables, physical property and equipment (PP&E), inventory, or intellectual property.

Delayed funding loans and revolving credit facilities ("revolvers") allow on-demand borrowing and usually receive variable or floating rates. Revolvers permit repeated borrowing and repayment. In addition to the credit risk, interest rate and prepayment risk are often "wrong way" because borrowing tends to occur at inconvenient times for the lender due to systemic distress. Trading in the secondary market may be restricted and illiquid, negatively affecting valuation.

Parcels of senior loans may be issued for leveraged buyouts (LBOs), mergers and acquisitions (M&A), recapitalizations, and other purposes. Investors in *participations* must rely on the seller to interact with the issuer while *assignments* allow access to the issuer. Parcels also face a variety of other issues, and losses may exceed the loan amount and can continue after the loan is repaid. Attempts at recovering losses may be costly. If loans are accessed through derivatives, like swaps, counterparty risk is a factor. Subordinated, second lien, bridge, and other junior seniority loans may have a subordinate lien on collateral pledged to more senior loans or may have different collateral. They may be uncollateralized as well. *Covenant lite* loans have fewer conditions than traditional loans.

Corporate debt with less than a year to maturity is referred to as *commercial paper*. Longer debt issues are called bonds, debentures, notes, and other names like "spread product" because they are measured by the difference ("spread") between their yield and a reference rate like LIBOR, SOFR, or the prime rate. Valuation depends on the ability to service the debt, hence the issuer's credit quality, and fluctuates with interest rates. Corporate bond indices are usually capitalization-weighted, favoring highly leveraged, lower-quality companies. Survivorship bias is common, as "fallen angels" (IG debt that are downgraded to high-yield) are removed from the index.[4] *Reconstitution* is an issue with most indices as holdings are added and delisted.

Corporate credit ratings segregate the market into investment grade and high-yield (HY), or “junk,” and these labels imply default probabilities. High-yield bonds are difficult to value and may contain calls, clawbacks, and other idiosyncratic features. Credit rating changes may lag the events that precipitate them, and fallen angels, as opposed to *original issue* bonds (bonds that were high-yield from the date of issuance), often recover within two years. The recovery results in a risk premium that can be harvested by buying and holding fallen angels because many investors are not permitted to do so. Segmented markets like this create inefficiencies and opportunities. A relevant dislocation occurred during the Great Financial Crisis of 2007–2008 (GFC) when HY debt recovered at a much slower pace than IG debt. Stressed, distressed, and bankrupt investments may require protracted and expensive litigation, and recovery may be incomplete or nonexistent.

Preferred stocks, similar to convertible bonds, are a combination of equity and bond features. They have a *par amount* and pay fixed-rate dividends that are a percentage of that par amount. Preferreds sit between stocks and bonds in the capital stack for both dividend payments and bankruptcy. Bonds are paid before preferreds, but most preferreds must be paid “cumulative” outstanding dividends (if skipped) before common dividends can be paid.

While indentured bonds must pay their coupon, corporate boards may defer preferred dividends (e.g., if the company is not profitable), making the preferreds sensitive to company risk on the downside. Non-participating preferreds have a fixed coupon, capping their upside, while participating preferreds may increase their dividend based on a formula specified in their prospectus. Adjustable-rate preferred stock (ARPS) vary with an interest rate reference. Like convertibles, some can be converted to common stock. Preferreds may also have sinking fund and redemption provisions preventing them from profiting much when rates fall significantly. Some preferred dividends are considered qualified dividend income, and taxed at a lower rate than interest.

Convertible bonds ("converts"), which can be purchased piecemeal and in funds, can be converted into equity at a specified ratio that dilutes holdings. This feature allows them to offer lower coupons than straight bonds, and they are usually subordinate to straight bonds. Converts behave as bonds when the stock price is significantly below the conversion price, and behave like stocks when the stock price is closer to the conversion price. The issuer can often force a conversion, limiting the potential for windfall profits. The potential for forced conversion creates a significant difference between synthetic convertibles, combinations of stocks and warrants, and true convertibles. If a company induces a conversion, converts can only add value over traditional bonds if converted quickly. A portfolio that is based on asset allocations will have to manage the varying categorization of convertibles and the resulting nonlinearity as they become more bond-like or more stock-like.

Contingent convertible bonds (*CoCos*) can only be converted (or have principal written down) after a triggering event, such as government bailout or a decline in the issuer's tier 1 capital reserves below a threshold. Converting to common equity (with or without dividends) will help cure a deficiency in regulatory reserves. While traditional converts dilute earnings per share, untriggered CoCos do not. Unlike conventional converts, CoCos that fail to pay a scheduled coupon do not necessarily cause a credit event. If a company's stock is lower than a specified level upon the conversion date, capital may be written down. In that case, CoCos may be worth less than par, possibly becoming worthless without recourse even if the company is not bankrupt.

Securitized Products and Derivatives

Mortgages and other loans (e.g., *whole loans*) can be securitized into products and derivatives to form a sprawling and complex category of fixed-income assets that have attracted a tremendous level of investment. These include agency (*agencies*) or non-agency residential mortgage-backed securities (RMBS), collateralized mortgage obligations (CMOs and tranches), commercial mortgage-backed securities (CMBS), covered bonds, asset-backed securities (ABS), including collateralized obligations (collateralized loan obligations [CLOs], collateralized bond obligations [CBOs], and collateralized debt obligations [CDOs]), among others.

In addition to the complex evaluation of the underlying pool of loans (seasoning, prepayment, default, convexity, etc.), the structures themselves are complex, often with multiple classes (*tranches*) with their own cash flow rules (*waterfall*) and seniority. They may contain complex guarantees (*covenants*) and trigger mechanisms. Assumptions, like the diversifying benefits of geographic separation and even well-intentioned legal technology, may not perform as planned, as experienced during the GFC.

The website for this book, www.QuantitativeAssetManagement.com, contains a flowchart of a large mortgage model whose complexity was achieved by layering smaller models for refinancing, turnover, curtailments and payments, defaults, and implied prepayments. It shows how intricate these models are and how layering can make them intellectually and computationally tractable.

Leveraged Loans

Leveraged loans offer substantial alternatives to bonds for the borrower. Though they are often secured, unlike many high-yield bonds, and are senior to bonds in default, most loans do not seek a credit rating, nor do they need to provide as much public information. Recovery rates for loans often exceed half of their value, while high-yield bonds may recover less than half. Private equity firms may use these loans to help finance the majority of a leveraged buyout (LBO). The loans then becomes the purchased company's burden. Even in circumstances other than LBOs, European companies prefer loans to bonds, while American companies often issue high-yield debt.

Structured Notes and Linkers

Structured investments, including hybrid instruments, are flexible, sometimes bespoke, investments with payoffs that relate to one or more reference (such as an index, rate, return, etc.), and these investments may modify those reference values substantially by including them as variables in a payout formula. Structured investments include securities linked to many references including credit, indices, commodities, weather, catastrophes, and inflation (linkers). These investments may be as simple as swaps or complex *tranched*[6] instruments like collateralized obligations (CBOs, CLOs, and CDOs) and other arrangements with complex features like credit enhancement, e.g., special-purpose vehicles (SPVs). The complexity of some of these instruments may even confound their creators and result in significant valuation errors. The bespoke and complex nature of some of the designs may make the investments challenging to exit—called “roach motels” because of the 1980s-era roach bait brand Roach Motel (“Roaches check in, but they don’t check out”). Linked securities may not pay if their holdings (such as a portfolio of swaps) do not pay, exposing investors to credit, default, and counterparty risk.

Inflation linkers are typically tied to an urban consumer price index (CPI) or retail price index (RPI) in some nontrivial way, such as including a deflation floor that preserves the principal and generally provides right-way risk, increasing in value as equities decline. A commonly followed indicator is the forward *breakeven inflation* rate, which is the expected inflation for some time, beginning at a future date, and is derived from two similar interest rate vehicles, one linked and one not. Some linkers, like United States Treasury Inflation Protected Securities (TIPS) generate phantom income that is taxed annually, even though the benefits are withheld until maturity in the form of principal adjustments.

These unrealized adjustments may result in an *original issue discount* that affects the funds’ gross income and forces it to make distributions. Distributions caused by unrealized income may make the fund cash poor and induce economically perverse forced liquidations. Inflation may affect real rates and compromise linkers’ insulation. Downward adjustments due to deflation may restate income distributions as capital gains. In addition to rate risk, linkers are also exposed to currency fluctuations and basis risk. Linkers have a solid appeal for retail investors, and some consider them to have a negative risk premium because of their popularity. At one point, half of the fixed-income assets in British and Canadian pension funds were held in inflation protected bonds.

Event-linked investment returns are contingent on a specific *trigger* event or absence of trigger, such as for catastrophe bonds. Triggered investments may lose principal. Risks include credit, counterparty, regulatory, and tax-related complications. Triggers are often designed to activate only during extreme circumstances and are challenging to model and predict, making these instruments difficult to value and exit. We continue our discussion of note triggers in Chapter 19 when we address hedging and risk transfer.

Swaps

Swaps are a low-cost, convenient, and efficient instrument that can be used to synthetically replicate a return stream, or manufacture one to a required specification. In essence, the swap counterparties agree to pay each other according to any formula they choose to write in the agreement. Typical forms of swaps include total return swaps, which replicate the return of an investment, credit default swaps (CDSs), inflation-linked swaps (ILS), and volatility swaps among a wide variety of others. Some swaps are exchange traded, while others are traded over the counter. OTC swaps involve counterparty risk, and their valuations incur a credit value adjustment. Swaps are frequently used to overcome legal and regulatory restrictions, like taxation, geographic prohibitions, or regulatory capital requirements.

Because swaps are merely money transfers, they are “off the books” or “off balance sheet,” and legal ownership does not transfer between parties. This arrangement can eliminate many hindrances; for example, if an investor wants to hedge a low-cost-basis asset, he can sell the synthetic returns stream of a TRS against the asset’s natural returns and be net market neutral, without ever selling the asset and creating a taxable event. Investors who are prohibited from investing in specific geographies may purchase a return stream that tracks that location’s market index. Ownership of an asset’s return stream does not have the regulatory capital burdens of owning the asset. Moreover, transactional efficiency is improved, since all spreads and costs are only “on paper” (or incurred by hedging the exposure and not realizing a taxable event). Swaps can be standardized and documented with an International Swaps and Derivatives Association (ISDA) master agreement, credit support annex (CSA), other standard forms, or they can be bespoke constructions. Swaps may involve transfers of principal or not, currencies, *resets*, *haircuts*, day-count conventions, and many other features.

The buyer of a credit default swap pays a premium periodically until a credit event (such as a default, restructuring, bankruptcy, repudiation, moratorium, or acceleration) is triggered by the *reference entity* or one of the reference entities in a basket of underlying references. CDSs are usually written for 5 years but commonly range from 1 to 10 years and synthetically mimic levered loans without the interest payments of those loans. CDSs can be used for *capital structure arbitrage* and *synthetic bond replication*.

Marketable Real Estate

Direct investment in real estate will be discussed later in the chapter along with alternatives and illiquids, but real estate TRSs and public *real estate investment trusts* (REITs) will be treated now. Private or private placement REITs benefit from lower perceived volatility because they are unlisted. Though direct investment has its uses, there are good reasons to use other vehicles. For instance, UCITS portfolios are prohibited from holding real estate because they must be valued daily, and real estate is appraised at infrequent intervals. Despite convenience and good tracking over the long term, indirect investment comes with caveats, including basis risk for TRS.

There are diversification benefits even with indirect vehicles, especially outside the United States, but exposures may be concentrated in population centers, especially for commercial real estate. Infrequent reporting and appraised values underestimate volatility (known as *volatility laundering* when done intentionally),[6] and non-market risks like liquidity, leverage, direct costs, and implicit costs. Implicit costs like mismanagement, require difficult and artful adjustments when comparing these returns to other marketable investments. For instance, applying an unsmoothing method to the popular FTSE Nareit US Real Estate Index greatly increases the volatility of returns.

Equity, mortgage, and hybrid (equity and mortgage) REITs are usually in the form of daily exchange-traded common stock. Like direct investments, REITs benefit from rents, which are stickier than earnings, and are distributed to provide the majority of gains for investors; REIT investors are usually not as concerned about innovation as common equities. As a result, the long-term benefits of public real estate investments are different from stocks. The public equity nature of REITs burdens them with a short-term correlation to equity markets (and should be used as long-term investments) and its associated volatility, as well as basis risk relative to their net asset value (NAV).

As closed-end investments, REITs can trade at a premium or discount. REITs invest in real estate concerns, including property and loans—but not directly in property unless they acquire the property indirectly, such as through default. REITs are restricted in many ways, including shareholder dominance, leverage, being required to earn at least 75 percent of their income from interest, and distribute at least 90 percent of their income (interest and other capital gains). As compensation, they do not pay income tax, and pass the tax on to shareholders who pay ordinary income rates (thus, investors should locate their REITS in a tax-deferred or exempt account).

International REITS owned by a US investor may be considered passive foreign investment companies (PFICs), and their *mark-to-market* (over *fair market value*, or FMV) may be taxed as ordinary income. Real estate operating companies (REOCs) are also publicly traded, but generally focus on commercial real estate and reinvest profits without benefiting from the tax treatment REITS enjoy. Difficulties with legal challenges, regulation, and tax status are also risks for real estate in general, and REOCs specifically.

Currencies

Currency (also known as foreign exchange, or FX) instruments are relatively liquid and uncomplicated (like FX forwards and swaps), emphasizing spot and forward rates (with or without an exchange, such as non-deliverable forwards, or NDFs). Forward contracts may have popular dates, which may create *gap risk* if their cash flows do not align with matched liabilities, or can be bespoke. Exotic currency products are also popular.

Unlike for some assets, information flow in FX markets is swift and efficient. Social effects, including heuristics and market psychology, dominate short-term trading. Technical analysis is commonplace, and may generate right-way risk with respect to equities. Fiscal (exchange rate and trade balance) and monetary (interest rate) policy and other economic and political conditions are drivers of FX returns but are challenging to forecast consistently. Foreign exchange rates are poorly correlated with equities and bonds, though some well-known currency pairs (e.g., AUD/JPY and AUD/CHF) are often used for “risk-on, risk-off” proxies like the equity/bond ratio. Developed markets mandates often are limited to the 11 nations in the G10, with the addition of some proxies like the South African Rand (ZAR) for EM countries, or highly cointegrated currencies as a proxy for less liquid ones (assuming basis risk). Leverage of 10 times or more is expected. There are many hedging options, and a 50 percent hedge is often used as an approximate benchmark for a hedge manager.

There are varied opinions on hedging and whether currency exposure is a fundamental component of geographic returns, or an unwelcome source of volatility without a reliable expectation of compensatory return enhancement. Some investors seek a pure expression of their forecast, and hedge as much noise as possible, while others emphasize the cost of this comfort and prefer to accept the variance as a more tolerable alternative.

Commodities

In much the same way that emerging markets are heterogeneous, so too are commodities (*hard assets*)—being grouped more because of their dissimilarity to other assets than their relationship to each other; gold and pork bellies are not the same. Commodities are often exposed to shocks (droughts, floods, labor disruption, etc.), but most of these are right-way risks for equities and other assets. Instruments used for investment in agriculture, energy, and metals are often standardized, liquid, deep, and are often decomposed into spot, future, forward, carry, and quality components (using conversion factors to equate grades).

Some investors use companies that mine, produce, distribute, use, or otherwise deal in commodities as proxies for the commodity itself, but the assumption that these investments behave like the commodity they are meant to replace can be misplaced. Companies whose businesses rely heavily on commodities usually have active hedging and trading programs, and their profitability may not track the commodity they produce or consume. In fact, hedged commodity producers and users may generate returns that behave contrary to what one might expect from an unhedged company. Indices like the S&P Goldman Sachs Commodity Index (GSCI) and the Dow Jones-UBS Commodity Index (DJ-UBSCI) may be correlated with inflation, but can be overwhelmed by highly volatile components like energy, and may be distorted by the peculiarities of the futures contracts that they represent.

While ETFs are convenient and usually have tight bid/offer spreads, they are not as simple as they might appear. In Box 4-1, we discussed how VXX was a poor substitute for VIX for long holding periods. Similarly, various oil ETFs underperformed spot, most notably USO. ETF exposure to commodities is most often accessed through grantor trusts, limited partnerships (LPs), or ETNs. Grantor trusts usually hold the physical commodity, and are taxed as ordinary income for short-term gains and the maximum rate for long-term gains. In the United States, 40 percent of an LP's futures profits are taxed as ordinary income and 60 percent at the long-term capital gains rate. LPs are popular energy commodities vehicles, especially for oil and gas investing. ETNs are unsubordinated unsecured debt, and are taxed as futures—except for currency exposure, which can be taxed as ordinary income. Leveraged and inverse ETFs generally track the reference with swaps and may have to liquidate to meet redemptions. They may be taxed as short-term gains even if they were held for the long term. This can be onerous; in 2008, REC distributed 74 percent of their NAV as short-term gains.

More direct engagement can be rewarding; tangible oil and gas drilling costs are completely deductible (depreciated over seven years), and intangible drilling costs are entirely deductible in the year incurred. Lease operating costs and expenses can be deducted as well. Participation can take place through mutual funds, MLPs, royalties, or operating interests.

Another popular hard asset is timber, whose reliable biological growth represents over 50 percent of profits along with resale prices and land value. Like the upstream and midstream operations in energy production, the management of timber properties can provide benefits beyond the sale of the product itself.

Alternatives and Illiquids

Alternatives ("alts") are a catchall category whose membership is not well defined (Figure 4-3A and B). Many people consider only liquid equities and bonds to be mainstream investments, but alts have long been a mainstay of investing. Most large institutions allocate to unusual income streams (including things like tax liens, rights storage, water, etc.), direct private investments, livestock, revenue streams (royalties and other residuals), viatical settlements, and other risk transfer vehicles.

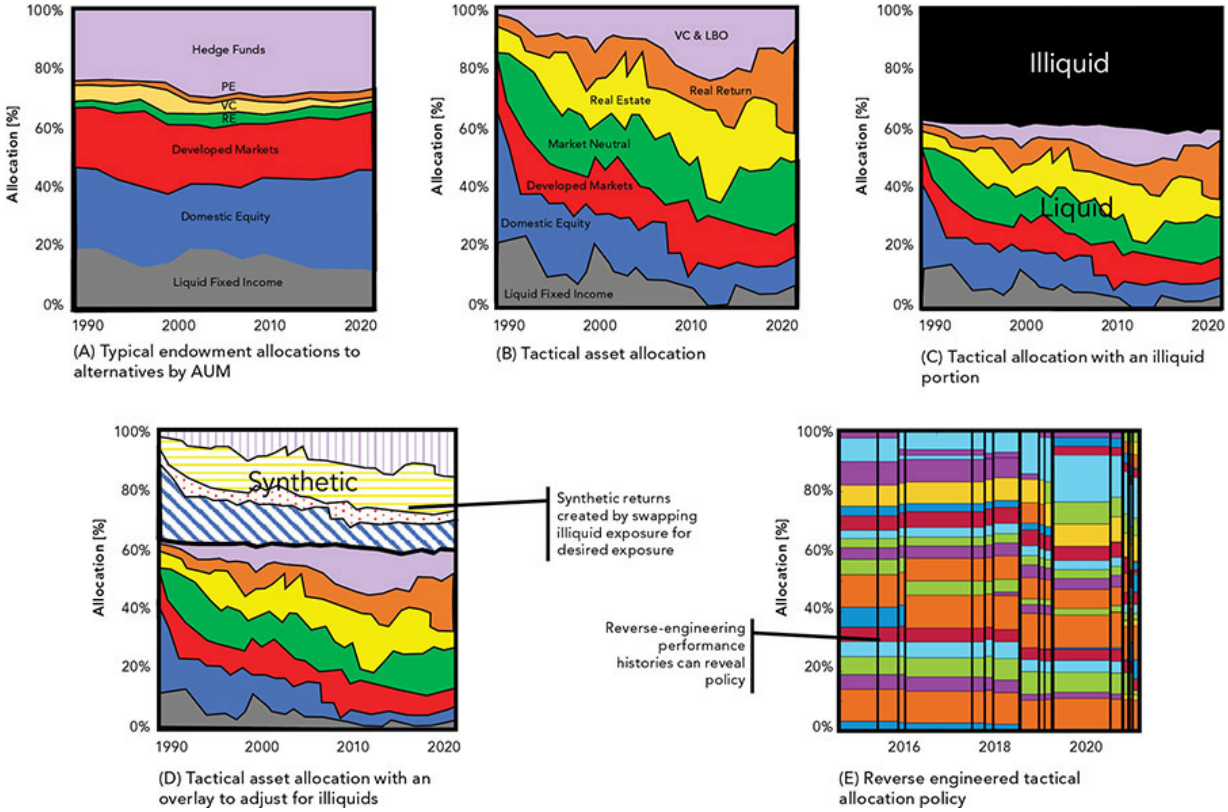


FIGURE 4-3 Illiquid assets constrain other allocations.

It may seem ironic that the most staid and timeworn investments, like real estate or gold, are considered alternatives. What alts often have in common is illiquidity, light regulation, lack of transparency, high fees, limited access, barriers to entry, and a lack of risk and performance data. An illiquid allocation can be disruptive if not addressed. A more dynamic portfolio, such as a tactical allocation (Figure 4-3B), cannot easily reduce illiquid exposures, forcing dynamic allocations to operate in a narrower range (Figure 4-3C). A similar difficulty presents itself when an investor is committed to a large illiquid allocation, such as the core business.

There are techniques to reduce exposure to alts while maintaining legal ownership and avoiding a tax event. Such techniques include using total return swaps (Figure 4-3D). A portfolio allocation is displayed in Figure 4-4E, which shows constant allocations reverse-engineered between rebalance dates (no drift). Alternatively, weights can be allowed to drift in between rebalance dates as shown in the other four panels of this figure. These schemes are useful and common. For instance, a business owner may want to reduce exposure to his concentrated income. Or a founding investor may want to diversify his holdings, but the low tax-basis of his shares make an outright sale prohibitive.

Another common problem caused by illiquidity is autocorrelation resulting from appraised valuations. It is common to resort to historical or cross-sectional (matrixed priced) estimates to fill missing data.

Hedge Funds and Managed Futures

There is a great deal of debate about active management, primarily focusing on distributions and averages of fund performance and risk. We discussed this earlier in Part I. Conclusions based on distributions and averages are sensible for most investors.

When investing in active funds, a rigorous investment and operational due diligence (IDD and ODD) should be conducted and maintained, and performance and risk must be monitored as closely as is practical (Chapter 18) and managed to the extent possible (Chapter 19). Accurate and timely data (directly or through an aggregator) and bespoke share classes, separately managed accounts, side pockets, and incentives can be negotiated for preferred investors.

In the same way that alts are heterogeneous, so are *commodity trading advisors* (CTAs). CTAs are only similar in their most general characteristics, like their legal structure. Knightian uncertainty and headline risk are severe concerns, and even funds that are marketed with low day-to-day risk (such as measured by value at risk, or VaR) may not be immune to catastrophic left-tailed return distributions or mean returns left of zero.

Strategies are diverse, including (but not limited to):

- **Equity hedge**, including long-biased, short-biased, long-short, emerging markets, market neutral, and statistical arbitrage (*stat arb*)
- **Macroeconomic** or *macro*, including discretionary or thematic, systematic, systematic and diversified, and diversified
- **Event-driven**, including mergers and acquisitions or risk arbitrage, activist, special situation, and event-driven

- **Fixed income**, including diversified, EM, mortgage and asset-backed
- **Fixed-income relative value**, including arbitrage, mortgage arbitrage, convertible arbitrage, capital structure arbitrage, and credit arbitrage
- **Multi-strategy**, or “multi-strat”

As technology advances, a growing portion of what once was considered alpha is being compressed into *smart beta* or even just beta. Many indices and replication strategies have been attempted, including some using stale holdings in 13F filings. There is still a good deal that has not been captured, specifically in high-volatility assets, vague causality, high-speed/high-complexity decision-making, and some forms of arbitrage. Crucially, good salesmanship and excellent communication are vital for success. “Soft skills” are important when getting started and in overcoming obstacles. These skills build goodwill and often make up for temporarily disappointing performance.

Smart Beta

Depending on how “smart” beta is, it could be expressed as an index or a sophisticated, active strategy. Even simple indices are plagued with imperfect compromises and biases. Generally, smart beta products are easily explained and understood, nondiscretionary, high-capacity, low-cost, low-drag, and diversified.

Some of the most popular indices are *capitalization* (“cap”) weighted, which favor size and can present significant distortions in sectors or even individual names. Cap weighting is vulnerable to bubbles, as in the dot-com bubble of 2001, and other temporal idiosyncrasies.

Different from equity indices, cap-weighted fixed-income indices tend to overweight companies that are debt-laden and may be more sensitive to stress. Negative systematic biases are easier to identify and avoid in these indices. It is embarrassingly easy to underperform a naive equally weighted allocation with a thoughtful one. Security biases are the bread and butter of most smart beta, most famously using Fama and French’s factors and the elusive holy grail of fundamentally weighted indices. The ephemeral and codependent nature of these relationships make selection and analysis confounding. In a form of Occam’s razor, many investors claim that a parsimonious set of common factors explains all things. Practically speaking, a quick arbitrage in the hand is worth many universal truths in the bush.

Private Equity

Many institutional investors have embraced private equity (PE). Private investments potentially offer substantial benefits for their many caveats, including an illiquidity premium, tax benefits of interest on leverage, a compelling narrative, and potential for active management for large shareholders. Influence can add value by forcing restructurings, including personnel changes and spinning off assets more directly than they could with public assets.

Allocations vary; typically, these investments are split between buyout/buy-in (mature companies, often called *venture capital* [VC] outside of the United States) and venture (startups without control) investments. The third most popular direct real estate investment is in growth (like pre-IPO or private investment in public equity [PIPEs]) and debt capital or direct lending. The remainder is invested in real assets (nonrenewable), funds-of-funds (secondaries), and co-investments (predominantly minority investors). Venture capital investing is often described by the maturity of the investment: angel, expansion, growth, etc. Investors often specialize in subcategories, like distressed investments or *mezzanine capital* (junior to all claims except common stock). Leveraged buyouts are a common specialty in the United States when acquisitions are often financed with more than half debt, deducting the interest.

Vintage. Deals, like loans, are often referred to by *cohort* or *vintage* because their performance is influenced by the economic conditions that existed when the deal was made. These investments usually take the form of a *limited partnership* (LP) with capital raised via closed-end funds. The fund manager takes the role of the general partner (GP), and investors are limited partners. Funds typically spend up to five years investing capital and another five returning capital (the *J-curve*). It is common to extend the fund beyond 10 years to 13 or even 20 to exit investments, typically through IPOs or bankruptcy proceedings.

Measures. PE investments are challenging to compare to marketable securities because they are quoted by the dollar-weighted *internal rate of return* (IRR), which is affected by flows in and out of the fund, rather than the time-weighted rate, or "total return." PE calculations are also complicated by uninvested capital, appraisal schedules (which grossly underestimate volatility without the benefit of public market diversification), and fees—especially fees on uncalled capital. *Secondaries*, though often considered PE, have different characteristics because of their improved liquidity. Statistics that rely on voluntary reporting are muddied by survivorship bias. A great deal of research challenges the dominance of PE returns and alleges that the benefits go primarily to fees and not to investors.[7] The enhanced risks (leverage, volatility, beta, opaque reporting, illiquidity) demand a premium, and arguments have been made that a substantial return is required to compensate for them. Illiquid investments, like PE, offer wrong-way risk, forcing investors to liquidate desirable investments in times of need (e.g., for filling margin calls during a crisis) and potentially further burdening them with capital calls at the same time.

Access. As with all actively managed investments, the best and most successful managers may generate enviable returns, but access to these investments can be limited to large and sticky investors, leaving more questionable investments to the remaining return seekers. A few exceptional managers may garner vast amounts of capital and produce high returns; the majority may suffer substandard performance. Similar to using equities as commodity *proxies*, listed PE (and PE indices like those offered by Cambridge Associates and Standard & Poor's) are a poor substitute, providing similar returns to public equity.

Internal Fund-of-Funds

Funds-of-funds (FoFs) provide valuable services to entities without the resources and scale to build a team and purchase the necessary data and technology. FoFs come with an extra layer of fees, making allocations to active management expensive. A manager may wish to offer a FoF to their clients as part of a wrap fee.

Pooling assets creates economies of scale in purchasing better resources and hiring better talent. More assets allow FoF investors to diversify more than many of the LPs could afford otherwise. Superior liquidity can come from allowing clients to sell their shares to other clients on a best-efforts basis and may result in monthly liquidity under normal circumstances. Because redemptions are only used as a last resort, investment continuity reduces costs and can provide access to soft-locked and even hard-locked funds—a key ingredient in accessing the best funds and beating the average investor (Figure 4-4).

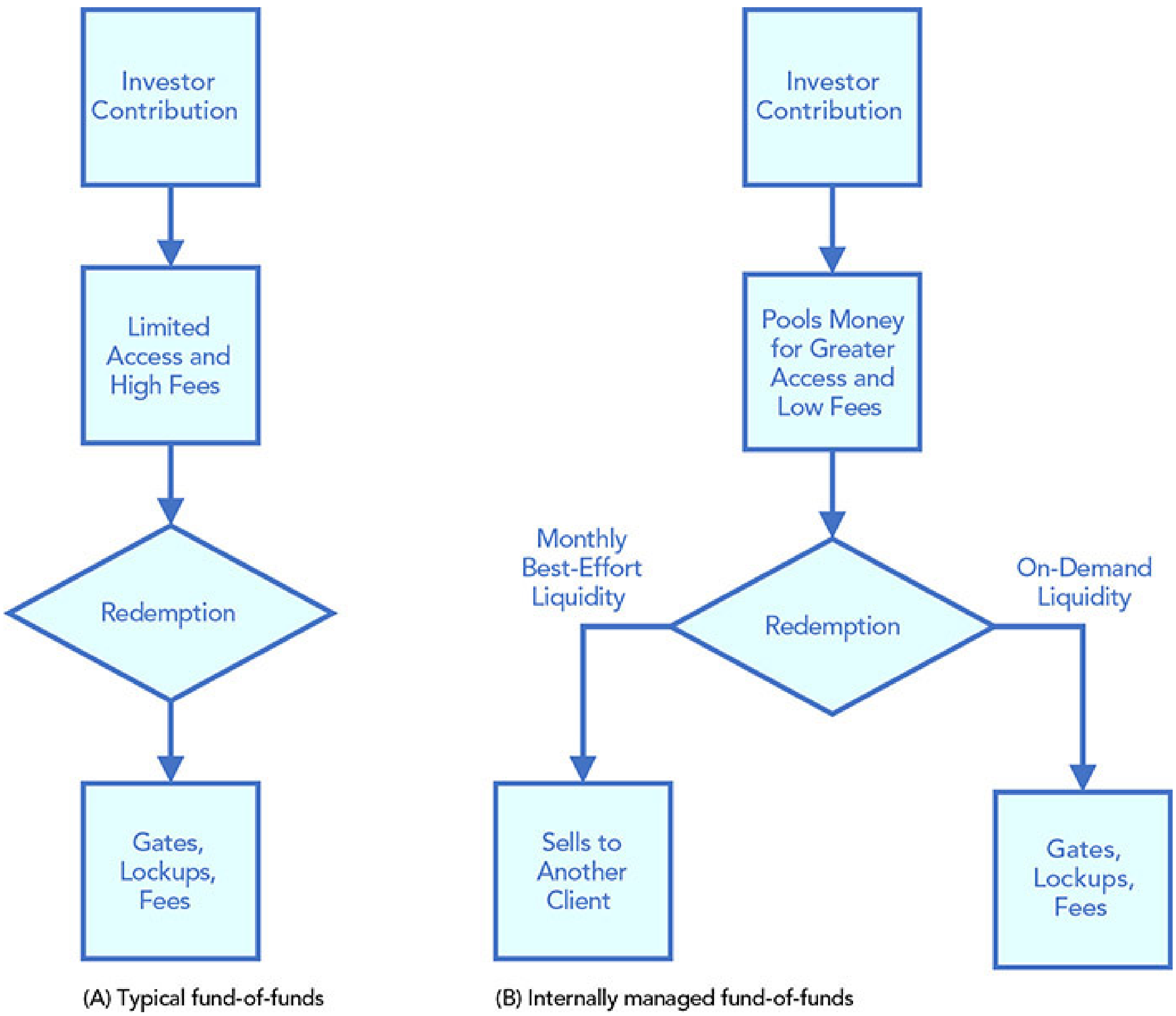


FIGURE 4-4 Internal funds-of-funds can offer many advantages for clients seeking active management.

Real Estate

This section focuses on direct diversified investment in real estate, rather than REITs or similar investments. Concentrated investments, like a primary residence, should not be considered in this category.

Like private equity investing, there are many kinds of real estate: residential, commercial, development (green, brown, grey), etc. Moreover, like private equity, direct investing in real estate requires a plan and a team for legal, environmental, financial, research and marketing, management, and so on. Oftentimes, this requires market and marketability research, land, planning, entitlement, financial analysis and financing, and possibly construction and leases—with an exit plan at each stage. Like any other life-cycle investment, it must be managed. Many proprietorship options are available with different degrees of taxation, control, and liability.

There are many dimensions to real estate, including residuals like air, surface, and ground rights. Purchases can be financed with debt and equity investments from institutions, financial entities, companies, REITs, and foreign entities. While other investments may be strongly influenced by global trends, international real estate investments, especially in emerging or frontier markets, can provide diversification and exposure to local markets. This focus may also suffer extreme currency fluctuations and high inflation that can overwhelm the benefits of the investment itself.

Infrastructure

Listed and unlisted infrastructure investment all share the trait of stability (similar to fixed income), often with little competition, but with high startup costs and burdensome regulations. These include private-public partnerships (PPP), private investment in public equity (PIPEs), the Public-Private Investment Program (PPIP), and private finance initiatives (PFIs) for transportation, energy, and social purposes.

The asset may be owned by the government and operated and financed privately (PPP) or owned and operated privately and financed publicly (PFI). Large projects tend to be long-term investments with steady and reliable income that grows with inflation or the economy. Governments often impose regulatory oversight. Frontier and emerging markets may involve regulatory and corruption entanglements. Default is rare, but renegotiation can happen during long and complex projects, with potential headline risk if mismanagement is perceived.

For example, oil and gas infrastructure investments are popular because they are poorly correlated with traditional assets and generally hold up well under inflation. Midstream investments (transportation, processing, storage), rather than upstream (production) and downstream (consumption), are somewhat insulated from commodity prices. Environmental political issues may be a structural headwind, as they face a tangle of tax contingencies that can have a significant impact on investor returns.

Art and Other Collectibles

Securitization has made art and collectibles like wine, paintings, and stamps a proper and mature asset class. Investment services, including advice, lending and residual capture, insurance, and tax structuring are available to serious collectors. Some domiciles offer lenient tax treatment for estates. Like many private investments, art can be illiquid with high transaction costs and uncertain valuations, but like a primary residence, value can be derived even when not monetized. Correlations are higher with real estate than with traded markets but are stratified by wealth bracket. Demand drops with economic activity in all but the highest levels. *Veblen* collectibles[8] exhibit a perverse positive elasticity, as increased prices enhance status and create demand.

Cryptocurrencies

Cryptocurrencies ("crypto" and other forms decentralized finance, or "DeFi") continue to be a topic of wide debate. "Bitcoin is not an asset, but a currency, and as such, we cannot value it or invest in it. We can only price it and trade it," according to New York University finance professor Aswath Damodaran.[9] He argues that financial instruments either generate cash flows (like stocks and bonds), have a fundamental use (like some commodities), can be exchanged for other things and may store value (like many currencies), or are primarily priced based on sentiment and scarcity (like collectibles). His framework identifies cryptocurrencies as currencies that have no value, only price. The utility of cryptocurrencies seems to be most relevant to the facilitation of decentralized applications that are resistant to tampering by institutions like governments, but are inferior to more centralized applications in many other ways. Political and regulatory risks are a concern for crypto trading, and innovation is strong and dynamic.

As financial instruments, they are highly speculative, socially driven, and dominated by volatility and sentiment. Adoption by sophisticated participants is increasing, but many advocates are retail-based and unfamiliar with more traditional, complex strategies and derivatives, enabling sophisticated and fast professionals to apply techniques that have ceased to be effective or legal in the more developed financial markets. Box 4-2 gives an example of cryptocurrency arbitrage.

Box 4-2 A Cryptocurrency Arbitrage Example

An example of a simple cryptocurrency derivative transaction is the Grayscale Bitcoin Trust (GBTC) premium (the difference between the market price of GBTC and the NAV, or the value of what GBTC owns). The price of the GBTC Trust can be greater than the value of the assets because it is more liquid and convenient than buying Bitcoin. The convenience of GBTC resulted in demand and manifested in a substantial premium until it peaked at the turn of 2020. The excessive premium quickly reversed to an equally large discount due to competition, including a Canadian ETF, and the ability to use Bitcoin as payment when transacting with companies like PayPal. The decline was exacerbated by pressure on the underlying (Bitcoin), which was more cumbersome to sell than GBTC, and the end of a six-month lockup, allowing investors who bought shares in a private placement to sell their holdings. Kyle Davies of Three Arrows Capital suggested that easy money contributed to the premium collapse.[10] He described the premium arbitrage trade as overcrowded leading to its collapse.

Volatility

Volatility is not the most useful measure of risk; there are many better estimates of danger and pain. However, its use is ubiquitous and standardized, allowing a common language to discuss the pricing of derivatives and structures. Volatility, particularly VIX replication, is sought after as a factor, and it can be purchased at a price even if the most available retail products are not suitable for investing. Variance swaps and other instruments, including bespoke exotics, provide pure exposure, though the cost may outweigh the benefit.

Risk Transfer

Similar to more common credit risk transfer securities like CDS and credit risk transfers (CRTs), risk transfer vehicles include insurance derivatives, insurance-linked securities, and contingent capital. Niche products may seem exotic and esoteric, but as with commodities, there exists a large and natural market for risk transfer vehicles in insurance companies and reinsurers. With these products, there is little excuse to be surprised by a left-tail event.

ESG Investments

Environmental, social, and governance (ESG), *sustainable investing*, and *socially responsible investing* (SRI) offerings have evolved into mainstream products. A great variety of metrics are still being devised; for example, the Europe-based national Sustainable Investment Fora (Eurosif) considers exclusion (negative screen), values (norms-based screen), selection (positive screen), thematic, integration, engagement (voting and activism), and impact.[11] Most scoring is based on hierarchical trees, normalizing for sectors to avoid gross misweightings, then mapping onto a Gaussian curve.

While this may seem similar to credit ratings, credit ratings are based on a specific absolute result—like the one-year probability of default—whereas ESG ratings are less focused and are relative, leading to wide disparities and instabilities.

It is not surprising that fuzziness is exploited by *greenwashers* and the ultimate investment value is difficult to agree upon. Many consider these investments to have *double bottom line* goals, and any negative economic value to the purchaser is considered a *greenium* worth sacrificing for the additional noneconomic benefits. Others argue that earning a profit and donating it to charity instead would provide the additional benefit of a tax write-off.

Strategies are flourishing, such as transition bonds used by low-ESG concerns to finance more sustainable business practices, and green bonds and social bonds used to finance ESG-friendly businesses. The impact of ESG on business profits and investment valuations are still being understood. Critical to this discussion are the effects of legislation penalizing undesired behavior or sessions, *reputational risk*, voting, shareholder activism, engagement, and ramifications imposed by society and governments. Some elements of ESG investing are commonly accepted as prudent investing, including a preference for "quality" companies with conservative accounting practices (especially in emerging markets where corruption is prevalent). We will explore some ESG models for scoring, climate risk, carbon credits, and other factors when we discuss security selection in Chapter 13.

■ ■ ■

This Chapter provided a tour of the multitude of investment vehicles and asset classes available to provide information and express opinions. More products were left out than were covered; many fill an entire career exploring just one niche. The intention was to convey the markets' breadth and richness beyond common stocks and bonds.

---

1. Ideally, factors and categories, such as asset classes, are independent of each other, but this is rarely true in practice. When the demarcation between factors or categories blur and the groupings influence each other, such as when a stock market crash in the United States affects stock prices in other markets, it is referred to as contagion. Often the relationship between factors and categories is *conditional*. For instance, stock prices around the world may normally have relatively low correlations but may be highly correlated during crashes.

2. Cost basis, *adjusted basis*, and *tax basis* are terms used to describe the initial value of an investment for taxation calculations. Early investors can acquire assets at valuations close to zero. Without proper management, practically the entire value of the asset may be taxable upon sale or other disposition.

3. Using the proper day-count conventions for different markets and geographies can make the difference between a profitable and a losing trade. Overlooking details can elicit an outsized penalty from an employer or client. It is easier to defend a decision that didn't work out than a careless error or lack of understanding.

4. As an example of survivorship bias in an equity index, after 10 years, only about 40 percent of the original the Russell 3000 stocks remained in the index. Approximately 25 percent of the original complement were left after 20 years.
5. Cash flows can be tranched (from the French word *tranche*, meaning slice) into complex parts and traded separately. Tranching and other forms of *financial engineering* can make an investment product attractive to a variety of investors—a sum of the parts that is greater than the whole.
6. Clifford Asness, Twitter post. 2022.
7. Interested readers can find links to this research at this book's website, notably Ludovic Phalippou's prolific work on the topic and Robinson and Sensoy (2013).
8. Veblen goods are high-status collectables whose prices do not follow the normal rules of supply and demand.
9. Aswath Damodaran, "The Bitcoin Boom: Asset, Currency, Commodity or Collectible?," Musings on Markets (blog), October 24, 2017, https://aswathdamodaran.blogspot.com/2017/10/the-bitcoin-boom-asset-currency.html.
10. Jen Wieczner, "The Crypto Geniuses Who Vaporized a Trillion Dollars," *New York Magazine*, August 15, 2022.
11. Eurosif *European SRI Study 2016*.

5

Financial Data

Deceptively Insidious

Traditional statistical and time series analysis methods often fall short when dealing with the nuances of financial data and practical knowledge that has been cultivated over generations. Experience is critical for many reasons, and consumers of financial analysis are likely to be extremely knowledgeable about finance, even if they are ignorant about quantitative methods.

Many quants are reluctant to climb the steep learning curve and focus on math instead, but this is a tremendous mistake. We have seen many brilliant people toil wastefully, and made to look foolish for lack of a critical but commonsense understanding of the problem.

Wall Street is not unique in using culture and jargon to exclude outsiders and protect their interests. Many practitioners find it nearly impossible to speak in everyday English when talking about finance. Few of the concepts are difficult once they are understood, and these concepts are essential to understand before analyzing our data—and long before making predictions.

In this chapter, we will discuss:

- General ways to organize financial data
- *Archival*, or *point-in-time*, data as a tool for model training
- Methods of adjusting data, including resampling, restating, and detrending

We will focus on cleaning the data and making it consistent. Before immersing ourselves in the economics, finance, and math of prediction, it is necessary to thoroughly understand and adjust the raw data. Avoiding pitfalls when using financial data requires a great deal of attention and domain knowledge. Specific and arcane adjustments are standard, and even data from the same corner of the financial world may have many different conventions that are befuddling to the uninitiated.

Organizing Financial Data

Before we can begin the analysis, we need to organize our financial data. There are many methods of data organization. We will focus on the hierarchy method.

The first tier of the hierarchy tree loosely follows the Bloomberg Professional API convention, which partitions data requests into historical, real-time, current, bulk, and fields. Different historical data is recorded at different frequencies. Some are only available daily or less frequently. Binned or bar data is available for various frequencies. And some instruments have portfolio, real-time, technical, and tick data. (This example does not include alternative or high frequency data.)

- **Historical** requests produce time series (historical and descriptive data on specific dates).
- **Real-time** data is not static and will updated constantly.
- **Current** requests produce descriptive data (and the most recent time series data).
- **Bulk** requests produce data in inconvenient formats.
- **Field information** requests return metadata describing the fields used by other methods.

Many other data services and conventions exist.

Time series data is the first category that comes to mind when thinking about financial data. Though time series data is usually stored in time-value pairs,[1] such as date and price, it is frequently transformed into cross-sectional and panel formats. Time series typically refers to a single instrument or index, like the stock prices for a single name on different dates. If we combine data from other names for a particular time, we refer to the resulting table as a *cross-section*, such as the financial ratios for several stocks on a single date. If we combine cross-sections for different times, we call that a *panel*.

Cross-sectional models are less likely to be underdetermined. The number of predictors is often much larger than the number of values that need to be estimated. Cross-sectional data is popular in some fixed-income models.

Descriptive data is more or less static and describes or labels time series, such as the Committee on Uniform Securities Identification Procedures number (CUSIP)[2] or asset class of security. Descriptive data lends itself to relational databases, since a small amount of descriptive data can be related to a large amount of time series data.

"Bulk" data, such as geographical or credit rating distributions, does not store well in tables. Bulk data can be time series, but is usually descriptive. Bulk data can be structured, unstructured, or semi-structured. Even highly structured data like regulatory filings recorded in eXtensible Business Reporting Language (XBRL) can be complex. Many well-designed tools exist to parse common data structures like XBRL.

Most financial data is unlabeled, in the sense that it does not contain notations. Bulk data may be a good source of inexpensive labels, such as National Bureau of Economic Research (NBER) business cycle date. Labeling often requires domain experts, so it is costly to acquire and may contain errors.

Time series data. As part of our analysis process, we will divide time series data into three categories:

- **Economic data,** which encompasses economic factors, such as inflation
- **Market data,** which addresses capital markets transactions like prices
- **Fundamental data,** which describes a company or other underlying entity like sales statistics

Detractors of quantitative analysis tend to believe that fundamental investing is more sensible than technical investing because there is some narrative involving causality, whether that story is true or not. While a top-down model might be more intuitive, the economy's effects on individual stock prices are not precise. Conversely, market data may directly impact prices, but the data may not be as persuasive as a consistent predictor of performance (see Figure 5-1).

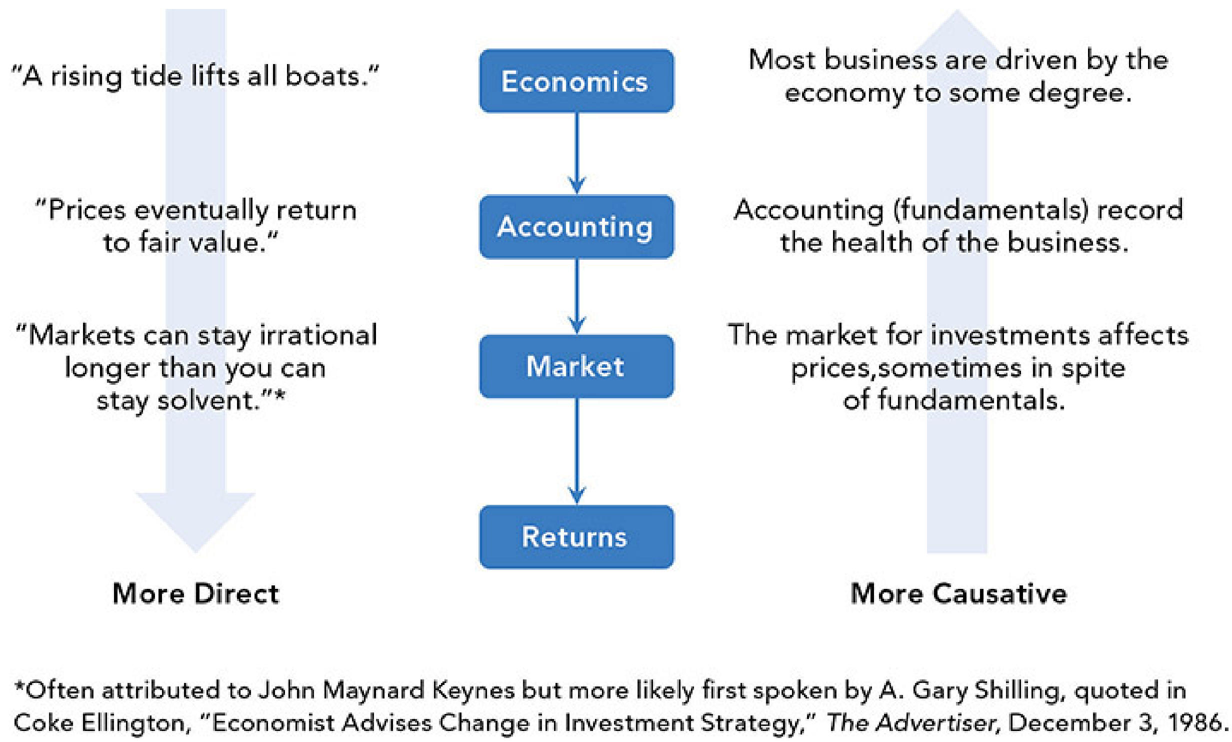


FIGURE 5-1 Deduction can flow in different directions.

Economic Data

Economic data can be used to forecast inputs for many kinds of analysis, such as a top-down asset allocation model. The type of data used and how it is used depends on the phenomenon being modeled. For instance, a credit model may need to look at *through-the-cycle* data to determine the cycle's full effect, while a tactical asset allocation (TAA) model may attempt to identify recessions within a cycle.

Economic data can be categorized in many ways, such as:

- **Employer-driven:** hires, payroll, openings

- **Utilization:** part-time work, marginally attached workers, the employment-population ratio, unemployment rate
- **Flow:** job finding rate, initial claims
- **Confidence:** hiring, quits, availability, inability to fill openings
- **Wages:** hourly earnings, cost index

Economic data may also be categorized to identify such areas as the business cycle (GDP, utilization, help wanted, housing starts, new orders, permits, money supply) and inflation (rates, money expectations).

Economic data is often multidimensional. For instance, there are several types of dates involved in economic data recording in addition to the specific date that a particular event occurred. Inflation data from January, for example, may not be announced to the media until long after the fact. Other relevant dates might include the date of the most recent revision and the date of the survey or forecast.

Each estimate and revision can create a market reaction. If an announcement or modification differs from analysts' forecasts (e.g., an earnings "miss"), it could cause a market reaction.

The yield curve is a commonly used as a forecasting measure for identifying imminent recessions. One way to measure the yield curve slope is to subtract the yield of the 2-year US Treasury note from that of the 10-year US Treasury note. This *term premium* is usually positive, and when it is not, the curve is called *inverted*. This term premium often leads recessions by about one year.

Another indicator of pending recessions is the corporate bond credit spread. The corporate bond credit spread, as measured by high-yield *option-adjusted spreads* (OAS), is not an economic indicator, strictly speaking, but a market measure. It estimates how much additional yield an investor would demand to take on the risk premium of a company that may not pay its bonds versus the yield on a US Treasury note of similar maturity.

Yet another example is the St. Louis Fed Financial Stress Index, which combines 18 data series in an attempt to identify financial events. Spikes in the indicator tend to identify major market events.

The Importance of Archival Data

It is essential to use only the data that was available at the time of the observation to avoid *lookahead bias*. While this may seem trivial, it is a common source of error. In many cases data is periodically revised, especially economic statistics, such as GDP and employment figures. Most time series databases only maintain the most recent revisions. Data is often revised as more information is released or errors in gathering, reporting, and trading are discovered. Most data sources overwrite prior data vintages, making the original reports inaccessible.

By contrast, archival databases maintain all versions of the data, so an analyst can have access to both the preliminary data that would have been available on the release date as well as the revised data released later. That gives the analyst the option of using either data set—or both, depending on the focus of their research.

For instance, if revised data is used rather than point-in-time data, the algorithm may be given a signal with future information (look-ahead bias) or miss a signal that had been edited out in the final version of the data (Type II error). When we discuss adjustments later in this chapter, we will see that unadjusted data can differ dramatically from adjusted data.

Mapping those disparities can offer an insight into how the economy, markets, or an investor may react under either scenario—or a combination of these consequences. This concept is referred to as an *agent*—a simple representation of how a person might make decisions, either rationally or with biases. Agents can be used to design interesting and complex response functions.

Here's an example: When modeling the response variable, we might consider that an agent might sell a stock if he loses a great deal of money, but otherwise would wait until he has owned it for an extended period before considering selling it (if $\Delta\pi < -10\% \mid \Delta t > 60$ days $\rightarrow$ sell). Some data providers only provide access to the actual release, latest revisions, and the latest forecasts.

How Archival Data Works

The concept of archival (point-in-time) data is straightforward, but it can be confusing. We distinguish between vintages and revisions:

- **Vintages** are a series of data that was known to be valid beginning on a specific date.
- **Revisions** are updates correcting data previously thought to be accurate (the *prior*) to what is now believed to be correct (the *update*).

Lexis diagrams are a useful tool for archival data. Though they were invented for demographic studies, these diagrams are useful for studying loans, private equity, and other data that varies by both vintage (birth date) and term (age).

Challenges of Archival Data

One drawback to working with archival data is that it is more complicated than traditional time series analysis, and it is sometimes challenging to apply machine learning to this kind of data. For instance, each time series must be represented by a matrix rather than as a simple vector so that each evaluation date has its history. Each observation has its own time series as well. Both are valuable drivers of asset performance.

Although archival data is complex, it provides a rich data set to help offset the relative lack of data in economic time series. Economic series are often reported monthly, while the asset prices that define our response are granular, making economic data far sparser.

Nowcasts. Since economic data is generally infrequent and asset pricing data is nearly continuous, economists have been inventing techniques to increase the frequency of economic forecasts. One such technique is referred to as *nowcasting*.[3]

Bulk and Descriptive Data

Descriptive data is often invariant or at least persistent, such as data that describes an investment's identity like its CUSIP number or International Securities Identification Number (ISIN). Descriptive data is often accessed in multidimensional bulk format. It can be used as a positive or negative filter, for instance, the *pricing source* of fixed-income data may indicate that it is *matrix priced* (estimated with methods such as interpolation or bootstrapping) rather than actually *quoted* by market participants.

Instead of returning a single data point, the API may return a structure of data, often requiring creative methods for storing and accessing the data. In many cases, the data can be flattened into a more standard format, requiring relational tables or repetition. Databases designed for complex storage, like MongoDB, can store this data more easily, or flattened data can be stored in traditional databases.

Descriptive data is often provided in bulk format. For instance, Bloomberg provides geographic ownership data as a table within an awkward structure. Each datum is offered as a table of uncertain size and composition, nested within the larger data set.

Market Data

Prices are primary market data but are plagued by several flaws. Differencing causes scaling errors; a $1 change in a $100 price is less impactful than a $1 difference in a $10 price. Many adjustments need to be made to market data.

Higher frequency data can be reported in *ticks*, representing discrete trades, but often transactions are aggregated into *bars*. A standard bar format specifies an interval, such as one minute, and the data packet includes summary statistics about the trades for that bar, like the first (open), highest (high), lowest (low), and last (close) price in that period. Other statistics are standard, like the volume, number of ticks, and total value in that period.

Implicit in most market prices is that the *order book*, or *stack*[4] lies behind each bid, offer, and transaction. It is assumed that many buyers and sellers would transact at the same or different prices. These bids and offers are arranged in a stacked format called an order book.

Most pricing is disaggregated and must be collected, cleaned, and organized, but some treasure troves of organized data exist in exchanges and other reservoirs. For example, all registered swap dealers active in credit and interest rate trading send data to Swap Data Repositories (SDR) as outlined by the Dodd-Frank Act and the Commodity Futures Trading Commission's (CFTC) real-time reporting rules. In accordance with these rules, the data is delayed based on the type of counterparty, clearing eligibility, and execution venue. Even SDR data contains many discrepancies and idiosyncrasies. For instance, paired trades are often reported separately and at different times to obfuscate their purpose and can sometimes be combined by noting the hedge ratios and relative prices. Many companies curate, maintain, and sell extensive data sets, including archival (point-in-time) data.

The amount of data available is another common problem. Analysts often want more history, but it may be unavailable. Long data sets may encompass many cycles, obscuring trends by averaging over regimes. Long trends can conceal reversion tendencies. The art in choosing a sample period lies in identifying a representative interval. Many methods are available to manufacture data if enough is known about its structure, but using unsupervised methods to identify those characteristics effectively is elusive.

Fundamental Data

Company analysis is usually divided between fundamental and technical. Fundamental data describes a company's operation or the behavior of whatever the financial instrument is referencing, such as a commodity.

Much of fundamental equity data comes from the income statement, balance sheet, or cash flows. There are many dimensions to fundamental data (both company and economic), such as analysts' forecasts. Two kinds of dates are relevant to forecasts: the estimation date, when the analyst made his prediction, and the announcement date, the date of the event being predicted. Data for fundamental analysis is commonly taken from several accounting tables: the *comprehensive income statement*, *balance sheet*, *cash flow statement*, *shareholders' equity statement*, and *the auditor's report* (Box 5-1).

Box 5-1 Regulatory Filings

Many regulatory filings are available on the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) maintained by the US Securities and Exchange Commission (SEC). Some of the most popular documents include:

- Current reports (8-K)
- Annual reports (10-K)
- Quarterly reports (10-Q)
- Quarterly holdings (13F)
- Unregistered offering (D)

Newer filings are reported in XBRL, though old filings may be in less convenient formats.[5]

Some of the key data that can be gleaned from EDGAR filings include:

- **Financial statements,** which are available in annual reports and 10-Ks (and may have been reported earlier on form 8-K).
- **The income statement,** which is also referred to as the profit and loss statement, the statement of operations, or the statement of earnings. It is usually released quarterly.
- **The numerator of earnings per share,** which comes from the bottom line of the income statement. Complex rules regarding the recognition of revenue can make the income statement tricky to analyze.
- **The statement of comprehensive income,** which is produced annually and expands on the income statement to include more sources of income and expenses.
- **The statement of financial condition,** or balance sheet, which is often released quarterly. It consists of the assets and financing, comprised of liabilities and equity.
- **The statement of cash flows,** which describes the receipt and use of cash from operations, financing, and investing.
- **The statement of shareholder's equity,** which explains how the number of shares has changed, such as offerings or repurchases of stock or options issued.

- **The auditor’s report,** which is helpful to judge the degree of good governance or of manipulation of these data.

Alternative data. There is also a tremendous amount of fundamental information available in alternative data. The vastness of this kind of data is difficult to exhaust, and much remains overlooked, which can create a rich source of alpha. Examples of alternative data include media releases and company guidance, news stories (including announcements of accounting irregularities and regulatory investigations), product announcements and patent filings, lawsuits, advertising, staff changes, and much more.

Survey Data

Survey data is available for some interesting analyses, such as forecasts of economic data, capital markets assumptions, earnings estimates, etc. The respondents tend to be qualified, and the questions are designed well. The surveys may lack the timeliness and frequency of social data, but in theory, the results will be more reliable and forward-looking.

Economic survey data is widely available for headline data, like the yearly change in Real Gross Domestic Product (Real GDP YOY%), and more granular data like consumer spending, government spending, private investment, etc. The forecasts and revisions are usually available. The forecasters and their firms’ names are often provided, so the accuracy and precision of predictions can be evaluated and weighted accordingly. It is easy to calculate population statistics like consensus and dispersion and compare “surprises” to market reactions.

Conviction data is less common, but information about timeliness, variance, win/loss ratios, information ratios, and other statistics may help form conviction. Conviction can also be found for surveys of estimates for company data, such as earnings per share (EPS), their securities, such as price to book (P/B), and indexes, such as weighted average maturity (WAM). There are also forecasts for capital markets summations and ratios like spreads (yield curve, EMBI+, etc.).

Sampling and Synthetic Data

Sampling data analysis is used to select and analyze a representative sample of data from a larger database to identify patterns and trends (see Box 5-2).

Synthetic data analysis is used when there is insufficient real data to accurately measure. To be effective, the synthetic model must replicate the same statistical properties as the real data set and perform in a similar fashion.

Box 5-2 Using a Gibbs Sampler to Change Frequency

One solution to infrequent data is to use a Gibbs sampler. The Gibbs sampler can use many higher frequency series (*coincident indicators*) like industrial production, business climate, economic sentiment, Purchasing Managers' Index (PMI), etc. to estimate a lower frequency predictor on dates in between actual reports, such as GDP announcements.

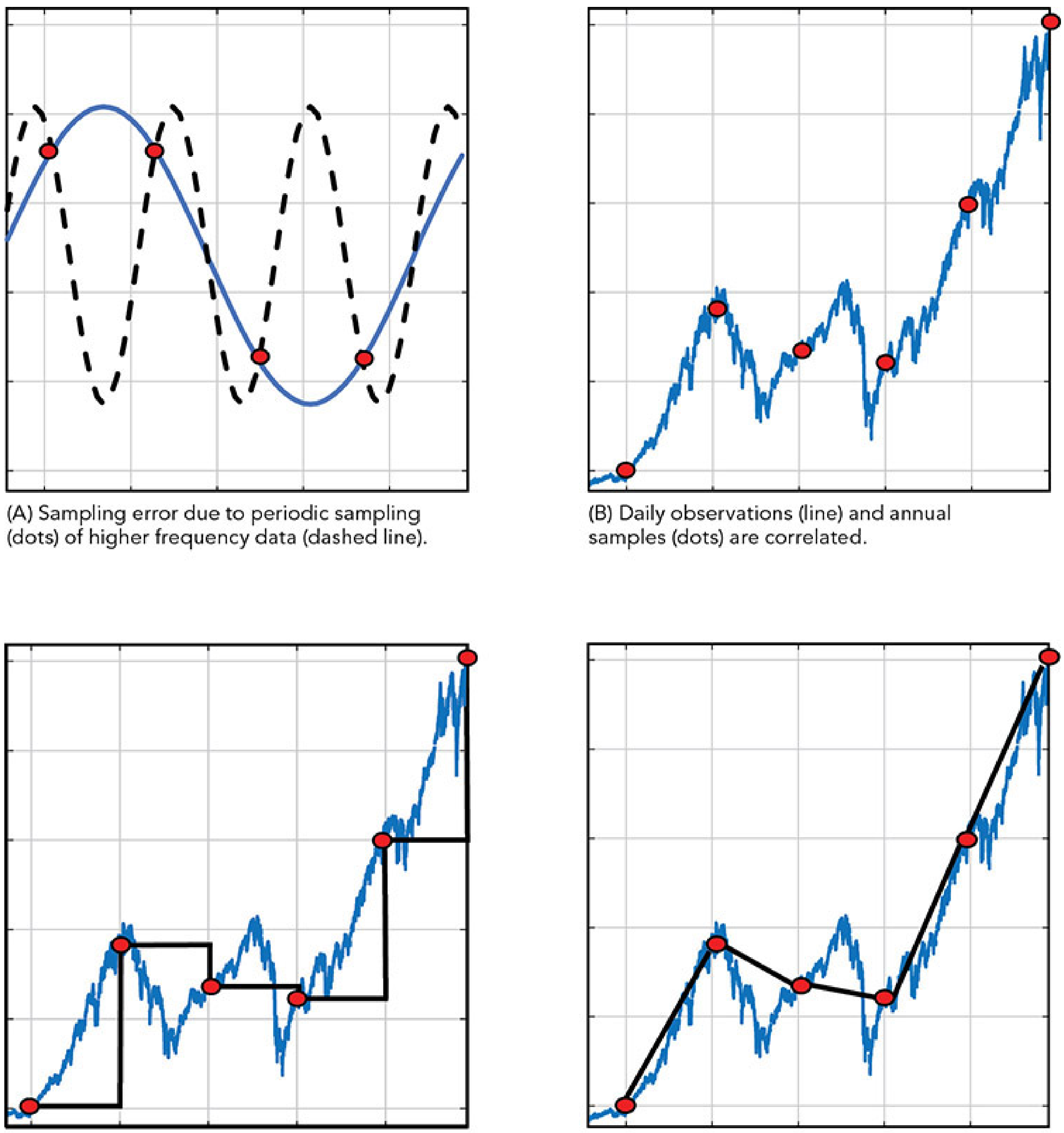
The sampler uses Bayesian probability (which uses expectations rather than frequency) to relate the coincident indicators to the predictor on the dates when the predictor is available and uses that relationship to forecast what the predictor's values should have been when it is not available. Crucially, it does not interpolate. It uses a relationship between the coincident indicators and the response.

Computer code that uses a Gibbs sampler to forecast GDP and other data is available at the book's companion website, www.QuantitativeAssetManagement.com.

Missing data and frequency. Not only is it difficult to fill in missing data, it is also difficult to *downsample* to a lower frequency. It can be even more challenging when comparing two data streams of differing frequencies.

Merely choosing values in a detailed time series to align with a less frequent one is often not an appropriate solution. Financial data is fraught with this problem. Frequently, appraised investments like real estate or private equity are assumed to be less risky than liquid assets because of sampling error. If these investments were valued with a higher frequency, they would typically be as volatile as their liquid counterparts.

Figure 5-2A shows how the wrong sampling frequency (dots) can imply an incorrect signal frequency (line). Of course, *upsampling* is even trickier. We will discuss methods for downsampling and upsampling in Chapter 6.



(A) Sampling error due to periodic sampling (dots) of higher frequency data (dashed line).

(B) Daily observations (line) and annual samples (dots) are correlated.

(C) Forward-filling copies the last known value and is poorly correlated with the data.

(D) Linear interpolation can be poorly correlated with higher frequency data.

FIGURE 5-2 Sampling and missing data-filling errors

Methods used in other disciplines to fill in missing data, like interpolation, are unsuitable for many types of financial data. Consider two perfectly correlated series in Figure 5-2B, one with annual observations (dots) and another with daily observations (line).

Forward filling (Figure 5-2C) creates the possibility of large estimation errors and confounds comparisons with correlated series. It also creates discontinuities at the observation points. *Backfilling* and interpolation are commonly used but improperly compound these problems by adding look-ahead bias because they incorporate future values.

Linear interpolation (Figure 5-2D) between the observations (piecewise linear line) is a poor choice. Not only is the implied time series far less volatile, but the estimated values are sometimes far from the true ones and have lookahead bias.

Confounding. Since there are 249 times more incorrect daily interpolated values than actual annual values, the errors overwhelm the correct data, and correlations are confounded. There are much better ways to evaluate data in between observations, such as those discussed in Box 5-2.

Smoothing. Another problem with resampling is artificially low variance. Smoothing creates the illusion of lower risk. Private Equity Index returns are similar to that of US Small Cap stocks, but much smoother. This is mostly artificial, due to the lower frequency and the appraisal method used in evaluating the private equity returns.

Regulatory filings. One area where a lack of reporting frequency is a persistent problem is the filing of 13F data by portfolio managers. Some companies (mutual funds, hedge funds, and many other kinds of financial companies) report some of their investments (long positions, vanilla options, ADRs, and convertibles but not bonds, foreign stocks, or shorts) every quarter with a lag of up to 45 days. Aside from the missing investments, the quarterly frequency poses a significant problem in estimating a manager’s holdings since there is no transparency between reporting dates.

Exacerbating the lack of frequency, portfolios can be “dressed up” at the end of the quarter to hide information. Detrending can create similar problems both in destroying relationships and interactions with other data and in injecting look-ahead bias.

Rolls and continuation contracts result when data constructed from multiple pricing sources is combined, like futures contracts that expire periodically. At each expiration date, or near to it, each old contract is replaced by a newer one in a process called *rolling*.

Several problems arise due to this combination of instruments, the most obvious being a discontinuity, called the *roll yield*. Roll yields can create unexpected effects. For instance, negative roll yields can cause a continuation contract to decrease in value while the underlying commodity is appreciating.

Even two contracts with the same underlying commodity and grade can have drastically different prices. When the term structure is in *contango*, future delivery costs more than current prices, reflecting the expense of storing and insuring the commodity. When in *backwardation*, prices are lower for future delivery implying the expectation of less demand or more supply.

The term structure's slope can cause significant effects that an inexperienced investor might not anticipate. The iPath S&P 500 VIX Short-Term Futures ETN (VXX) decays over time, ensuring that an investor will lose money even though the underlier, the CBOE Volatility Index (VIX), does not. This is due to the shape of the VIX term structure and the hedging required to maintain the VXX contract. Interestingly, it is not easy to make money shorting VXX either; we explored this in Chapter 4.

Continuation gaps may be further complicated if different instruments need to be combined. For instance, bootstrapping a yield curve may require many kinds of interest rate instruments to gather data for other tenors like fed funds, LIBOR, bills, notes, bonds, futures, and swaps. Neither the roll nor the slope of the term structure is trivial to deal with, as Box 5-3 illustrates.

Box 5-3 Roll Examples

To show how calculating even a simple roll can get complicated, let's look at the roll between a penultimate US Treasury note (the "old" or "off-the-run") and the new current note (the "current," "on-the-run," “when-issued” or WI). We consider the curve, carry, liquidity (including supply and stripping as well as demand), and accounting for weekends and holidays (“bad days”). High-demand interest rate instruments have a noticeably lower yield than those with similar maturities, as shown by the dips in yield (higher prices) near the maturity dates of the on-the-run instruments.

Normally, the yield curve is upwardly sloping. A longer dated interest rate instrument will have a higher yield due to the term premium (except for local distortions due to things like high demand for a particular maturity date). That differential is called “the curve.” The curve may be negative or *inverted*, e.g., if the economy is expected to deteriorate.

The off-the-run curve is usually imputed using swaps and other instruments less idiosyncratic than coupon-bearing investments. The liquidity adjustment is mostly due to demand for the newest instruments with maturities closer to the *original* or stated maturity. For instance, a 10-year note with 10 years to maturity may be a close match to hedge a 10-year mortgage. The previous 10-year with a little more than 9 years left may be less of a fit for that purpose.

Supply, due to the quantity of issuance and the decrease in float due to stripping, may also enter the calculation. The liquidity effect is evident by observing the reductions in yield at the original maturity dates (2, 3, 5, 7, 10, and 30 years). The most current issuance enjoys the most generous liquidity premium, and the premium decreases with each prior issuance until only a smooth curve effect is obvious. Carry considers the cost of lending or borrowing and is equal to the coupon less the financing (*repo*). If the forward rate is calculated, such as for a when-issued security, carry must be considered. Finally, errors in calculation due to day-count methods or “bad days” may need to be removed to calculate the shape of the curve.

Treasury Inflation Protected Securities (TIPS) add another dimension requiring an adjustment for inflation seasonality and the deflation floor. If we adjust TIPS, we must consider the seasonal adjustment in the Consumer Price Index (CPI) between the previous and new issuance’s maturity since the inflation calculation is not seasonally adjusted. Additionally, since the final payment for a TIPS note cannot be less than par (an embedded par put at maturity for the owner), that must be considered.

Equity rolls can get complicated as well. When calculating the roll for equity futures, we must consider future market events like Federal Open Market Committee (FOMC) meetings, borrowing costs, dividends expectations, and demand, such as using the CFTC's Commitments of Traders report.

Corporate Actions

Corporate actions such as *dividends* and *stock splits* are common and easy to adjust for. By removing or adjusting for these discontinuities in our data set, we can solve some problems in calculating our strategy's return in a backtest. But take care not to use adjusted prices as raw signals for our factors since live prices are unadjusted. A model trained on adjusted data may not behave well when using a raw feed because it will not be acclimated to the dislocations of unadjusted data.

Data collection of corporate actions faces some of the same types of challenges that the archival (point-in-time) method faces. In both cases, data collection can be difficult, machine learning can be spotty, and results can be skewed by small inconsistencies in the data collection. If the algorithm does not know how to adjust incoming data, it may be tricked by corporate actions.

There are many kinds of corporate actions that recur frequently. For instance, there were over 20,000 dividends in 2015, in addition to all the other events.

Adjustments of this kind revise past data retroactively, so that the data series' current price matches the price trading in the markets today. For example, when a cash dividend is issued, the stock price decreases by that amount on the *ex-dividend date*, which means an adjustment must be made to reset the stock's historical price data to reflect the value of the dividend.

Similarly, when a stock dividend is given by a company, the number of shares increases, so the stock price before the issuance must be adjusted up by the percentage of additional shares created for the dividend. Stock splits are handled similarly to stock dividends, except the previous prices must be reduced. The number of shares available is called the *float*. So, if there is a two-for-one stock split, the stock price must be halved because the company is now composed of twice as many shares. *Reverse stock splits* work the same way, except the number of shares decreases and the price goes up.

Spinoffs remove some of the company from the parent, leaving less value, so the share price must be reduced by the spinoff's value as a percentage of the original company's value. *Mergers and acquisitions* do not require adjustments, since the acquiring company is paying for the acquisition. Likewise, no adjustment is needed for *buybacks* and share *repurchases* since they are being paid for.

Other Practical Matters

The data we collect and the implied values we calculate are only estimates and subject to noise. Market expectations may be unbiased forecasts, but they are not "best guesses" by market participants. Some investors regard the "wisdom of crowds" to be a reasonable forecast, but various market forces combine to produce an implied value, and no individual participant may think that value is a good forecast.[6]

Distributions, not estimates. It is best to use distributions instead of point estimates and summary statistics (such as means and standard deviations). It is also important to pay attention to the relationships between values rather than to just consider the values in isolation. Care must be taken to craft models that are insensitive to inevitable data issues. Similarly, even technology problems must be planned for. When downloading large sets, be prepared for network failures, data limit breaches, and other disruptions (Box 5-4).

Box 5-4 Chain of Custody

In electronic warfare, a seemingly insignificant feature could be exploited to great effect. Data scientists are not usually motivated by life-and-death decisions, but the same impressive work can be done with financial data. Always retain raw downloaded data so it can be reexamined.

An arbitrageur often earns his keep by torturing data to expose the embedded patterns that continually change and are ephemeral. In high-frequency trading this kind of work can be even more challenging and profitable, though the sophistication of high-frequency techniques is often limited by the need for fast execution.

There are many examples of how this manifests itself, but here is one example. When a trader posts a bid or offer, it conveys information—it shows supply and demand and can be exploited by an adversary. To obscure his interest, a trader may want to submit a hidden dark order or pegging interest (PI). A PI is an order that trails the best bid or offer (BBO) with a cap. If the best bid or offer moves beyond that cap, the pegged order will not rest at that price; it would be reported at the best price within the capped range, outside of the BBO, and buried in the order book. Since the New York Stock Exchange (NYSE) displays orders where they rest, including PI orders, they would appear whenever the BBO moved away from the cap. Because this tiny detail was not disclosed by the NYSE and could allow information to leak, it resulted, in part, in a $14 million fine levied by the SEC in July of 2020.

High-frequency trading advantages sometimes rely on arcane or subtle bits of information. Moreover, this is not often directly actionable information, e.g., the level displayed is not the cap but the closest price within the capped range. If the PI cap is outside the BBO, no one will transact there. It is just a puzzle piece, part of a mosaic, a slight advantage in filling in a model of the order book. But that is the power of ingenuity enhanced by data science.

Every bit of data contains information, including its sequence. Another example of the value in retaining data relates to a regularly scheduled download. A database administrator may routinely sort a time series database by time. This sort can be detrimental if the database is large and time queries are not needed often. If intraday data is updated over several days for the purpose of retaining end-of-day marks, some of the data (the data for the current day which has not yet closed) will be assigned midday prices instead of closing prices, making them invalid. The next day’s update would provide a second price for that date which would be the correct close. If the date is recorded (but not the time of day or update time) and the database is sorted by observation date, it would be difficult to determine which data is correct.

> The subsequent download would produce a redundant date but with the correct closing price. If the database remained unsorted, it might be possible to tell which duplicate elements were right since the most recent price is most likely to be the true close. If sorted, the download order may be destroyed making it difficult to ascertain which closing price is the correct one.

Storing, cleaning, and transforming data can be the most time-consuming part of quantitative modeling. Even a well-formatted comma separated value (CSV) file can be wrought with difficulties. Figure 5-3 shows distributions of some factor data provided by a hedge fund. The distributions are quite different from each other and contain a good deal of insight into the factors' character.

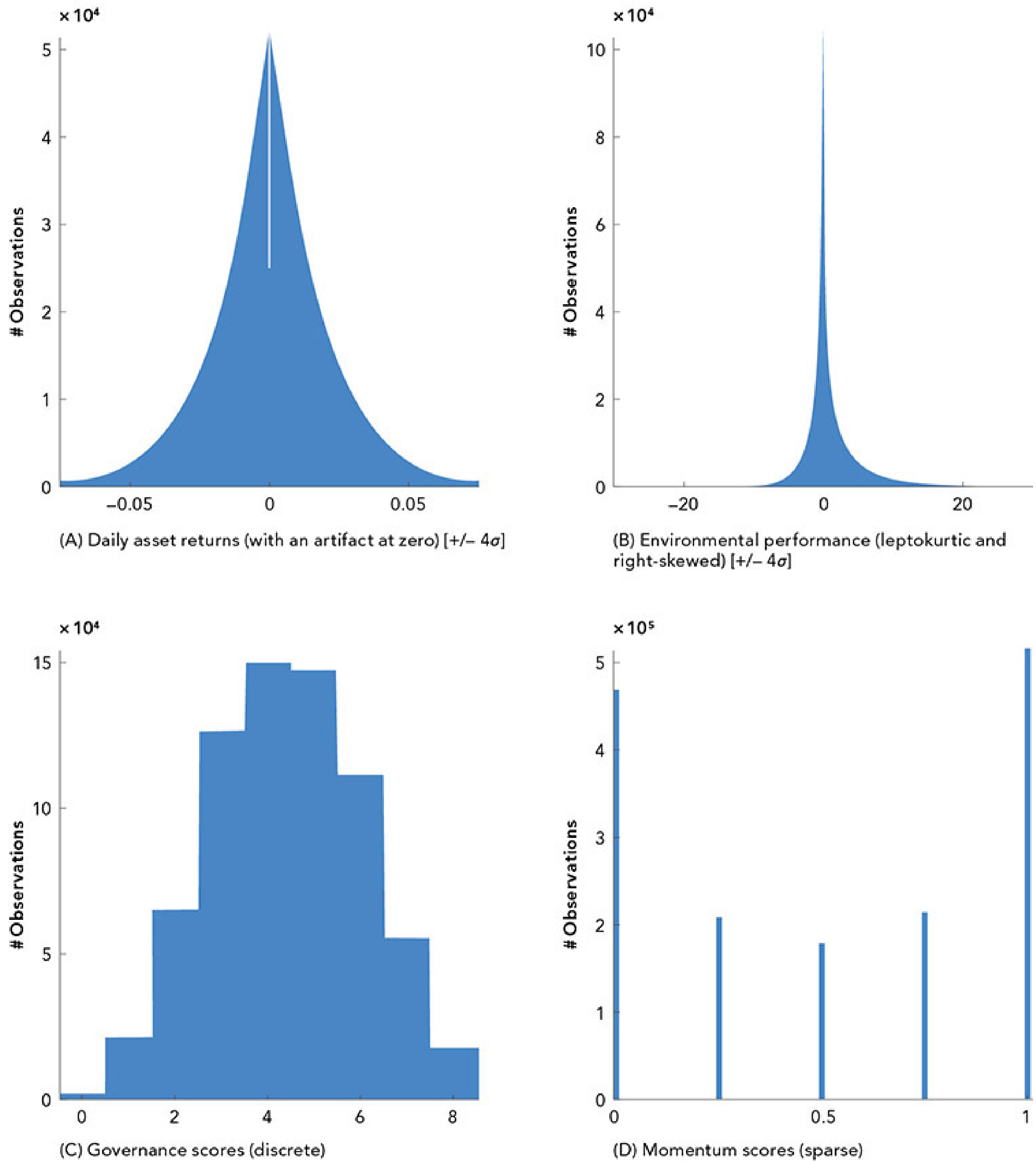


FIGURE 5-3 A single data set can provide a great variety of distributions.

By plotting stock returns' values versus their observation dates in Figure 5-4, it becomes evident that two different data sets were combined. Stocks 1 through about 1,250 have long histories. Stocks above about 1,250 have shorter histories of about 425 days. Some problematic adjustments needed to be considered before using the data. Box 5-5 shows a case study in problematic real-world data.

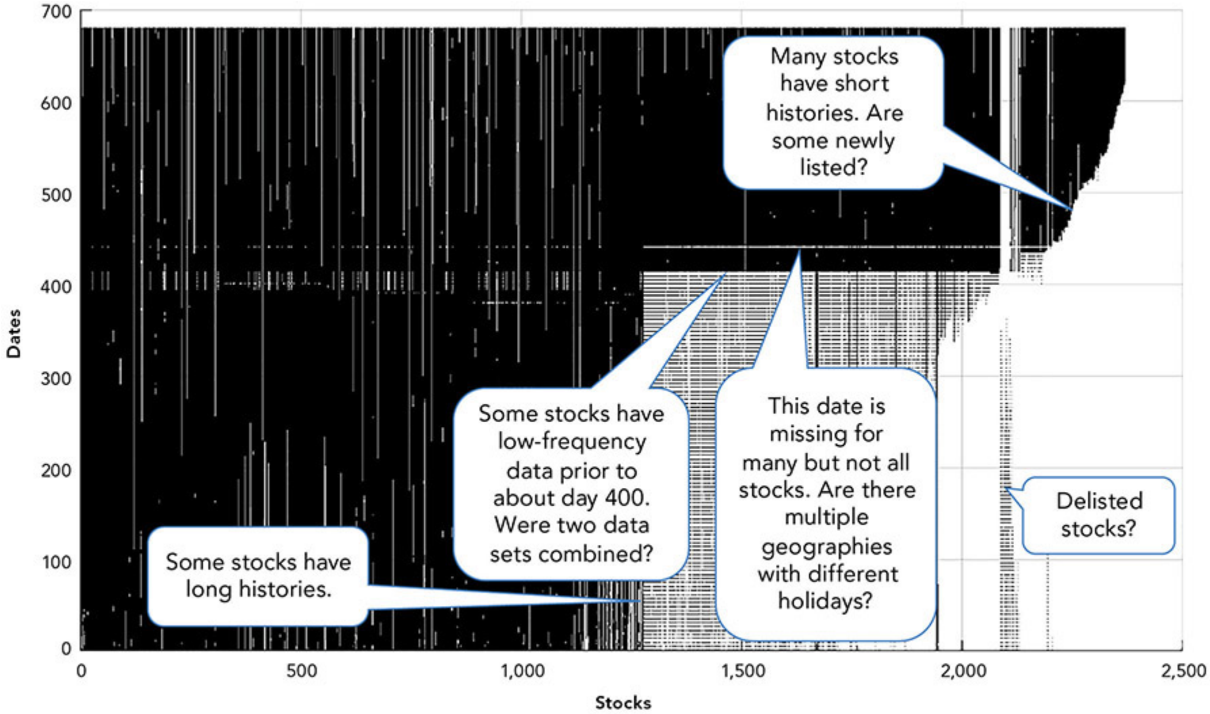


FIGURE 5-4 Some stocks report daily data for the entire set, others include a combination of daily and lower frequency, and still others appear to have been new issues or delisted.

Box 5-5 Case Study: RIA Data

This registered investment advisor (RIA) data case study is based on a project assigned to master's students at Columbia University as part of a Quantitative Asset Management course. We acquired trade and performance history for 101 robo-advisor funds, including fees and costs.[7]

We used this data to analyze the robo-advisor's performance attribution, the efficacy of their tax-loss harvesting and thematic funds, and the rebalancing methods they employed. Using a combination of performance attribution, simulation, and backtesting, the students were able to analyze past performance of the various investments and attribute success to allocation, timing, sectors, and other factors; analyze the past performance of tax-loss harvesting and thematic strategies and their adherence to their mandate, including the tax-alpha decay of tax-loss harvesting strategies; and perform "what-if?" analyses to see where strategies could be improved, for example, opportunistic rebalancing, changes in rebalancing frequency, etc.

With this information the firm could inform clients of the benefits and drawbacks of their investments and propose new ones; add detail to the existing performance newsletter and commentary for the firm; propose improvements to the robo-advisors; and develop their own strategies and defend them with research.

Classifying funds was problematic because we did not have access to the funds' holdings—only the names of the ETFs and mutual funds that the advisors owned. Classifying these required the use of two APIs and a classification model.

We were able to determine fund classifications and benchmarks for some funds. For others, we used the asset classes, geographies, industries, sectors, and ratings of their holdings and used a hierarchical method to build a taxonomy tree, classifying each investment.

Determining the portfolios' custodian, tax status (qualified or non-qualified) and theme (tax-loss harvesting, social investments, etc.) was also problematic. The manually entered labels were inconsistent, so we used natural language processing analysis and a classifier.

We cross-referenced and augmented this analysis with the holdings and transaction data. For instance, portfolios that hold municipal bonds are unlikely to be qualified (tax-deferred). Transaction data was also useful for classifying portfolios, e.g., high trading volume might indicate tax-loss harvesting strategies.

Policy. Since we did not have access to the portfolio managers, we needed to reverse engineer the policy portfolios and rebalancing frequencies. Essentially, rebalancing involves buying low and selling high to realign the portfolio manager's portfolio with their intentions. For tax-loss harvesting, it is the opposite—selling low.

We did not know the target allocations. We estimated the target allocations using several indicators. The transaction data provided us with a set of possible rebalance dates (the dates when transactions occurred), but not all those dates were true rebalancing dates. Frequently, investments were purchased or sold to manage cash flows, such as dividends or fees, or they were transacted over a period rather than all at once.

There are many ways to determine the weighting of a portfolio. Estimation methods could be as simple as returning to a set percentage of assets or as complicated as employing an optimization function or a machine learning algorithm. By visual inspection, it appeared that the schemes for this universe were simple, and so we tested for several methods. The triggers and the new weightings defined the methods.

We assumed that predetermined rebalancing may be triggered several ways, the simplest including:

- **Calendar.** On predetermined dates.
- **Bands.** When the deviation from the benchmark exceeded a buffer, called the *band*.
- **Triggers.** When risk metrics were exceeded.

The new weightings may be:

- **Benchmarked.** Returned to the same percentage weight allocation as before.
- **Band.** Returned to the band that triggered the rebalancing, which would be a small and infrequent transaction and thus less expensive than benchmark rebalancing.
- **Tolerance.** Returned to a tolerance level that lies in between the benchmark and the band. This would create less frequent transactions.

Band and tolerance rebalancing are also called *opportunistic* since they only rebalance when deviations are large. They are also “lazy” because there is a lag between the variation and the rebalance due to the buffer.

■ ■ ■

Avoiding pitfalls when using financial data requires a great deal of attention and domain knowledge. Specific and arcane adjustments are standard, and even data from the same corner of the financial world may have many different conventions that are befuddling to the uninitiated yet so evident to a practitioner that they would go unannounced in a casual conversation, such using disparate day-count conventions in different parts of the yield curve. Before immersing ourselves in the economics, finance, and math of prediction, it is necessary to thoroughly understand and adjust the raw data.

---

1. In a time-value pair, each row in a table contains a time and a value. For instance, one column may contain dates and a second column corresponding prices.

2. Many identifiers are not static at all. Tickers, CUSIPs, and other keys can be reused if a company no longer needs one (e.g., if it merges with another company or ceases operations), and some companies have multiple identifiers.

3. You can find code for an economic nowcasting model in the online resources for this book at www.QuantitativeAssetManagement.com. GDPNow is a nowcast produced by the Atlanta Federal Reserve using a Kalman-filtering technique and a dynamic factor model. It is not a traditional nowcast, but instead a frequent estimate of the current quarter's GDP, yet to be officially reported.

4. A bid or offer may represent the best bid or offer at that time and can be relied on for a specified number of shares or monetary amount. Worse bids and offers may lie "below" the best bid and offer, giving the quote "depth," and may accommodate larger transactions at worse average prices.

5. A seasoned data analyst may welcome inconvenient data because his competitors may be unwilling or unable to include it in their analyses.

6. James Surowiecki, *The Wisdom of Crowds*, (Doubleday; Anchor, 2004).

7. The data for this study was provided by Backend Benchmarking.

6
Features
Separating the Wheat from the Chaff

While there is a common perception that the modeling phase accounts for the lion's share of the predictive process of any project, adeptly selecting the appropriate features can be far more important to the final outcome. Features provide the information required to train models and predict outcomes by equipping the algorithms with the data necessary to detect patterns. Typically, the more precise and refined the features are, the more reliable the predictive potential of the process will be.

Similar to data types, features may include:

- **Cardinal** quantities whose order may matter, like returns or credit ratings
- **Quantities** without a meaningful order, like coupon frequency
- **Ordinals** for which magnitude is not important but order matters, like ranks
- **Categories** with neither quantity nor order, like industry name

In this chapter, we will examine data exploration; sampling; visualization; the selection, processing, and curation of features for our model; and related techniques used in preparing the data for analysis. Modern models may be ingenious and intractably complex, but they are rudimentary compared to human cognition. There is a trade-off between control and ease of use; the more ability an analyst has to affect the outcome of a model, the more informed and careful he must be when using it. Since most implementations lean toward control, small changes in the format of predictors can significantly affect model performance and can confound an otherwise robust model.

Most models must be fed highly processed data that is selected, organized, and conditioned so it is presented to the algorithm deliberately and clearly. *Multicollinearity* (correlated predictors, also referred to as the *substitution effect*) and *autocorrelation* (a predictor's observations correlated with themselves, such as momentum and mean-reversion) can foul many models, especially *high-bias models* that rely on strict assumptions.

Missing values can cause wholesale deletion of valuable data in common software like Scikit-Learn. Even when missing data is addressed, the method must account for patterns and information that can be gleaned from their absence. Irrelevant or extreme data can drastically skew results, but it is not always easy or wise to eliminate it, as is the case with windfall profits in investing or catastrophic losses in insurance underwriting.

Usually, data and model choices involve a delicate trade-off between accuracy and precision (bias and variance), simplicity, tractability, and interpretability. The subtleness required often defies automated solutions, requiring a data scientist who is "in the loop." *High-bias models*, like many used in economics and statistics, are robust and tolerant of imperfect data because they impose a relationship or distribution assumption. *Low-bias* models frequently overfit due to their data-driven nature, producing the common problem of high precision with low accuracy, especially when the number of observations is low relative to the number of predictors (*underdetermined*).

Here we will focus on the feature engineering process, discussing the preparation of data for analysis. The techniques required may be similar to those used in analysis, such as clustering, but the purpose is different. This Chapter will concentrate on machine learning for the imputation of missing data, rather than for prediction. We will defer our discussion of other statistical and machine learning techniques to Chapters 10 and 17 in order to isolate the most technical and perishable discussions from the rest of the text.

Feature engineering converts data into effective features. Domain knowledge is dominant at this stage of predictive modeling, especially as in relation to portfolio management. Financial information is often complex and nuanced without any clear way to measure or interpret values, and it is susceptible to manipulation, which can mislead or deceive. Many managers have great respect for domain knowledge and little trust in analysis that disregards it. Excellent analytics packages are widely available, and they are becoming increasingly specialized.

Due to intense competition, pioneering applied financial research is rarely made public outside of academia, and some of the most advanced innovations, discoveries, and tools are closely guarded. As a result, most practitioners "roll their own" software, and outsiders only see more generalized advances made in other fields, leading some to believe that the algorithm can do the heavy lifting. Human involvement in feature engineering is even more beneficial than in other stages of analysis because domain knowledge, intuition, and the acuity of human visual processing and pattern recognition are more powerful than artificial analogs.

Machine learning is multidisciplinary, so there are a variety of terms for the same or similar concepts. Many terms come from either mathematics (such as topology) or computer science, and a great deal of work has been developed specifically for biology (bioinformatics), neuroscience, image classification, and textual analysis (natural language processing or NLP).

As shown in Figure 6-1, we begin with a description and exploration of our data set, dominated by visual tools, and then select which elements appear to be important. Some time-consuming but essential cleaning and organizing of the data is almost always required, and it often involves difficult choices and trade-offs.

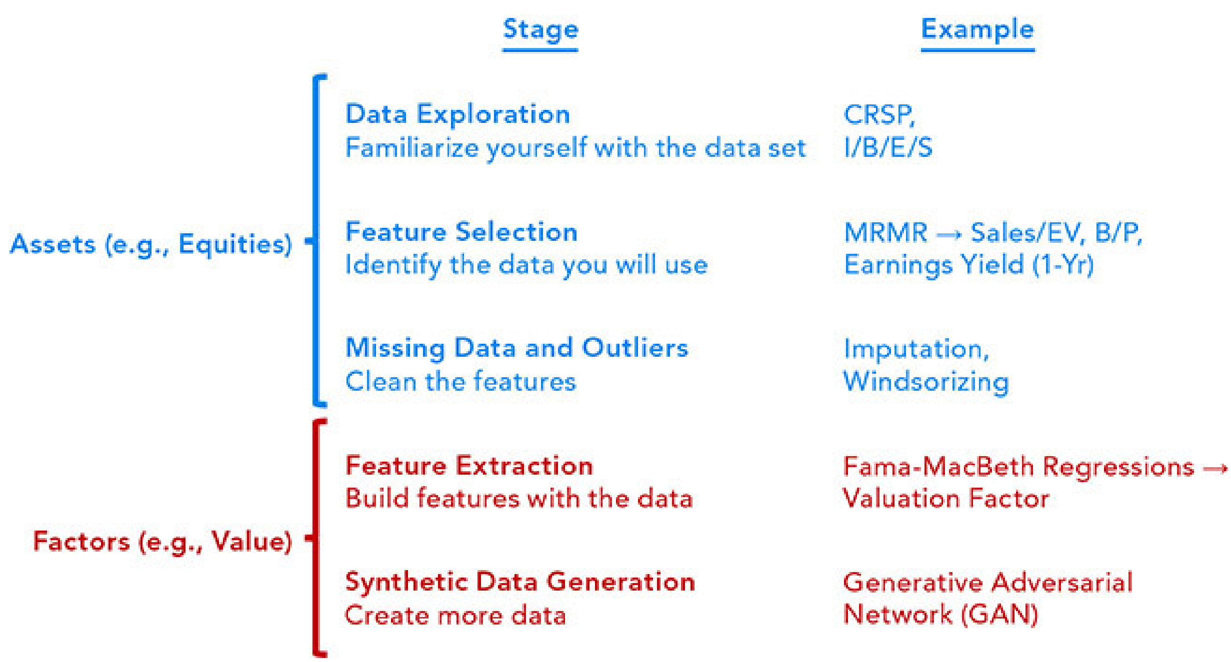


FIGURE 6-1 The feature engineering process

Once the data set is familiar, we conduct a more thorough inspection of combinations of information and interactions between them. Most financial analyses (except for some tasks like high-frequency trading, investor analysis, and ticket processing) lack enough relevant data; there are several imperfect solutions to this need, including generating synthetic data.

After our features have been engineered, we will feed that data to a model that is *mechanistic* (finding the steps in a process), *inferential* (understanding populations), *predictive* (forecasting a specific value), or *causal* (determining outcomes due to changes). It is common but misleading to use causal language to describe inferential or predictive analyses (Figure 6-2).

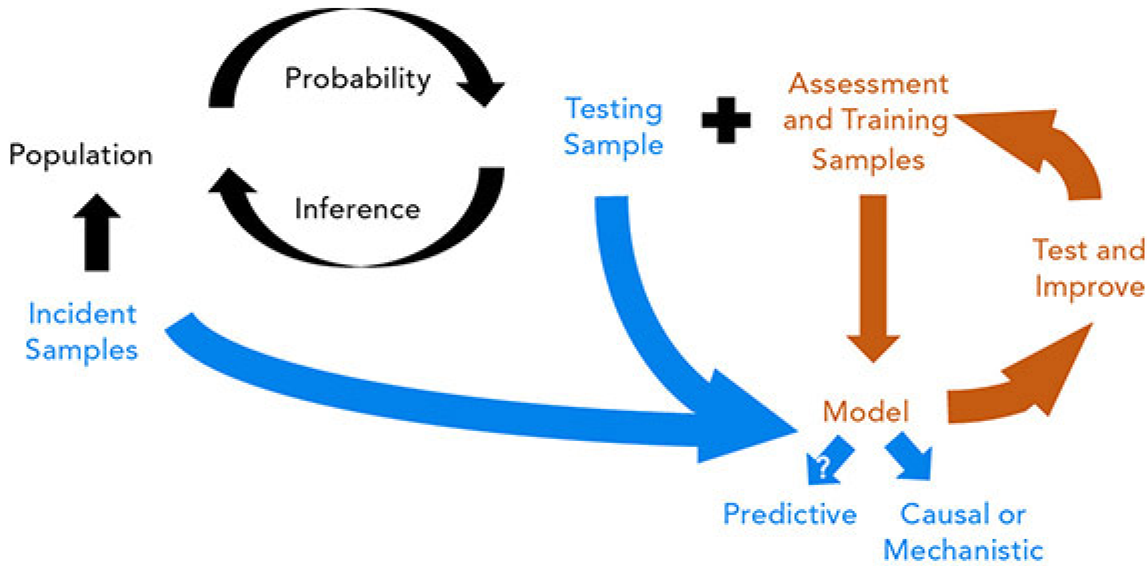


FIGURE 6-2 Partitioning data is crucial when creating and tuning models for good out-of-sample performance.

Data Exploration

Even a small company may gather far more data than it can process. Prior to feature engineering, data may be stored in a *data lake* in a raw, unstructured format, or in a more organized *data warehouse*. A *dictionary* usually accompanies the data to describe it, and a *codebook* describes the methods used to collect it.

Every operation performed on data removes some information that may be needed later, so it is important to retain the original format and keep an audit trail and *chain of custody* to track what has been done. It is best to include all operations in the training, testing, and production phases to ensure that in-sample and out-of-sample results are as similar as possible, and that the process is consistent and repeatable.

Sampling bias is a concern even when the sample is drawn randomly. Raw data allows us to refer to the purest form of the data—the *single source of truth* (SSOT) or *single point of truth* (SPOT). Auditing the data can help determine parameters for the study, such as the effective sample size. For example, data can be added incrementally to see at what point sample distributions converge to population distributions.

Ideally, the relative variability and *rate of discovery* (ROD) will converge asymptotically, as shown in Figure 6-3. In Figure 6-3A, the variance of the sample dampens as the number of samples increases to a representative amount. In Figure 6-3B, the rate of discovery (ROD) falls as larger samples encompass most unique instances in the population. In Figure 6-3C, the distributions of the sample become similar as new samples resemble other samples. Finally, in Figure 6-3D, the distribution of samples becomes more like that of the population as enough data is included.

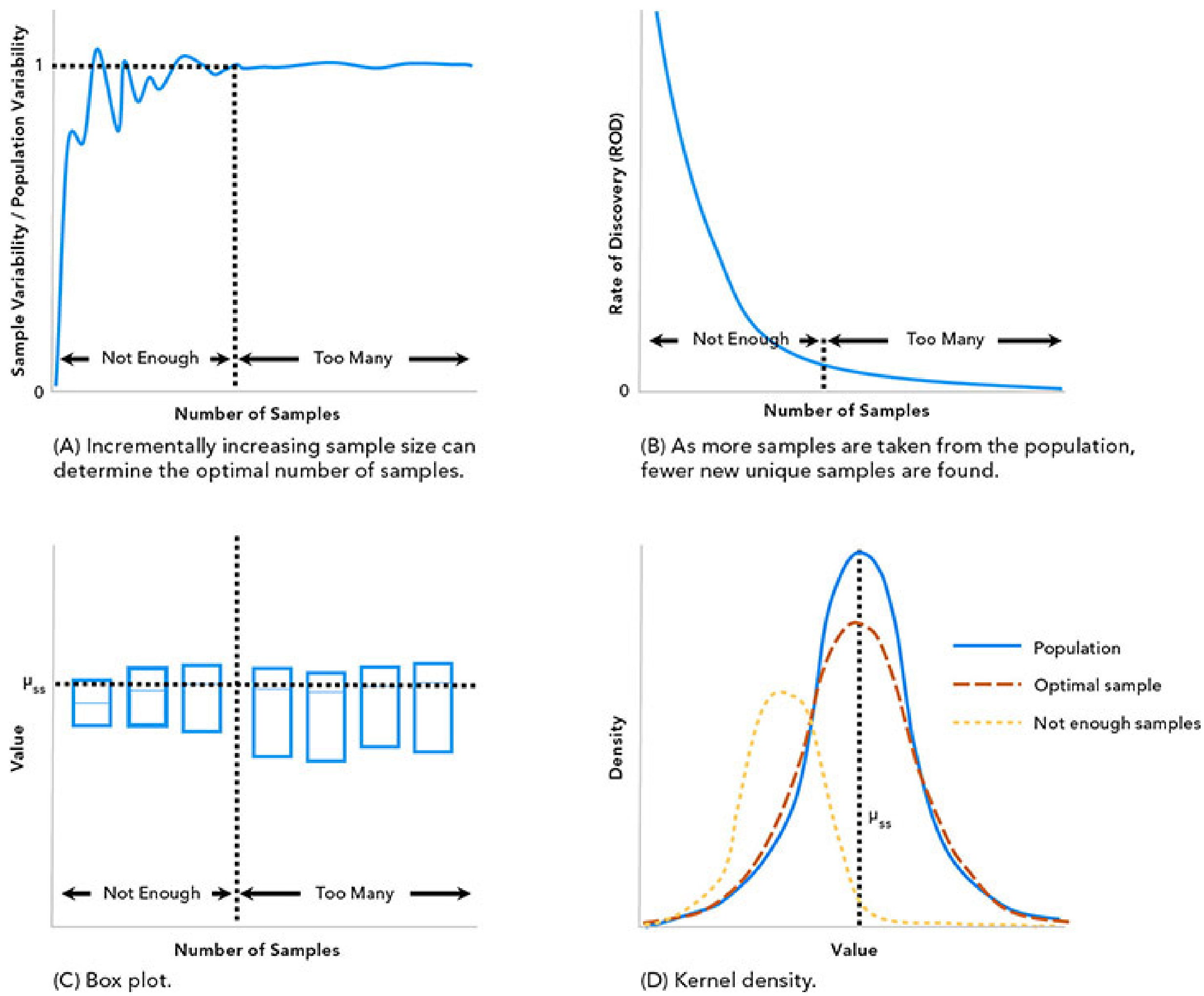


FIGURE 6-3 The (A) relative variability and (B) rate of discovery (ROD) will converge as the sample size approaches optimality, also shown with (C) box plots and (D) kernel density.

An audit may be helpful to determine if more or different data is needed before much time is wasted. Be careful to add only informative data; adding uninformative data can make models less predictive. Recall our discussion in Chapter 5 about how interpolation or filling can overwhelm an analysis with incorrect and confusing information. Even warehoused data reflects the biases of those who created it. For example, the relative ease of collecting some data may cause a librarian to bias the sample. A canonical example involves surveying by telephone, which excludes would-be respondents who cannot afford a phone. A researcher whose objective is to curate representative data may bias the sample in a helpful way by systematically removing confounding information. The audit's purpose is to survey the data to discover the character of the raw data set. This differs from the rest of feature engineering, which alters the raw data to make it easier to address the questions that inspire the model.

Sampling

The population of a study must be split into subsets to facilitate robust modeling and predictive results. The available data should be divided into at least two sets: an *assessment set*, used to train the model, and a *test set*, used for the final evaluation. The assessment set may then be further split into multiple samples, each of which should set aside a portion for training and testing. If there is enough data, a third partition—a *validation set*—may be set aside for parameter tuning, so that model performance and parameter tuning are not conflated.

Based on a survey of the population and the nature of the data, splits may be achieved by sampling in various ways. Random sampling could provide effective sets for building models (assessment) and testing the models in the most straightforward case. The assessment data can be further divided into a series of assessment samples by methods such as resampling or bootstrapping.

Stratification (e.g., segregating into classes or quintiles before sampling) is a common way to produce samples that are similar to the population from which they are drawn. Thematic models, like environmental, social, and governance (ESG) investing techniques, often favor investments in some sectors over others, such as technology over industrials. Take care to avoid bias; for example, bucketing debt issuers by size may result in *adverse selection* with many bonds for large, prolific issuers (and high debt ratios) and few for smaller issuers.

Bucketing by sector, credit quality, or some other factor, and then sampling within each bucket, will maintain the exposures relative to the benchmark. Frequently, buckets are not of equal size, but are instead selected to maintain sizes in proportions similar to the benchmark portfolios. Oversampling poorly populated classes is one way to trick a model into noticing a class that might otherwise have been ignored.

Nonrandom divisions, like holding out the most recent data in a time series, or holding out data that corresponds to a stress set, may be more suitable depending on the goal of the study.

Missing and minority data. Most financial time series have a great deal of *majority data*, such as small price changes, and a small amount of important data, such as large rallies and routs. The rarest and most valuable data tends to be the *minority data* that look like majority data—such as small price movements that warn of large ones to come. A similar concept is called *overindexing*, where one segment of the population has a stronger signal than the overall population. *Underindexed* segments have a weaker signal. For instance, one might say "Millennials are overindexed by 20 percent on ESG investments," meaning that they prefer ESG investments by 20 percent more than the overall population.

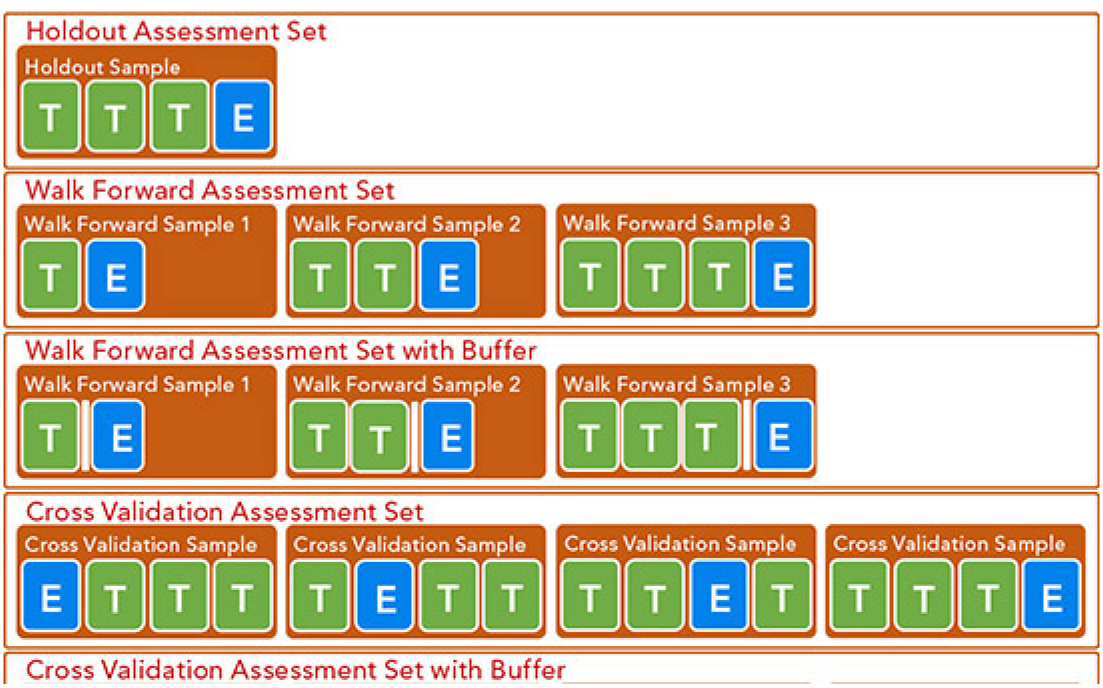
Oversampling, undersampling, and combinations of both are common ways to manage imbalanced data. Oversampling may include random oversampling, support vector machines (SVM), synthetic minority oversampling technique (SMOTE), *borderline* methods, and adaptive synthetic sampling (ADASYN), which uses a density metric. Many other techniques are commonly applied, including random undersampling and variations of nearest neighbors (condensed, near-miss, edited, neighborhood cleaning, Tomek links, etc.). Specialized techniques, like yield curve fitting and bootstrapping, are designed to work in higher-order space using domain knowledge. Computer code for oversampling is available at the book's companion website, www.QuantitativeAssetManagement.com.

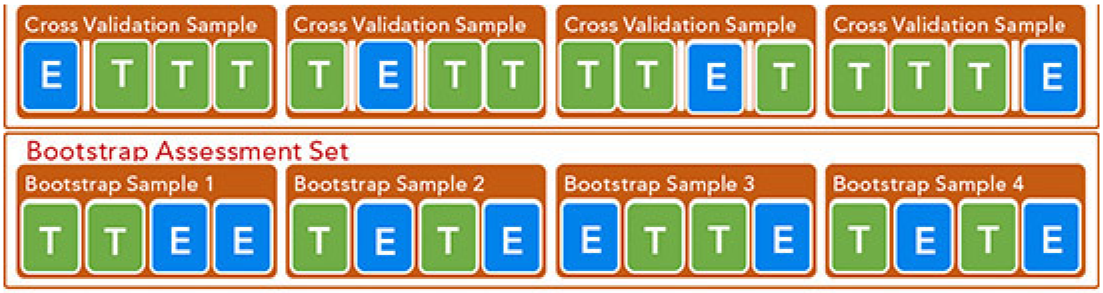
The key is to oversample high information data without introducing artificial patterns, while undersampling noise and redundancies. This helps emphasize examples of minority classifications along the border of a majority class. These general-purpose methods work best if the data is transformed so that the space is as simple as possible.[8]

For applicable implementations, a *cost matrix* can be an intuitive and targeted tool for prioritizing imputation goals. A cost matrix weights the cost of different categories of misclassification, much like a confusion matrix reports them. The costs can be implemented through resampling in the methods themselves (such as altering the impurity score, usually a Gini index or entropy, to advantage the minority) as penalties during training, or in wrappers for combinations of unmodified methods, such as ensembles, *bagging* (overbagging, underbagging, or overunderbagging), *boosting*, and *stacking*.

Each assessment sample can also be divided according to the chosen validation method: *holdout*, *cross-validation*, *bootstrap* (also called out-of-bag sampling), etc. Special techniques have been developed for time series data, like incorporating a buffer around the test set to limit look-ahead bias. The buffer must be large enough to avoid overlaps between labels—for example, five-day returns need a gap of at least five days, before the test set. However, the buffer should include that interval plus an extra margin of error between the test set and the next training set to avoid bleed-through. Figure 6-4 illustrates the difference between sampling data (Figure 6-4A) and selecting it for training and validation (Figure 6-4B). Several of the diagrams in Figure 6-4B employ buffers.

(A) There are many ways to split data.





(B) There are many ways to select data (T=Train, E=Test).

FIGURE 6-4 There are many ways to split and select data.

Ideally, all the steps after splitting raw data into sets through prediction should be inside the assessment loop, to capture as many sources of potential error and bias as possible. The splits themselves often create bias, but data will not be split in production, so the split analysis is done before the assessment loop. Figure 6-5 illustrates the stages of assessment and how the data partitions are used. The assessment set is split into an assessment sample test set and assessment sample training set. The split is used to evaluate the models, and then to aggregate (e.g., vote) on preferences. Later, the final selections of features and models are tested using the test set, and the process is repeated with new samples from the assessment set.

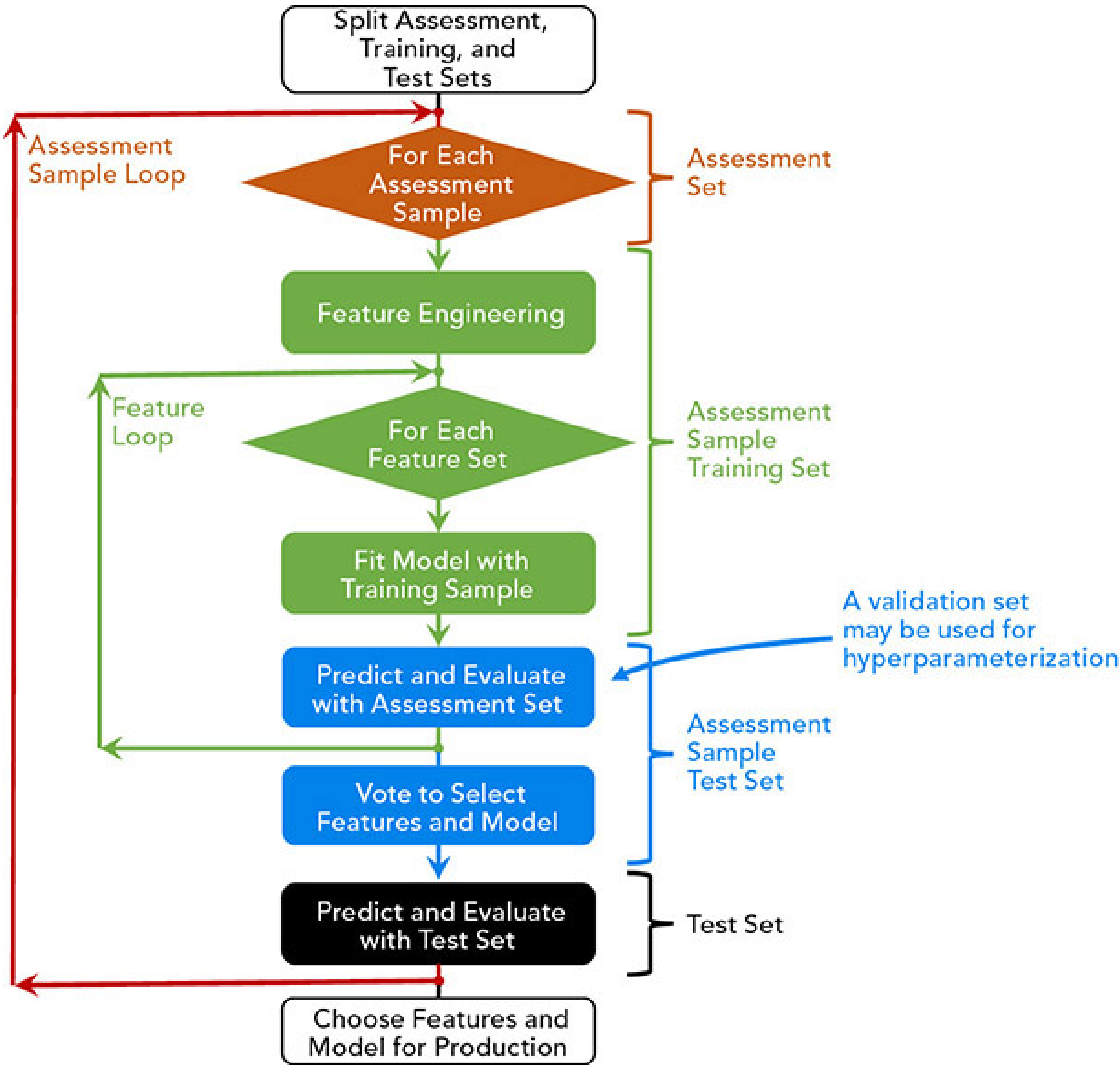


FIGURE 6-5 Properly guarding against overfitting requires three data partitions for assessment, training, and testing.

Visualizations

The first examination of a data set is likely to involve labor-intensive exploratory profiling of characteristics and visual, tabular, and summary statistics. A deep dive in areas of interest, like potential artifacts and patterns, often leads to manipulation and cleaning using tools or scripting. More sophisticated regressions and clustering techniques may be used to reveal relationships and interactions. Software such as Python, MATLAB, R, OpenRefine, Tableau, and Excel are commonly used for this task.

Visualizations are a vital tool at this stage. When initially exploring the data, the goal is to test many economically intuitive or causative variables as possible quickly. Be sure to take care when preparing charts for others. Make sure the format shows the phenomenon directly and conveys the message without requiring further explanation. Keep reproducible code and data snapshots for each chart, record the parameters, software versions, pseudo-random seeds, and other details. Solicit feedback to identify unexpected problems, such as palettes that cannot be seen by colorblind people.

General tools such as scatter plots for observations, box plots, violin plots, histograms for distributions, and kernel density plots that compare multiple distributions are frequently used at this stage. These visualizations can expose errors clearly and without need for further explanation. For instance, the limitations of population statistics are famously revealed by Anscombe's Quartet data set,[9] and more recently (and whimsically) Matejka and Fitzmaurice's Datasaurus Dozen data set.[10] They show strikingly different populations with the same mean, standard deviation, and correlation.

More specialized visual formats effectively relate a precise narrative without tedious explanations but require the viewer to be familiar with their specialized design. Some examples include probability plots (probplot) to compare multiple distributions; pairwise correlation plots to compare predictors; flattening projections to examine dimension reduction; heatmaps, dendrograms, and other clustering visualizations; and mosaic plots for categories. Simple transformations and conditioning can help create intuition for the data, including smoothing methods and various time lags for time series and faceting. Insight into the variation and noise in a sample can be aided by whisker plots, confidence intervals, jittered scatter plots on top of box plots, and the like, or Bland-Altman (or Tukey mean-difference) plots. Highly specialized plots for time series and financial data are commonplace as well, such as trellis plots to compare data, rolling (or moving) window charts, and Market Profile[11] charts.

Rather than relying on summary statistics like the chi-squared test, more in-depth examinations of categorical and tabular data, like contingency tables and correspondence analysis, can help reveal relationships, associations, and redundancy through principal components. Partial regression plots (also called added variable plots), adjusted variable plots, and individual coefficient plots, display residuals from two or more linear regressions to uncover nonlinearity and heteroscedasticity, among other things. Automated tools like OpenRefine and TensorBoard can be helpful in this area.

Feature Selection

Machine learning algorithms often cannot identify even superficial relationships, much less hidden ones that are not superfluous or spurious. Models also get stuck on local solutions and initial conditions (such as seeds) can dramatically change the outcome. To find a balance between beneficial results and interpretability, different compromises must be made for different methods. Uninformative features can confound SVMs and nearest neighbors, and even features of different scales and centers can hinder many models. Collinearity can befuddle high-bias techniques like generalized linear models (GLM). Some models that are more robust rely on methods that are difficult to interpret.

Filters are a quick and easy class of techniques to reduce the number of features, but they may not consider interactions or the type of learning applied to the features. Features often interact with each other or with conditioning variables, like economic states, justifying the time required for more sophisticated feature selection methods. *Inference tests* may be required to better determine which features are more valuable than others. *Wrappers* evaluate feature efficacy outside of the model. *Embedded* methods are more comprehensive but can be burdensome; they may use a sophisticated model and embedded methods integrated with the model itself so they can be tailored to the model’s particular needs.

Features can be evaluated *stepwise*, in a backward fashion, sometimes called *recursive feature selection*, eliminating features, or work forwards, adding features iteratively. Stepwise methods have many caveats; Type I errors are more likely to cause correlated predictors and multiple tests.

It is sometimes prohibitive to select an optimal feature from a large set using filters, and far more consuming using wrappers or embedded selection. The most sensitive machine learning techniques can take a great deal of time. If evaluating all combinations is prohibitive, a random selection may a better choice, keeping a consensus score for each feature over many trials.

A branch-and-bound search, nesting, and recursion can be employed in clever ways. Many sophisticated optimizations have been developed, including the genetic algorithm, Bayesian optimization, gradient descent, stochastic optimization, and simulated annealing. Optimizations often have a preferred stopping criteria. When a threshold is needed, an option is to use the significance of a feature composed of pseudorandom numbers or a randomized version of another feature, either by mixing up the order or drawing from a similar distribution.

Interactions and Conditioning

Interactions complicate relationships, but are almost unavoidable in finance. Experimental design (*design of experiments*, or DOE) is used to tease out these relationships using methods like randomized trials. DOE teaches that interaction is predictive only when:

- The factors are predictive in isolation (*heredity*).
- The interaction can only explain part of the response (*sparsity*).
- The stronger the interaction, the less predictive that interaction is (*hierarchy*).

When more than one factor affects a response, the factors may interact with each other with or without a correlation between the two factors. When the factors drive the response but fight each other, they are termed *antagonistic*; without much interaction, they are classified as *additive*; when they act together to amplify the response, they are called *synergistic*. (See Figure 6-6.)[12]

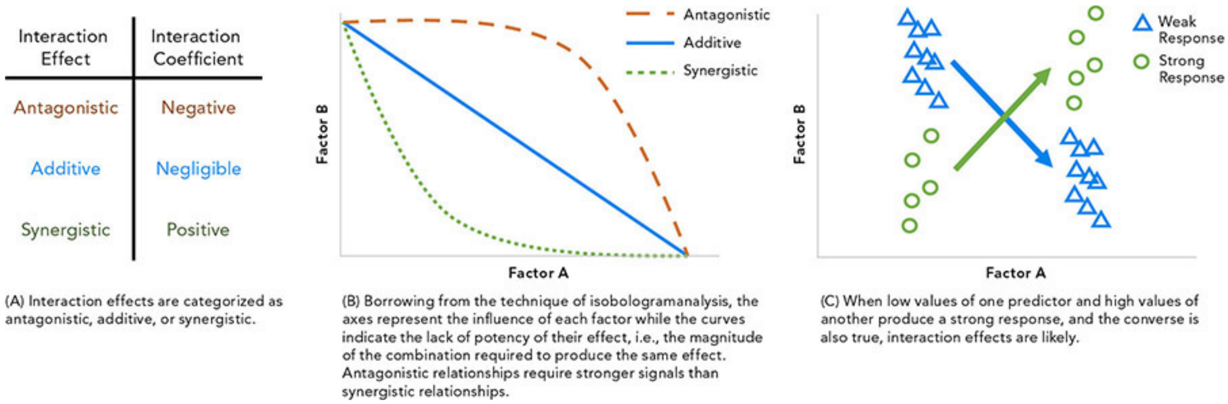


FIGURE 6-6 Identifying interaction effects

If our model is parsimonious and intuitive, such as a classical economic model, our interactions will be few and well defined. But if we take an automated or scattershot approach, we may produce too many terms to evaluate. To complicate the task, the interactions are sensitive to the preprocessing procedure and may be created (as an artifact) or destroyed (resulting in information loss) as the features are engineered. If interactions are identified in the raw data, they can be added as additional features to augment the preprocessed data.

It may be easy to eliminate terms that are weak predictors or interactions that produce few identifiable observations. Care should be taken when using standard methods to account for degrees of freedom and Type I and II errors. For instance, consider adjusting p-values using the Bonferroni, Benjamini-Hotchberg, or Benjamini-Yekutieli methods, or using more powerful techniques to avoid p-hacking and other shortcomings.

Many machine learning methods are well suited for data sets with interactions, including tree-based methods, nearest neighbors, support vector machines, neural networks, and many others. Some less flexible methods employ various *penalization* techniques to wean down large feature sets while preserving predictive power. These techniques will be discussed in Chapters 10 and 17.

It is also common build a secondary model to analyze the residuals from the primary model to identify features and interactions that were excluded. Feature importance techniques including partial dependence, Individual Conditional Expectation, and the “H statistic” can be helpful at this stage, especially if the interactions are numerous.

Preprocessing, Missing Data, and Outliers

Missing and imbalanced data is a perplexing problem in finance. We will discuss it as it applies to preprocessing and, soon after, imputation using machine learning.

Missing Data

Methods for resolving missing data are related to outliers, such as trimming, Winsorization, zeroing, or capping. The standard method of substituting, backfilling, or averaging is inappropriate for many reasons, including the ever-present specter of look-ahead bias. Even carrying forward, substituting the training set mean, median, start or end data, or random data sampled from the distribution are choices that rely on fragile assumptions. These methods must be included in the *training-testing loop* and repeated with each iteration and “in production.” With *cross-validation* and similar techniques, *stratified sampling* should be considered to ensure a representation of minority class data.

A more significant concern is determining why the data is missing. The nonrandom absence of data may be an essential predictor and a source of selection bias. If missing data is replaced, this should be recorded, and the result should be examined for predictive information.

For example, an additional categorical feature can record whether data has been filled, smoothed, substituted, imputed, or otherwise altered, and that indicator can be used to determine if those altered values are more or less predictive than the rest of the sample. Structural or other systemic causes may be a source of debilitating bias.

Visualizations like a heatmap can reveal patterns and frequencies of the missing data. Co-occurrence analysis can help identify combinations of predictors that result in missing data. Some models like neural networks, and many models used in Scikit-Learn, effectively delete entire rows when one of the cells in that row is missing data.

Other methods, like classification and regression trees (CART) and naive Bayes classification, deal with missing data directly, such as with surrogate splits. Not all these methods are equally effective; some ignore missing elements, and some create a “missing” category that treats all missing elements the same when they might be better assigned a value like a random number drawn from the population. Finally, some assign a numeric value that may have some artificial ordinal influence unrelated to any genuine relationship or order.

It is often much easier to modify data in several subsequent processes rather than rolling some or all of the processes into one elegant and tortured function—euphemistically called coding “golf” (writing using the minimum number of [key]strokes) or “WORN” (write once, read never) code. Intricate techniques are often used, like imputing missing values using machine learning.

Outliers

Outliers (also referred to as *deviations*, *exceptions*, or *novelties*) in new observations outside the training set can be a result of data entry errors, precision or equipment limitations, sampling errors, manipulation (clipping and censoring), and even intentional misinformation. Misinformation frequently occurs when point-of-sale employees are asked to fill out lengthy applications or questionnaires, as in the robo-signing foreclosure crisis of 2010.

Extreme values may be items of interest, such as windfall profits or catastrophic losses. These values make many distributions in finance non-normal or *fat tailed*, and frequently violate the assumptions of statistical techniques. The *king effect* describes the situation when some legitimate values overwhelm a distribution. It is easy to unintentionally exclude these valuable insights when they are more appropriately treated as a minority class in an imbalanced data set.

For instance, if we categorize daily returns between –2 percent and 2 percent to be “neutral,” a model that classifies all daily returns as “neutral” would be correct 93 percent of the time (for S&P 500 adjusted returns between 1927 and 2021). While that would be an excellent level of accuracy for a stock prediction model, it would provide little information.

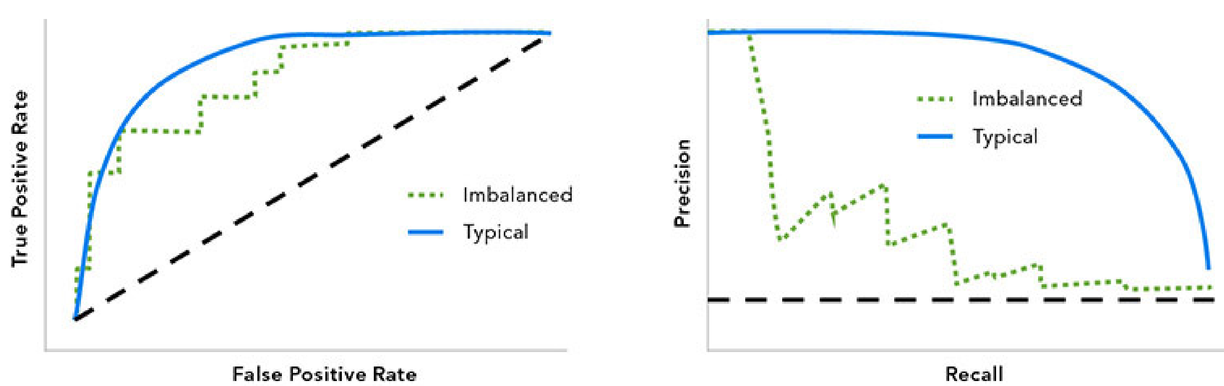
Some machine learning techniques allow customizable penalties for errors based on class, but can become increasingly problematic when upsampling or oversampling the minority class using Markov chains, autoencoders, and other methods, and downsampling or undersampling the majority class by removing values. Both methods are unlikely to be effective with minimal numbers of observations; take care to minimize sampling bias, especially near category boundaries. Techniques like substituting correlations or deviations for features are possible solutions.

Detecting outliers is easier than dealing with them. Often an examination of the distribution of a class using kernel density or extreme value theory is sufficient, but more sophisticated techniques, such as using the reconstruction error of autoencoders or decision trees, may be necessary. As with other elements of statistical and machine learning analyses, outlier detection tools have been developed specifically for time series.

Common metrics can be misleading for the classification of imbalanced data:

- **Threshold metrics** such as accuracy, error, sensitivity, specificity, g-mean, precision, recall, and F-measure
- **Ranking metrics** such as true positive rate, false-positive rate, ROC, AUC, and Kappa
- **Probabilistic metrics** such as RMS, log loss and cross-entropy for the logistic regression algorithm, and Brier score

One noteworthy difference in these methods is whether they measure the majority class or the minority class, which can influence their effectiveness when analyzing large imbalances. Receiver operator curve (ROC; based on both classes), precision-recall (which uses the minority class), and area under curve (AUC) are all relatively unbiased (Figure 6-7).



(A) Typical vs. imbalanced receiver operator curve (ROC)

(B) Typical vs. imbalanced precision-recall curve

FIGURE 6-7 Typical vs. imbalanced receiver operator curve (ROC) and precision recall curve

Even unbiased measures can be deceiving when data is severely imbalanced or skewed. If imbalanced data is used with a model that does not calibrate its probabilities, (like support vector machines [SVMs], trees, and nearest neighbors), take care to calibrate the probabilities explicitly. Other models (like logistic regression, LDA, naive Bayes, and neural networks) calibrate automatically. Make sure to set aside data for calibrating probabilities, and don't reuse that data. Stratified sampling may be prudent for calibrating imbalanced data. Isotonic regressions are often used to scale well-behaved probabilities, and Platt scaling uses a logistic regression for more extreme values.

Unsupervised outlier, novelty, and anomaly detection (also called one-class classification, unary classification, and class modeling, respectively) identify outliers using only a single class, and may delete data considered outliers. SVM and positive-unlabeled (PU) learning methods are most popular for one-class classification, but other methods also exist.

Feature Extraction

Feature extraction involves transformations to make features more salient and their effect more pertinent, dimension reduction to depreciate noise and redundancy, and feature expansion to make relationships clearer. Expansion also reduces the computational burden required to discover them and the possibility that they will not be discovered. The ratio of features to observations should be minimized to avoid overfitting and to reduce processing time. Transformations as simple as normalization and regularization can have an enormous effect. Akin to data mining, a "feature explosion" created by indiscriminate transformations is common; several software packages are freely available for this—notably, highly comparative time-series analysis (hctsa) for MATLAB and Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh) for Python.

Feature crosses, combining many features in many ways, can allow high-bias models to learn combinations they may not have otherwise tried. It can also decrease training time, even in low-bias models, but at the risk of a feature explosion and spurious combinations. Combining debt-to-equity and Global Industry Classification Standard (GICS) sectors may distinguish leverage inherent in different businesses, but combining day of the week with debt may be useless.

When extracting features, it may become evident that additional data sets should be appended from other sources, possibly outside the data lake. Expansion may be as simple and relevant as extracting the month and day of the week from the date so the model can find monthly and weekly seasonal patterns more easily. Time is thorny, however, considering differences in time zones and seasonal variations across hemispheres.

Using time as a variable becomes more complex in finance than in most other disciplines because of a plethora of day-count bases and other technical details. Feature templates, feature stores, and automated feature transformation tools can help suggest and perform extraction, but they create a risk of spurious or uninterpretable solutions and feature explosions. Tools like volcano plots (statistical significance versus size of change) can be helpful.

Crosses may require quantizing, or other transformations, to compare different predictor types, but in doing so, the encoding can be problematic. For some circumstances a fine bucketing is advisable (such as for credit ratings), while in others, a coarser division may be more predictive (such as investment grade [IG] versus speculative [HY]).

Ordinal Transformations

Feature extraction frequently involves both univariate and multivariate transformations. Transformations may reduce or expand the predictor set. A variable may be modified, multiple variables may be combined, or variables may be split or combined in multiple ways creating more variables than the original set.

While many analytical techniques do not require simple transformations, the modifications will almost always make the analysis quicker and help start the algorithm in the right direction, similar to starting an optimization close to the global minimum or maximum. It may seem that many of these transformations are trivial, but by leaving them in the hands of the analysts, the algorithm designers added additional ability to control the workflow. For instance, it is common to scale predictors so they all have a similar range (such as –1 to 1, or 0 to 1). In some circumstances, an analyst can emphasize a predictor by increasing its range, effectively weighting it more heavily. The risk of this flexibility is the possibility of unintentionally emphasizing predictors and biasing the data.

Nearly all modeling requires cleaning and rescaling. Linear relationships can benefit from these adjustments, though they can make interactions more difficult to identify. Econometrics techniques often involve univariate transformations, and their tools for tasks (such as detrending and deseasoning) are ubiquitous. In finance, people who rely heavily on univariate transformations are often called *chartists*, *technical analysts*, or *market technicians*. Simple multivariate transformations, like price to book (P/B) or return on equity (ROE), are standard practice for fundamental analysts.

Examples of transformations include:

- **General,** such as scaling, normalizing, regularization, smoothing, weighting, differencing, filter, power (such as Box-Cox, pgot)
- **Categorical and clustering,** such as encoding, hierarchical, k-means and k-medoids, density based spatial, spectral, Gaussian mixture, nearest neighbors, hidden Markov models
- **Dimension reduction,** such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), factor analysis, nonnegative matrix factorization, classical and non-classical multidimensional scaling, Procrustes
- **Feature selection,** such as minimum redundancy maximum relevance (MRMR), neighborhood component analysis (NCA), Laplacian score, RelieF, sequential
- **Time series and econometric,** such as lead/lag, frequency (such as Fourier, spectrum) deseasoning, detrending, decomposition, lag operator polynomial, specialized filters (Hodrick-Prescott, stable seasonal, S[n,m] seasonal)
- **Technical,** such as oscillators (accumulation/distribution, Chaikin, moving average convergence/divergence or MACD, stochastic,

acceleration, momentum), volatility (Chaikin, Williams %R), volumes (negative/positive volume index, relative strength index [RSI]), rate of change (Bollinger, moving average, on-balance volume, price and volume trend, Willams Accumulation Distribution Line [ADL])

Ordinal transformations that involve a single feature without considering the interaction with other features can be effective for linear models. These transformations can expose or simplify a feature's relationship with the response, make minority observations more salient, enhance information (such as replace rankings with probabilities), or decrease bias (such as replace probabilities with rankings).

All these methods should be performed as part of the assessment loop to help discover and minimize bias and overfitting. Splitting features can provide context and domain knowledge that would be difficult for an algorithm to discover. Decomposing convoluted features, like converting dates into days of the week, holidays, and seasons, can provide insight that algorithms would find challenging. Even if an algorithm were successful, time spent on splitting data would detract from other learning.

Ranking, binning, discretization. Ranking can moderate the effect of outliers on a distribution and help models like trees outperform more sensitive models like SVM.

Discretization (quantizing) can be useful to partition data more effectively, such as binning data by density (such as trading volume or tick counts) instead of fixed time periods (such as five-minute or daily "bars") so that active periods are emphasized even if they occur infrequently. Other methods include clustering, entropy-based bins, and expectation maximization.

Some models, like linear and logistic regressions and neural networks, do not handle correlations well, while others like partial least squares do. Converting coordinate systems and scales (nominal, ordinal, ratio, and interval) can make features more meaningful. Optimizers can combine scales efficiently, such as capping one asset and limiting the ratio of two assets all at the same time.

Unit changes, such as earnings per share (EPS) rather than simply earnings, are commonplace in financial analysis. Normalization and standardization can help compare scales with less bias. Probability smoothing (including Lagrangian, ELE, Add-Tiny, and Good-Turing) can help reweight features to reflect intuition or domain knowledge.

Normalization and standardization. Normalization adjusts the range of a feature's values to make them more uniform or to emphasize features. Standardization can alter the shape of a predictor's distribution by scaling or centering the data. Since most classifiers are optimized for sparse data, take care not to transform a sparse data set into a dense one by moving the center mass from zero to a nonzero value. Because the distribution of factors changes over time, it helps to rescale their distributions periodically in a walk-forward cross-sectional analysis. Sigmoid functions, the Box-Cox, and Yeo-Johnson transformation can bring extreme values closer to the mean. Winsorization and other thresholding techniques can minimize the effect of outliers.

Logarithms, reciprocals, power functions, transcendentals, and a host of other relatively simple operations can make features vastly more effective. Kernel-based functions, functions with hinges (used in rectified linear unit, or ReLU, activation functions), and polynomials, especially splines, can reduce the complexity of a relationship. Polynomial and piecewise linear approximations are common in transaction cost modeling.

Deseasoning, detrending, and other econometric methods help modify the distribution and make the data less offensive to a method's assumptions. Something as simple as differencing values in a predictor can eliminate a time-varying mean and help make a feature more stationary.

Categorical Encoding

Encoding is common in programming when data needs to be transformed into a new format to be understood. Many models, with the notable exception of trees and naive Bayes, cannot accept categorical inputs and require converting categorical predictors into a numerical one.

Encoding may convert a category into an ordinal number, which may incorrectly imply a relationship (for example, between healthcare and mining). Many methods have been devised to avoid this, like converting categories into binary numbers, using frequency or another distribution statistic, a quasi-random number, or hashing.

Binary encoding can describe categories without order by creating new binary predictor variables (or "dummy" columns in the design matrix) for each class of a categorical predictor. For instance, red, green, and blue can be represented as 100, 010, and 001. This solution is called *one-hot* encoding because all combinations of binary predictors for each category contain a single one (*hot* state) while all other predictors for that row must be zero (*cold* state).

Binary encoding is overdetermined because one fewer column can be used if a class is designated as the default class (zero in all binary predictors), such as 10, 01, and 00. *Dummy coding* may be more efficient, but also makes results less explicit, since the predictor columns are no longer one-to-one with the categories. This may complicate learning for some models, like neural networks with softmax layers. *Effect coding* is like one-hot encoding, but uses –1 instead of zeros for the reference value, which allows some models to capture interaction effects.

Encoding can become burdensome when categories are sparse, creating too many columns in the design matrix. Categories may be combined to reduce the number of predictors. For situations in which there are many dense predictors and order is not appropriate, feature hashing (the *hash trick*) can be used to assign a quasi-random value to each category. It is possible for a hash to assign the same number to more than one category causing a *collision* that merges categories (*aliasing*).

Since a prevalent class may collide with an outlier, a majority class may overwhelm a minority class. There are many ways to control collisions, such as aliasing similar categories (for example, combining classes before hashing rather than cryptography) using a locally sensitive hash (LSH), or using a locality preserving hashing (LPH).

For incident models, which accumulate data in a streaming fashion, a comprehensive list of all class values may not be apparent when training the model. One solution is to encode additional placeholder class names or to assign new classes to a default category, like "other." It is also possible to use a log transform to map classes so that bins become larger as the number of predictors increase.

Target rate (similar to *leave-one-out*) encoding replaces one-hot indicators with a population variable, like the mean of a numerical variable or probabilities found by counting the observations in each category. Some measures, like an odds ratio, can incorporate joint and conditional probabilities. This number must be taken from a hold-out, or out-of-fold, sample that is not used for training or testing.

Transformations, like logarithms, and shrinkage methods may be helpful to make distributions of target values more normal. Many encoding methods, like one-hot and dummy encoding, imply that the predictors are independent. This is usually not a problem, because many models use regularization and the default value for missing data. A similarity matrix, populated with some distance measure like correlation, is used in embeddings (hidden layers)—for example, learning by gradient descent.

Multiple incongruent encodings can be combined to allow different types of data in the same meta-model. For instance, we could combine models that use numerical data with models that use visual data.

Temporal Transformations

Financial time series are a specialized form of data (and financial machine learning is its own specialized discipline), because financial data is behavioral and time-varying. Time series create complex problems not found with most other kinds of data.

The independent variable, time, also called the *displacement or index* variable, imposes a dependency and order (*autocorrelation* and *multicollinearity*) on the data that thwarts many convenient sampling techniques due to the time and effort required. It is common for each time series feature to require its own model, necessitating an inordinate amount of time spent evaluating each feature.

Other characteristics common to time series, like noise, trend, seasonality, autocorrelation, and non-stationarity, violate the assumptions of many models. For instance, many time series are related in a way that makes them not orthogonal.

Identifying and stripping these features away from time series is resource intensive and time-consuming and complicated by frames of reference. What may be structural (trending) to one analyst may be cyclical (reverting) to another. Seasons differ from mere cycles because there is a mechanistic cause. For instance, the periodic *window dressing* effect is attributed to deleveraging balance sheets for regulatory filings.

While the season may differ from company to company, the cause is direct and unambiguous. In financial modeling, multiple unrelated features often have the same autocorrelation function, creating spurious relationships.

It is not usually necessary for data to be strictly stationary, but enough to be so in the second order, or in a wide sense. Regardless, it is important to predict the signal by adding the trend, cyclicality, and seasoning back in at the end of the analysis.

Assuming the independence of previous values (lack of historical memory—called the Markov condition, property, or assumption) can greatly simplify a problem. Because time is continuous and most analyses are not, it is common to bin data into "bars" described by statistics related to the bars' distribution, such as open, high, low, and closing price or volume-weighted average price.

There are many other ways to partition bars, including:

- **Event time,** popularized by J. Peter Steidlmayer's Market Profile
- **Bins** determined by volume
- **Ticks** to separate slow periods from active periods

Even if the displacement variable is kept in the time domain, the spacing may not be uniform due to weekends, holidays, and time zone differences. Day-count bases ameliorate some of these problems but create others. Many time series techniques assume uniform displacement, as does much of the math in quantitative finance. Practical solutions are frequently convoluted with loops and exceptions.

Figure 6-8A shows how a time series may allocate more importance to some regimes than others because they are spread over a larger interval. Converting the time axis to bucket the observations by regime (Figure 6-8B) may induce an analysis to assign equal importance to the regimes. Similarly, the distribution of values can be disaggregated (Figure 6-8C) and evaluated by regime or combined (Figure 6-8D). If the modes of the regime distributions are close together (I(θ)), it will be difficult to disaggregate them.

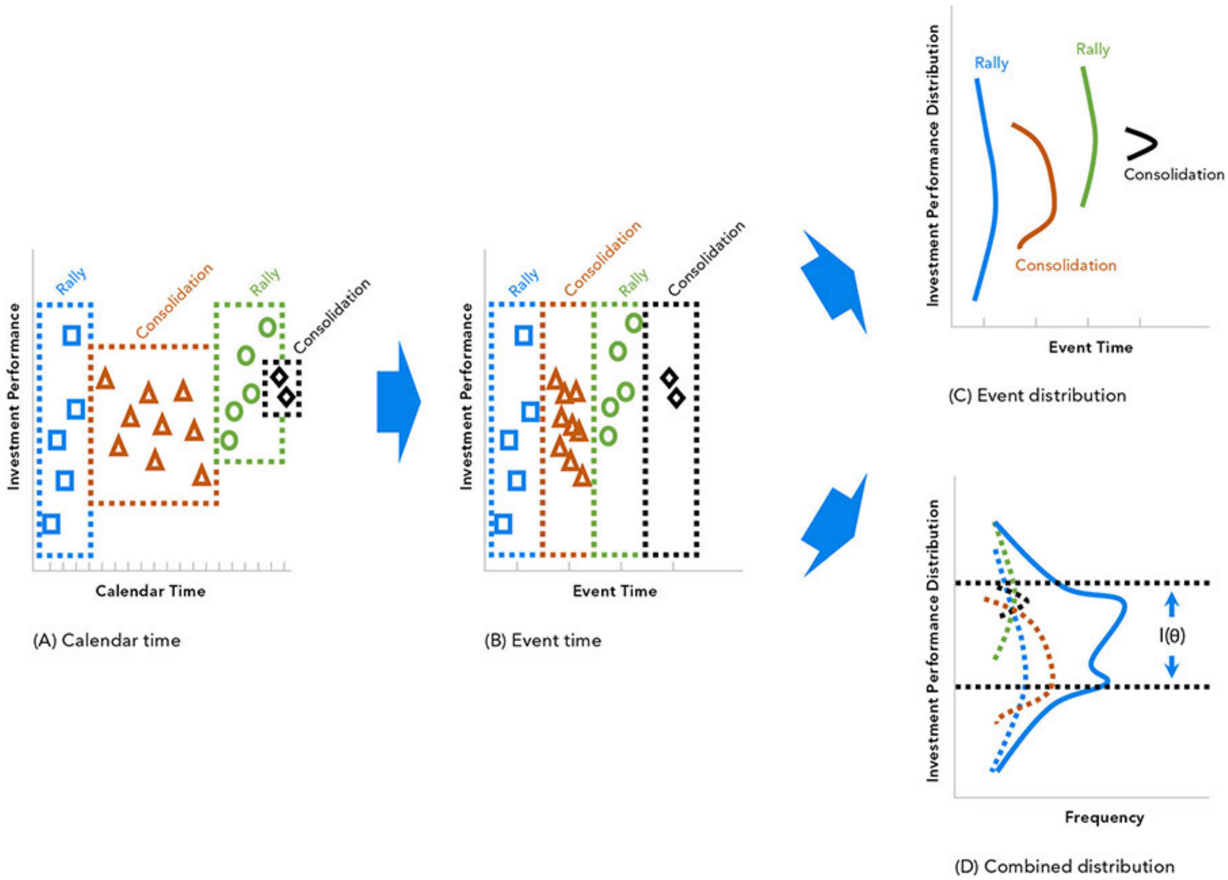


FIGURE 6-8 Calendar time to event time

Dynamic time warping (Figure 6-9) is used to compare multiple series whose displacement variables do not align. The distance between different series' displacement variables for the same event has been used successfully in classification.

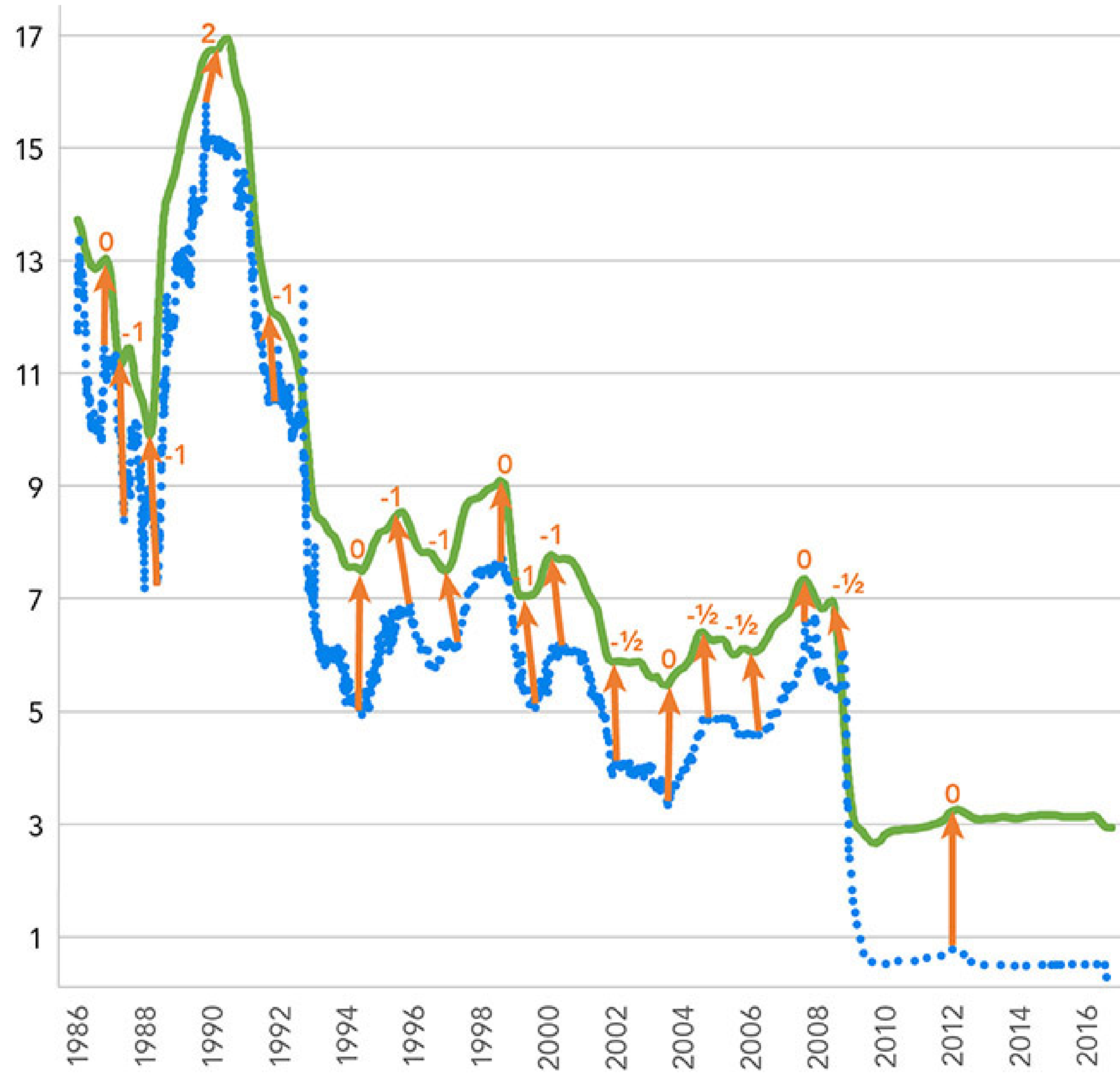


FIGURE 6-9 Dynamic time warping

Many research areas outside of finance have been studying time series both for inference and prediction, and some financial analysts draw on these fields. An operation as simple as differencing can eliminate random noise. Differencing and ratios are widely used and practical transformations. Spectrum analysis, especially Fourier decompositions (which approximate functions as a collection of sines and cosines), can identify seasonality and other cycles.

Tools and visualizations to aid these techniques include correlograms[13] that plot correlation versus lag (including autocorrelation, [ACF] and partial autocorrelations [PACF]), cross-correlograms, cross-lagged diagrams, and unit root tests (to test for more than one trend). Simple operations like differencing can make correlograms more useful, such as to identify seasonalities.

In contrast, methods not explicitly tailored to time series may be harmful. For instance, smoothing data to fill in missing elements may introduce look-ahead bias and cause patterns to be artificially repeated and later recognized as if they were organic. One remedy is to add noise (colored or white) when creating synthetic data. However, the appropriate quality, frequencies, and amplitude of the noise are difficult to determine.

Frequency-related time series analysis has historically used filters (low-pass to remove high-frequency trends, high-pass to remove low-frequencies, and notch to retain middle frequencies). Filters can use a host of functions, including moving average and other smoothing filters, exponential growth, logistic, logarithmic, and power-law. Filters often use lags and moving windows (also known as rolling periods).

To prevent "bleeding" (look-ahead bias), be careful about the overlap between these periods, possibly skipping data in between, and the effect of non-uniformly distributed samples. Beware averages that result in artificial oscillations due to the Slutzky-Yule effect. Similar operations, like emphasizing when a rolling change exceeds a threshold (such as 5 percent per week) or when a rolling change exceeds another rolling change (such as when a 7-day moving average "crosses" a 30-day moving average), are also common.

Conditional Factors

Economic intuition is frequently described using states that can serve as conditions for factors. These conditions can be expressed as categories, probabilities, or Boolean (dummy) scores and can be imposed on a model (high bias) or derived. For instance, values for conditions can be imputed from a decision tree structure by evaluating that factor's importance in the tree.

Economists and historians often write of conditioning based on the business cycle but, to paraphrase Max Beerbohm, the market does not repeat itself; economists repeat each other.[10] Different assets are assumed to behave cyclically as the business cycle progresses. Figure 6-10 illustrates the average valuation for different asset classes over the last five business cycles as defined by National Bureau of Economic Research (NBER) recession dates, not including the 2020 COVID-19 recession. Variation about the average can be extreme, and uncovering relationships from unfiltered economic data can be difficult. Even identifying the current state of the economy using factors like growth, inflation, earnings, monetary and fiscal policy, position, and flows can be daunting, and cycles are only named after they have passed.

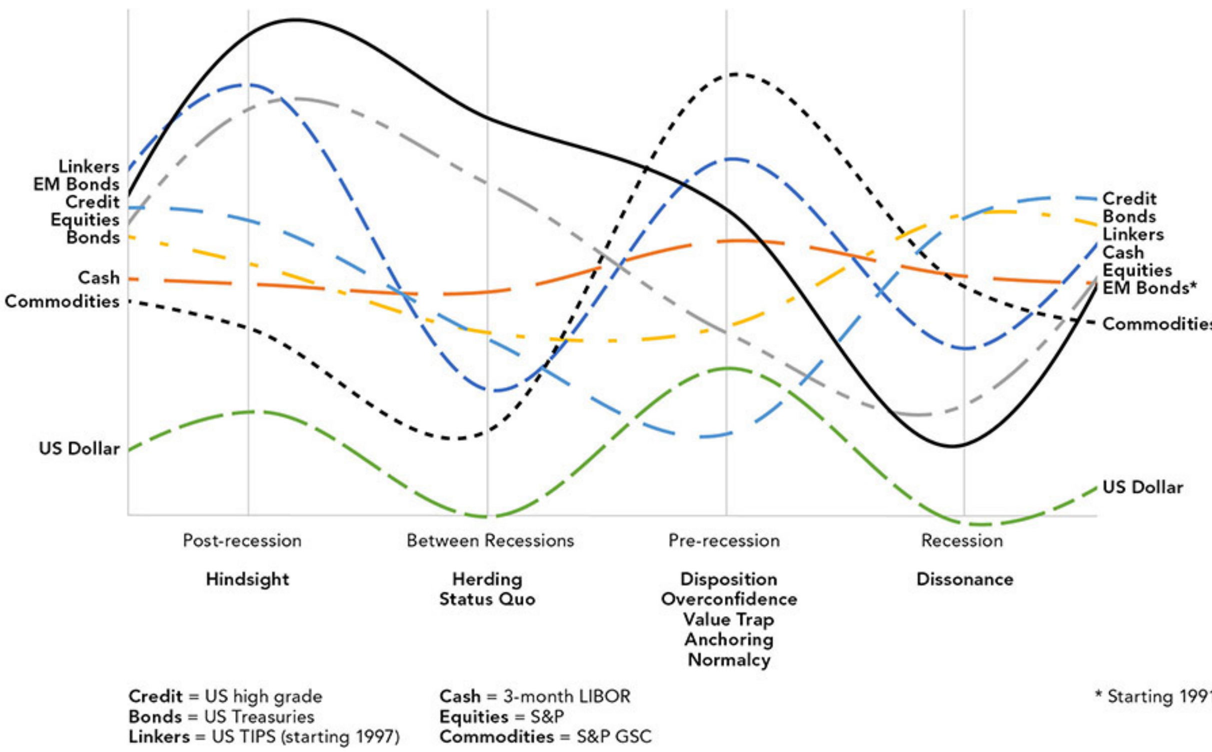


FIGURE 6-10 Average returns of asset classes since 1975 aligned and scaled by NBER recession date

No two cycles are the same, and idiosyncratic factors strongly influence all. There is a common socialized narrative and wisdom—buy financial and industrial stocks and base metals when the economy is growing—but a proven systematic macro investing strategy is much more complicated than it appears.

As we will discuss in future chapters, the business cycle is only one of many conditioning features. The Australian dollar/Swiss franc cross has been commonly used as an indication of risk appetite, long before China became so influential in commodities. The Volatility Index (VIX), the Cyclically Adjusted Price/Earnings Ratio (CAPE), the Baltic Dry (shipping) Index, dividend yields, Treasury-to-Eurodollar (TED) spread, the shape of the yield curve, “Doctor” copper, and crude oil/gold are among many common factors used for this purpose, and each factor can be placed in a cyclical state: peak, decreasing, trough, or increasing.

A conditioned model can combine the states of each factor to predict the performance of each asset. A rising dividend yield and falling price/book and price/earnings are well researched factors for a falling market (and the converse is true).

Identifying predictive factors and their states with enough foresight to act is just part of the challenge. Variations in training samples frequently blur the boundary between what is too early and what is too late, while intermittently conflicting indicators may confound the models.

Statistics and Machine Learning

Statistics and machine learning permeate many topics. While these methods use features as inputs, they are also used to identify pertinent features and combinations of features and to fill in missing data. There are many ways to organize these methods: those that affect a single feature and those that account for interactions, methods that expand the feature set and those that reduce it, simple methods and complex ones, high-bias methods and low-bias methods, methods and metrics, training and testing.

The Bias/Variance Trade-off

While sampling methods can introduce unwanted bias (such as bootstrap) and variance (such as k-fold cross-validation), the bias and variance caused by the model itself can be substantial. A model’s sensitivity to these shortcomings can help an analyst choose which models are most suitable. Perturbation experiments are one way to determine sensitivity.

Less flexible, higher bias models, like linear and logistic regressions and partial least squares, tend to be less affected by variance and are used when theory can guide a model’s structure, while neural networks, nearest neighbors, and trees can be misled by outliers.

Linear and Kernel Transformations

The simplicity of linear models permits a thorough understanding of the theoretical and functional aspects that are not tractable for more sophisticated models. Some prefer combining models by transforming nonlinear features to allow their use in linear models (stacking), aggregating simple models (such as bagging or voting), and feeding the results of one model into the input of subsequent models (such as boosting).

Stacking is like boosting with weights, and may facilitate the preprocessing of features, reducing the computational overhead of the more complex learners. It is sometimes possible to preprocess the most complex learning and create an agile (or linear) process to adapt to online data between expensive calibrations. Meta-labeling is similar to boosting but uses subsequent layers to condition previous models, such as weighting them by confidence.

A simple transformation can involve operations like polynomials and transcendental functions, or the kernel trick to represent a relationship in a lower dimension. The sigmoid function and other methods to separate or condense data are common, as are coordinate transformations, like distance calculations.

Encodings that add information (such as binary to probabilistic) or simplify relationships (such as one-hot encoding) can significantly enhance model accuracy and reliability with modest computational requirements. There are an overwhelming number of combinations including binning, frequency encoding, weight-of-evidence, and information value, standardization, Z-scores, distance from a centroid, and threshold values.

Automated feature transformers, like MATLAB’s transform function, can generate and rank many combinations, but they may not be intuitive, and their predictive power may be spurious compared with expert knowledge. When convincing an investor to trust a model, economic intuition and theory are paramount.

Dimension Reduction and Feature Importance

If domain knowledge (such as GICS) is not sufficient to effectively reduce the feature set, overlapping and non-overlapping clustering techniques (such as simple k-means or canopy clustering using expectation-maximization) or partial least squares (PLS) are the obvious choices for dimension reduction. It is important to note that unsupervised methods are oblivious to the response variable and may not have predictive power, while those that are aware of the response, like PLS, must be carefully monitored to avoid overfitting.

Often, these techniques reduce the feature space and training time but damage the accuracy of the models being trained. The loss in accuracy may be justified by greater efficiency. Supervised discretization and adaptive quantization are also options, but be careful to discard the data used for this analysis prior to training and testing the model.

Using dimension reduction techniques like PCA (including radial basis function [RBF], aka Gaussian, and polynomial kernel PCA), ICA, nonnegative matrix factorization (NMF), or clustering is tempting but comes at the cost of interpretability. They often reduce the effectiveness of analysis rather than enhancing it. Some methods, like low-rank matrix factor approximation, nonnegative low rank, or minimum torsion techniques attempt to strike a balance between dimension reduction and interpretability.

Principal component analysis (PCA) and singular value decomposition (SVD) are often used to orthogonalize (decorrelate) linear factors by choosing factors that explain the most variance in decreasing order. The techniques are limited to second-order dependence, and marginalize important factors that are not volatile. They do not usually work well with nonlinear factors, they are expensive, on the order of $O(nd^2 + d^3)$, and they usually require enough memory to contain the entire data set for inversion (though distributed, map-reduce, and other implementations are available).

The resulting factors are unlikely to be easily interpreted or to be supported by any economic intuition, but the reduction in the number of factors should speed up training times. Though PCA factors are orthogonal, they are only independent when they are jointly normally distributed. Since correlations can vary significantly over time, PCA may change factor choices and ordering frequently and violently, reminiscent of unstable portfolio optimizers.

Reducing outliers with logarithms or other transformations and filters can lessen the erratic nature of the technique. Independent component analysis (ICA) differs from PCA in that it removes high-order dependencies in addition to correlations, and produces factors that are equally important (though not orthogonal). As such, it provides no ordering of factor importance.

ICA is also only appropriate for linear relationships and is commonly associated with the “cocktail party phenomenon”[15] of isolating one conversation in a noisy room. Whitening and zero-phase component analysis (ZCA) combine standardization (unit correlation with a feature) and decorrelation (zero correlation with other features) to remove linearities and are often used prior to ICA analysis to reduce the computational burden. They do not reduce the number of features, since PCA and SVD are unaware of any prior knowledge. Probabilistic latent semantic analysis (pLSA), a case of latent Dirichlet allocation, also called latent Dirichlet indexing, uses a Dirichlet prior in a Bayesian analysis.

When features are nonlinear, it is sometimes possible to reduce dimensionality by making linear approximations of local neighborhoods in the manifold using nonlinear embeddings or manifold learning. Many methods exist for this, including:

- **Linear methods** such as various PCA methods and linear discriminant analysis
- **Projections** including Sammon mapping and kernel PCA
- **Manifolds, multidimensional scaling (MDS), and embedding methods** including locally linear embedding (LLE), maximum variance unfolding, Laplacian eigenmaps, isomaps, and local tangent space alignment (LTSA)
- **Neural networks** including autoencoders with modest hidden layers and regularization or dropout
- **Stochastic and probabilistic methods** such as t-distributed stochastic neighbor embedding (t-SNE) and Gaussian process latent variable models (GPLVMs)

There are many tools for feature importance as well, like chi-squared tests, least absolute shrinkage and selection operator (LASSO), MRMR, neighborhood component analysis, out-of-bag random forest and classification trees, sequential feature selection, and ReliefF and RreliefF PCA ranks components. Some of these methods can rank features.

Modification to methods, like Gower's distance measure for nearest neighbors, or calculating p-values, allows multiple feature types to be compared. Some methods that appear to rank do not do so relevantly; for instance, neuron weights do not provide rank independently, but only as part of the network, though some inroads have been made, such as independent significance analysis.

Reduction Metrics

Regardless of the technique, an objective for determining importance is required, and balancing bias with variance is a challenge, as is imbalanced and sparse data. Confusion matrices are easily biased when data is imbalanced, which becomes clear when they are represented as mosaic plots.

It is sometimes necessary to focus on one class at the expense of others, exacerbating imbalances. Accuracy, as a measure, is easily confused by a dominant class. Likewise, precision (positive predictive value) and recall have limitations and, unfortunately, solve the converse of the problem; recall describes whether the model can determine if an event occurred conditioned on coincident and past data, when what we are interested in is whether events occur conditioned on the model results. These models are only correct if the prevalence is a true representation of the population.

Fleiss' kappa adds a threshold for chance classifications. Some models, like neural networks, behave poorly when the objective function is not continuously differentiable.

It is important to diligently track experiments to identify how results were obtained and to avoid out-of-fold estimation (look-ahead bias) or excessive testing. Excessive testing can be managed using family-wise error and a false discovery rate, and may be addressed with the Bonferroni adjustment (or correction) or similar. Ablation studies work by subtracting and evaluating the quantity and quality of model degradation. Akaike information criterion (AIC), deviance information criterion (DIC), Bayesian information criterion (BIC), minimum description length (MDL), Watanabe-Akaike information criterion (WAIC), and leave-one-out cross-validation ablation (LOO-CV) studies may be used to compare feature importance with model results, using likelihood.

The chi-square test can help determine if the predictors and response are picked at random. Analysis of variance (ANOVA) can help determine the independence of categories. Correlation coefficient and mutual information are also used to identify the differentiation between variables. Gini and entropy (including Kullback-Leibler divergence) are popular for evaluating features, but not appropriate for imbalanced data. Cutoffs are often an issue, hence the ROC and precision-recall curves, and the associated AUC. If all the models use the same assessment sample and the same hyperparameter metric, many comparisons can be made between models, provided the comparison is adjusted for correlation and excessive testing.

Wrappers and Embedded Selection

The preceding methods can be used for large feature sets; wrappers can also be used for this purpose. Wrappers use model performance to evaluate features. While it can be efficient to compare features outside the model, wrappers can determine how they interact in the model.

Wrappers can be expensive, and it is usually necessary to use wrappers on a subset of features. It may be most efficient to use simpler methods to reduce the set, then use wrappers on the most promising subsets.

Financial features are often related in a nonlinear and contextual way. The nonlinearity can be addressed with low-bias wrappers, but interactions between features may remain undiscovered if the feature set is reduced to make the exploration tractable, especially if conditioning requires dummy variables.

Embedded feature selection (also known as implicit or minimum description length, or MDL) is affected through penalization rather than exclusion during cross-validation. Unlike the earlier methods, embedded feature selection is used during the modeling phase. Alternatively, the embedded selection can be applied prior to modeling, and the penalized features can be removed before the final model is used. After selecting features using an embedded method, the selected set can be used in an entirely different kind of model.

A standard embedded method is regularization in conjunction with regression using the *Lp norm* $(\Sigma|\alpha|^p)^{-1/p}$, which is sensitive to centering and standardization. When p is 1, this is called the *L1 norm,* used in LASSO, which can apply zero weights to features. When p is 2, the *L2 norm* (Euclidean norm, least squares, or *ridge regression*) may assign minimal weights to features but will not exclude them entirely.

Ridge regression can be used to manage large numbers of features and correlated features, unlike LASSO. *ElasticNet* is a weighted average of L1 and L2 and works well with correlated features, assigning them the same weights. Variations include the bridge regression.

Other models, such as decision trees (especially random forests), can be used to select features through embedding for use later. It is tempting to use some models, like neural nets, for feature selection, but their preferences are specific to that instance of the model. As a result, their preferences would only make sense in that precise combination. They cannot easily be interpreted in isolation to select features outside of the embedded method for use later.

Imputation

Imputation of missing, *truncated* (clipped), and *censored* values (missing but bounded) is a challenging task, often misapplied to the detriment of the model. The most common method is to remove values entirely, often the within the corresponding data in other predictors. Frequently, the entire row of a design matrix is deleted because of one missing value, as Scikit-Learn does.

Other implementations, like Weka, handle missing values explicitly, but the method employed may not be suitable. When missing values are replaced, take care to substitute a value that will neither bias the model in an inappropriate way, nor ignore the actual value that should have been there; the “missingness” may be structural.

For instance, when assigning security types to assets, all the cash-like instruments may not have values, as may a random selection of other assets. It is common to replace all these missing values with a constant:

- A mean, median, mode, frequency, or a number randomly sampled from the predictor's distribution to match the population,
- A dummy indicator, such as a NaN, random number, or
- A special category, beginning or ending value.

These substitutions often confuse the model, not only biasing the data but also sending a strong confounding signal to the algorithm. Replacing missing values may also turn sparse data into dense data by moving the mean, e.g., from a mean of zero to some non-zero number.

If a significant fraction of missing values were cash-like, a dummy indicator or special category would provoke the algorithm to classify most or all missing values as cash-like. If most values are randomly missing, some of the other methods might provide a weak signal that may lead the model to neglect the values with minimal bias. Adding noise to these values may weaken them as well.

If bias is desired because the missing values are structural, an additional dummy variable to identify it may help. The dummy value would be the only thing that is actually known about the information, and thus has relevance as a signal.

If bias is unwanted, comparing predictions with and without the missing values may indicate whether they are introducing bias. This requires thought, planning, and testing—not a casual replacement or fill. Collinearity, autocorrelation, and risk of look-forward bias make time series challenging to work with.

Imputing should be performed in the resampling phase and before other transformations like standardization, since it may affect the predictors' distributions. Unfortunately, it is challenging to transfer learning from one sample to the next. Since the response variable is the feature itself and must not contain the response of the ultimate model, data analysis does not need to be performed out-of-fold.

To replace missing values with ones that fit a linear model, multiple imputation by chained equations (MICE) attempts to train several models, one for each missing category, in a stepwise fashion. Many models can be used, and logistic regression is a favorite for categories. Wrappers and embedded models can be used with great effect for missing value imputation; nearest neighbors is a common choice, as are trees (predominantly random forests and, for large feature sets, bagged trees), though trees may not extrapolate as well as some other techniques.

Synthetic Data Generation

Synthetic data can be used to solve many problems. Mechanically, financial data must often be modified or combined for use in models. Chapter 4, "Financial Data," discussed futures contracts and other investments that can be combined to create a single return stream. When adjusting data for continuity, as is often the case, calculating cash flows simplifies the process and makes it intuitive—"follow the money." Earlier in this chapter, we discussed oversampling and imputing missing data; these are synthetic representations of existing data.

Some of the fundamental problems in quantitative investing include a lack of data, particularly for specific *stress period* scenarios and rare or hypothetical events. Actual data prevents the ability to conduct repeatable scientific experiments, such as randomized controlled experiments, because the market is behavioral and ever-changing.

Synthetic data generation offers a solution to this. It provides control over the data's assumptions, including its distribution, as well as its probabilities and other characteristics when a great deal of knowledge is missing from actual data. But its strength is also its weakness: the assumptions made are questionable, compounding the already distressing lack of trust in backtesting methods.

While the data can be used for scientific experiments and generated to suit the interrogator's objections,[16] a model relying on synthetic data may not sit well in principle. The sponsor or client may not provide the opportunity for a rebuttal, but rather merely reject the model and refuse to discuss it further. Using historical data may not be more predictive of future events but has the benefit of being a commonly accepted compromise.

A frequently used form of synthetic data, backfilling with a proxy portfolio, can be applied to extend data sets to an earlier inception date. There are many imperfect ways to do this, including substituting another investment before inception, using a basket of similar investments in a hypothetical portfolio, or approximating performance with a model. For instance, a Gibbs sampler could approximate a time series using coincident indicators. Another method is to use a model of the time series itself to generate Monte Carlo trials. There are simple methods of generating stochastic returns correlated with constant drift and volatility, such as MATLAB's Portsim function, that uses Cholesky decomposition and is simple enough to code in Excel. Several promising attempts to use generative adversarial networks (GANs) were built explicitly for time series estimation but their complexity exacerbates the problem of explainability and interpretability for audiences that are not quantitatively adept.

■ ■ ■

In this chapter, we reviewed the critical task of curating features for our model and explored general techniques, including some designed for time series.

Next, we will survey some features used in financial modeling that are specific to this task. Finance has a long history of quantitative analysis and has developed an intuition for features that are at least partially self-fulfilling. Many of these features and techniques were designed with limited computing power in mind. Machines have made new techniques tractable and new data sources both accessible and manageable. It is prudent to study the past with the understanding that we are not so much evolving financial analysis as we are constantly reinventing it, aided by the ongoing creation of new, disruptive technology.

---

8. It is common to inaccurately generate interaction effects when using synthetic data. The venerable Cholesky method is a simple example of an attempt to include interactions (covariances). More sophisticated methods, like generative adversarial networks (GANs), can do a better job. Models often only work well within their anticipated "use cases." Extending models beyond their "envelope" may not work, e.g., using univariate synthetic data in a multivariate model may ignore interactions and produce inaccurate results.

9. Francis, J. Anscombe, "Graphs in Statistical Analysis," *American Statistician* 27, no. 1 (1973): 17–21.

10. Justin Matejka and George Fitzmaurice, "Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing," ACM SIGCHI Conference on Human Factors in Computing Systems, 2017.

11. J. Peter Steidlmayer and Steven B. Hawkins, Steidlmayer on Markets: Trading with Market Profile, 2nd ed. (Wiley, 2007).

12. For a more detailed discussion, see Max Kuhn and Kjell Johnson, *Feature Engineering and Selection* (CRC Press, 2020).

13. Correlograms are also called a periodograms if used with Fourier analysis when plotting power spectral density versus oscillation frequency.

14. “History doesn’t repeat itself. The historians repeat one another,” in Max Beerbohm, “1880,” 1896.

15. Adelbert W. Bronkhorst, “The Cocktail Party Phenomenon: A Review on Speech Intelligibility in Multiple-Talker Conditions,” Acta Acustica United with Acustica, 2000.

16. It can be compelling to generate scenarios that conform to the audience’s assumptions and forecasts. That way, critics would agree with the premise of the simulation if not the entire implementation.

7

Financial and Economic Factors

Isolating the Drivers of Risk and Return

During the feature selection process, the analyst typically maximizes the importance and dissimilarity[1] of predictors. A small and thoughtful set of predictors is best, but many people mine a large number of predictors[2] and their variations. Identifying the most effective predictors is a difficult task when most features interact. Care and domain knowledge can be critical to the model's ability to comprehend salient relationships.

It is important to note that financial time series features require even more care than many other kinds of time series. Economic, archival, market, fundamental, and survey data all demand special attention, especially when transforming and combining variables for feature extraction. For instance, machines often make mistakes that could have been easily avoided by a human. These mistakes may be amplified by the scale and speed of algorithms.

Practical code is ugly, and so are practical factors. Profitable features are niche, ephemeral, technical, and often complex. If the strategy's edge is not structural (e.g., proprietary customer flow may be a structural advantage), then advantage most likely emanates from experience, skill, sweat, and often speed and infrastructure. Many sources of edge are time-consuming and difficult to come by. If the rewards are too great or the access too broad, any potential edge may be arbitraged away or lost to frictions.

By narrowing a model's focus and addressing a market that has a high barrier to entry and modest rewards, it may be possible to create an edge that competitors are reluctant to reach for. Most traders are focused on a specific investment category because it is easier to be an expert in a narrow field. Specialization proliferates feature dimensionality by increasing the number of detailed predictors and inhibits general and comprehensive modeling because of the increased quantity of dimensions, e.g., an all-asset manager may resort to linear extrapolation for nonlinear assets like managed funds and derivatives, while a specialized investor may use more precise factors like Greeks.

Feature Complexity and Hierarchy

Financial features are multidimensional and difficult to organize into a comprehensive hierarchy or taxonomy. Quantitative analysts and strategists must choose an imperfect structure that suits their studies.

Style factors, such as value, may have different meanings based on such features as the economic cycle, geography, industry sector, or investment type. To manage the conditional dependence of these definitions, we can organize these factors into a tree and construct a model that includes a layer for each dimension. Layering is more prone to overfitting than bagging, but the intuition[3] behind the design is superior to that of a more technically sound model that is not based on theory.

Subdividing and isolating features into economically intuitive, digestible, and internally homogeneous subspaces facilitates machine learning by removing confounding factors. Model outputs can be layered to produce actionable and interpretable recommendations. While layering intuitive hierarchical factor categories is one of many methodologies, it has the advantage of simplifying complexity while being easily explained to laymen.

Layering mirrors many non-quantitative processes, so it is easy to justify and accept as sensible and effective. By applying many different models to many different feature spaces, analysts can focus on what works for each circumstance. Examples of this include support vector machines (SVMs) for corporate bonds and random forest for US large-cap pharmaceutical stocks, longitudinal models for regimes, and cross-sectional models for security selection. When evaluating a large and diverse universe of investment vehicles, there is no practical way to avoid elaborate designs, so it is best to embrace the need for complexity and maintain an organized and systematized methodology to manage it. Complex models do not need to be difficult to build or understand.

Even within the relatively simple representation of Figure 7-1, the model can be constructed in many ways, such as multi-factor models for each asset class (equities, fixed income, credit, etc.) and subclass within each class (utilities, energy, financial, etc.). Each class and subclass model may have its own response variable. Often a factor will be interpreted differently in each feature subspace; for instance, "value" may be interpreted as price-to-book for equities, change in yield for rates, and based on purchasing power parity for currencies. Even the response may vary by subspace.

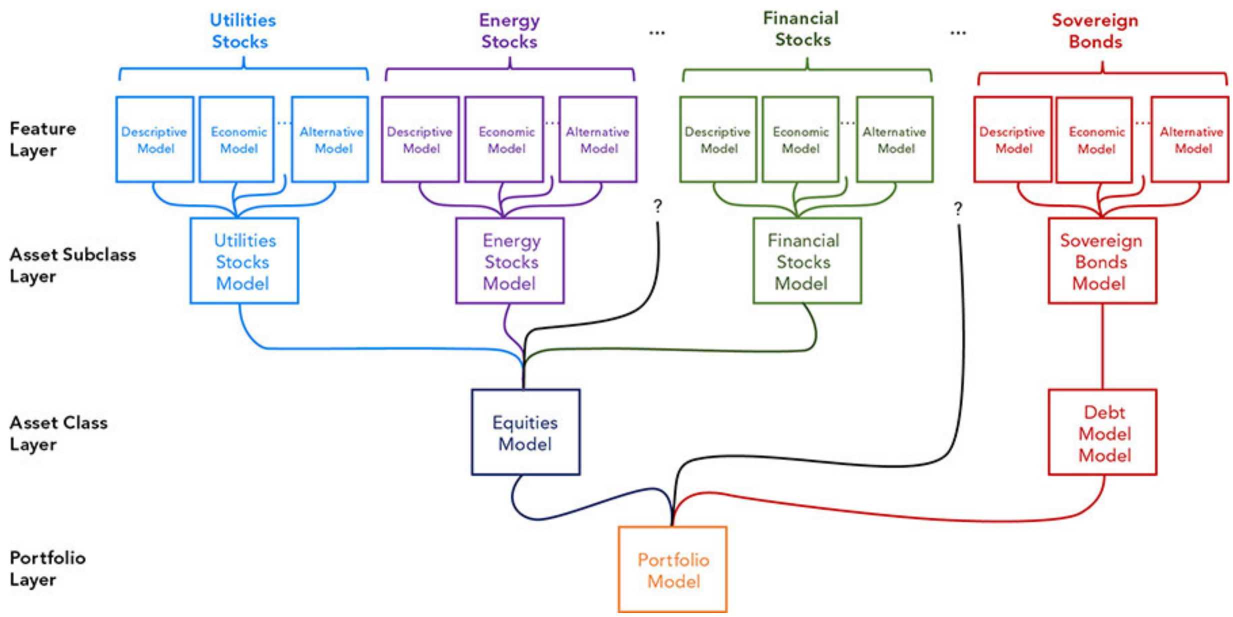


FIGURE 7-1 One way of layering models in an economically intuitive hierarchy

These models do not need to be sophisticated; they can be simple averages of signals. For instance, the stocks versus bonds decision can be made up of an average of consumer sentiment, crude oil, the US Treasury benchmark, gross domestic product (GDP), and forward earnings estimates.

Component indicators are fraught with nuance and idiosyncrasy. Any economist can discuss seasonality, deflators, chain-linking, and a host of other common adjustments they routinely employ for their data. Thoughtful managers apply many transformations before averaging, such as Z-scores, moving window techniques, lags, thresholds, discretization, weights, and percent change calculations.

It is typical for a straightforward predictor to be tortured by transformations. In a relatively straightforward situation, an analyst might enhance an indicator with an 18-month moving average of yearly percent changes and a 3-month lag. She may then classify the results into above 3 percent and below –5 percent. As complex as that may sound, transformation procedures are often far more elaborate and may be intuitive to financial analysts. This complexity of factor specifications can accumulate over several iterations of common-sense refinements. When viewed in their final form, they may seem arbitrary and contrived.

Frequently, managers data mine thousands of indicators and then tune or calibrate them, inviting overfitting. Only then do these managers use intuition to justify the empirical results. We are far more likely to achieve predictive and reproducible results by beginning with an economically motivated thesis and then designing treatments to make the indicators more salient.

Though they may take many forms, one response is usually chosen for each model (such as Sharpe ratio for hedge funds or duration times spread for corporate bonds). In Chapter 6, “Features,” we discussed how to combine models with different objectives, even those of different types. A model’s outputs can be used as inputs into the next tier of the layered models for each asset, and then for the portfolio as a whole. Many analytic methods can be applied to the outputs and when calculating backtest results.

As we discussed with Figure 7-1, a simpler top-down decision tree structure for Global Macro Tactical Asset Allocation (GTAA) may begin with coarse decisions, such as stocks versus fixed income, currencies, and commodities (FICC). As the tree is traversed, each node in each layer evaluates progressively more granular layered decisions that predict the dominance of one asset class, or subclass.

A branch terminates when a recommendation is indeterminate, unstable, or uncertain. The level above the terminator determines the security selection. For instance, if the model cannot decide which emerging markets (EM) country to prefer at the EM node, it terminates and selects an EM index as its security and avoids the country selection.[4]

The form of the response variable (performance forecasting, classification, risk forecasting, etc.) will greatly affect how the features are engineered. Metrics that predict common stock direction may not be suitable for predicting default probabilities of senior debt, even as it relates to the same issuer.

It is easy to tailor predictors to the response in a single model, but when managing a large and complex system of models, it is tempting to generalize and simplify where possible. An analyst may be tempted to use the same GDP input for both a common stock and senior debt model when a more thoughtful and bespoke approach may be required. Or perhaps more granular credit score classifications and thresholds would be better applied to return prediction than a default forecast. For instance, a transition matrix or directed graph (state transition diagram) is commonly used when incorporating ratings data and credit scores, especially when analyzing retail loans.

Many metrics that appear easy to quantify are deceptively behavioral, like measures of risk and return. Behavioral effects can be subtle; for example, forecast momentum is often driven by herding and career risk and can be conditioned on economic regimes. Forecast complacency can be more prominent during expansions and loss aversion (the disposition effect) may be common during contractions. Measures of behavior, like trends and dispersion, can help identify and predict behavior. Agent-based models, such as those that model incentives, can be used effectively to model behavior.

Descriptive Features

Descriptive features identify and categorize predictors by:

- **Class, domicile, and similar categories,** such as asset, class, data frequency, or reporting frequency
- **Geography** (domicile, revenue source, risk source, supply chain exposure, etc.)
- **Business type** (sector, industry, exposure, etc.)
- **Relative metrics** (assets, enterprise value, market share, etc.)

Descriptive data can be a powerful predictor because it is based on domain knowledge. Even if some labels are not precise, there is an intuition behind them and a tendency for others to make decisions based on them, adding to their predictive value through herding. Usually, descriptive data does not vary frequently and serves the same function as “dummy variables” in statistics. While descriptive data may be static for long periods, this data varies occasionally or misleads. For instance, industry classifications or allocation percentages may appear static but usually change over time. Data of this kind is sometimes only available as “current” (the most recent data point) instead of as a point-in-time, or even a historical time series. Analysts sometimes make this descriptive data static out of convenience. For instance, it is common for an investment advisor to try to win a client from a competitor (called a *takeover*). The client often only provides the current portfolio holdings, not a history of holdings.[5] Another example involves analyzing hedge fund risk and performance based on holdings data that does not vary between 13F filings.[6]

Many seemingly static fields, like geography, are particularly nuanced. The origin of an entity can be based on legal domicile or the geography most correlated with risk, revenue, or appreciation. It can even be derived from the company’s supply chain. The purpose of the analysis and response variables should determine how the predictors are defined. Geography most correlated with the returns’ variance may be suitable for some risk attribution, while the geography most correlated with sales may be better for revenue prediction. Frequently, descriptive data is limited to a single value, like the single most correlated geography, when a more complete list of allocations among geographies would be much more appropriate.

Economic Features

Economic features are a favorite classification for fundamental investors who emphasize domain knowledge and intuition. These features include categories like:

- Fiscal and monetary
- Political and trade

- Unemployment
- Production, utilization
- Sentiment
- Inflation

We emphasized the importance of the archival (point-in-time) nature of these statistics and the critical need to address revisions as subsequent data, provoking a response of their own, rather than corrections to prior data. Economic statistics like the business cycle are frequently mistaken as easily interpreted primary drivers of asset valuation and pricing. They are neither easily interpreted nor are they clearly drivers of returns. Economic factors are important, but other more available, frequent, and direct factors encompass much of their effect without many of their disadvantages. Most economic statistics are infrequent and lagged, which makes them applicable for only the longest investment horizons. These long horizons are often a mirage because much shorter evaluation periods can derail long-term plans.

Because economic indicators are frequently described as drivers of return in conversation and the media, they are excellent for persuasive narratives and interpretable models (Figure 7-2). They can be used in a layered model to recommend investment vehicles (Figure 7-3). Extensive feature engineering is often required, such as detrending, volatility adjustment, and Z-scores.

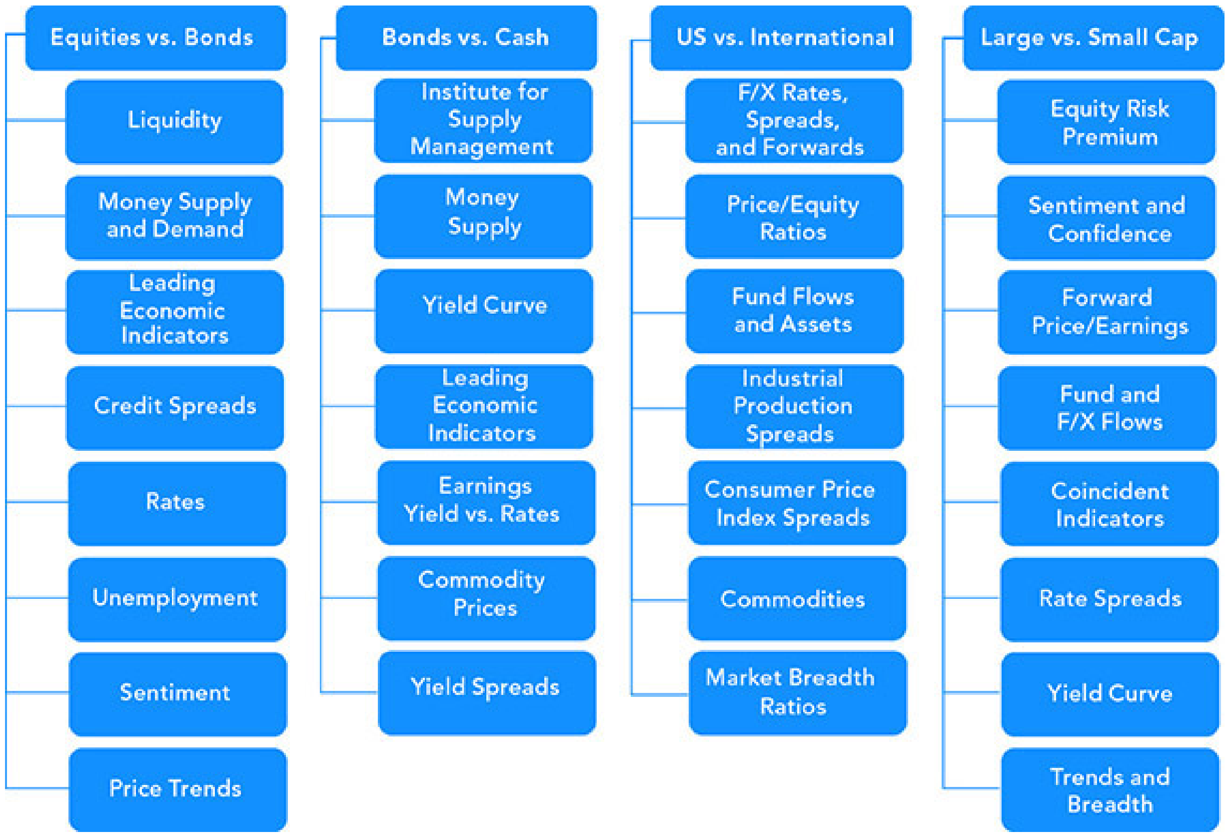


FIGURE 7-2 Paring economic drivers with binary investment decisions

Fixed Income	Cash	Inflation	Expectations	Duration	Credit	Liquidity	Geopolitical
Cash	✓						
CDs	✓		✓				
Inflation Swaps		✓	✓				
TIPS	✓	✓		✓			
Treasuries	✓		✓	✓			
CIPS	✓	✓		✓	✓		
Corporates	✓		✓	✓	✓		
HY	✓		✓	✓	✓	✓	
EM Sovereigns	✓		✓	✓	✓	✓	✓
IG (FX Hedged)	✓		✓	✓	✓		✓
Global HY	✓		✓	✓	✓	✓	✓
Infrastructure	✓	✓	✓	✓	✓	✓	✓

Equity	Dividend Yields	Earnings Growth	P/E	Sector	Style	FX	Geopolitical
Dividends	✓						
Futures	✓	✓	✓				
Indexes	✓	✓	✓				
Styles	✓	✓	✓		✓		
Sector Indexes	✓	✓	✓	✓			
DM	✓	✓	✓			✓	✓
EM	✓	✓	✓			✓	✓

FIGURE 7-3 Factors translate into investment vehicles.

We discussed how default risk models can be improved by simplifying their classes. Separating investments into high-yield and investment grade rather than a multitude of credit ratings can emphasize the salient features of the ratings, making it easier for the model to predict important outcomes. Similarly, economic data can be more effective when used to identify regimes than they can be as predictors.

A great many popular techniques for detrending, deseasoning, and similar actions have been devised, and many economic statistics embody nuances and adjustments that are important to understand. The temptation to use “an algo to rule them all”[7] is strong, but it would be foolish to ignore the accumulated wisdom of generations. Modern methods can analyze more complex relationships than traditional econometric methods. Econometric methods rely heavily on theory, which avoids overfitting and spurious outcomes. Moreover, like fundamental accounting data, many economic numbers are nuanced and require extensive domain knowledge to analyze properly.

Machine learning can find unanticipated relationships by relaxing some of the inappropriate assumptions required by traditional models, but that is not where the largest benefit is found. Currently, the greatest benefit of using machine learning for economic features is in applying revolutionary applications of new techniques with inconvenient data, rather than just incremental improvements using outdated data.

Computers can process and infer conclusions from alternative and extensive data. They can now address the shortcomings of traditional economic methods including latency, infrequency, and inaccuracy. In addition, mobile phone metadata, satellite images, and other invasive data sets are available to track behavior. The technique of nowcasting attempts to forecast economic statistics in real time using data like online purchases. Techniques like the Gibbs sampler[8] can be used to construct missing data points in economic time series using coincident indicators (variables that affect the statistic).

Another interesting area where machine learning can be applied is in causal models, popularized by Judea Pearl[9] and his peers. These models predict cause and effect and not merely tendency. We will discuss causal models in Chapter 10.

Cross-Asset Features

Assets influence each other extensively. These interactions violate the assumptions of many simple models and are an important reason to consider more sophisticated techniques. Nearly all asset classes and financial factors are related to each other and are sensitive to:

- **Economic** features like releases, surveys, estimates, and revisions
- **Commodities** including metals, grains, and energy
- **Equity** prices including global indices, sectors, and geographies
- **FICC** prices and relationships like borrowing rates, term structure, OAS, credit spreads, and currencies

Spanning trees that use distance measures like minimum variance or mutual information are among the many tools used to try to delineate effects. Features that span classes are often used for timing and rotation techniques.

Style Features

While factors are often precisely defined, styles are sometimes more ambiguous. Common examples of style features include:

- Value and quality
- Term structure and carry
- Momentum and mean-reversion
- Growth
- Volatility
- Size

Researchers like Fama and French describe precise formulations for their styles, but their work is frequently described incorrectly by less formal practitioners. For instance, the “small” or “size” factor usually refers to Fama and French’s “small minus big” factor (SMB), which is much more precise than merely favoring small stocks. In their 1993 paper,[10] they describe a procedure involving sorting stocks based on the June median of NYSE companies, taking the average returns of three small portfolios (small value, small neutral, and small growth), and subtracting the average of three big portfolios (big value, big neutral, and big growth).

Implementation details are critical, and playing fast and loose with relationships can make models unstable and unreliable. Complexity may burden a smooth narrative, but precision is essential for quantitative modeling and prediction. Style factors are often only effective under certain conditions or when their values are extreme or dislocated from cross-sectional measures. Complex relationships may confound simple models.

Transformations, like those that identify relative value or trends, often distort or destroy information. Relative, trend, or cross-sectional analyses may eliminate important absolute levels that define inflection points. The “fallen angel”[11] effect is a good example of how some credit transitions of the same magnitude can be far more impactful than others because of their levels. Examples include transformations from investment grade to high-yield "junk" or from late to delinquent.

Dissimilar styles like growth and value or momentum and reversion are frequently paired together. These pairs are meant to increase the opportunity set so that the manager can generate returns in both the "on" and "off" regimes. Often, managers are only effective investing in one of the styles. For example, they may be good at being long stocks but not good at shorting. Test the efficacy of the model in all modes and attribute the performance of the manager under various conditions. However, it is important not to assume that the absence of a positive signal is a negative one, or that shorting is as easy as being long.[12]

Style definitions vary with asset type. Value is measured differently for common stock and high-yield debt. Momentum is significant in many asset classes, but mean-reversion is more prevalent in commodities. Practitioners are often successful because they find a niche where their strategy works rather than trying to find a universal factor or model. Many successful traders focus on a specific subsector of a specific asset class.

A strategy's limitations can be its strength. For instance, a particular niche may provide enough income for a mid-tier professional manager but not so much that it attracts superior competition from top-tier professionals. The "SOES bandits"[13] of the eighties and nineties built their strategy on this concept.

Equity Value and Quality Factors

Value and quality for common equities are some of the most easily recognized style factors. They are commonly thought of as an analog of fundamental analysis and the opposite of growth. There are several definitions that identify low prices and good accounting ratios, such as book value, earnings, and cash flow, but there is no universal definition. Quality, like value, emphasizes strong and reliable earnings, conservative accounting, and low debt, but without the attention to price.

Identifying quality stocks can help investors avoid low-price stocks that may have serious deficiencies. Low prices are often low for a good reason, giving many *deep-value* stocks the moniker *value traps*.

Though evoking a compelling narrative for most people, US value stocks can underperform quality stocks for long periods and with excruciating drawdowns. Though value's mean return can be higher than quality's, it is less dependable with fat tails, importantly large losses in the left tail. Because of inconsistent performance, Morningstar and others have replaced their more traditional factors, such as value and growth, with a more complete menu of factors.

Following are some of the key issues analysts face in attempting to set up an effective model:

Standards. Accounting standards such as Generally Accepted Accounting Principles (GAAP) or International Financial Reporting Standards (IFRS) can be incomparable and may provide significant challenges to cross-sectional analysis. Decisions regarding where and how to record accounting figures may alter measures of value, revenue recognition, impairment, and other contributors to value and quality factors. While liquid developed markets can provide more reliable and accessible data, emerging and frontier markets are where quality premiums are largest.

Emerging markets. Accounting and other governance scandals are far more common outside of developed markets, allowing the quality factor to reduce risk in addition to adding to performance. Ignoring the "E" and "S" of ESG investing is sometimes referred to as standard value and quality investing due to the importance of governance. Quality indices increase risk and return for American and world indices, but increase return while lowering risk for EM. Poor quality EM stocks tend to be so damaging that the risk premium cannot compensate for the added hazard. Unsubstantiated narratives of EM and frontier market opportunities may create unwarranted demand in these markets, enticing investors to invest with a negative risk premium. Inefficiencies may hinder the market forces that would otherwise eliminate perverse incentives.

Agency conflicts. Good governance is important when aligning company management with shareholders, including vetting potentially manipulated earnings announcements. However, analysis can be skewed when companies give confusing signals for certain operations like buybacks and borrowing. Muddying the waters further, analyst recommendations are sometimes influenced by potential investment banking deals. Ratings agencies may be compromised when they are compensated by the companies they rate. False predictions are forgotten along with modest accurate ones, but prophetic forecasts are celebrated and can make careers. These agency conflicts create distorted prices.

Incompatibilities. Accounting ratios do not compare well across geographies or across sectors. Corporations, banks, and utilities respond differently to competition (market share, invention, productivity), operational challenges, debt burden, and other constraints. Ratios have different relevance and are difficult to compare. Asset ratios may be more relevant for staples, industrials, and energy, while income ratios may be more predictive for utilities. A company's maturity also affects the relevance and meaning of ratios, making context as important as regime. Factors may not be ordinal; for instance, costs may be excessive or too miserly and tend to vary widely based on the stage of development, the sector, the industry, or the geography.

Combining factors can cause entanglement problems and may require procedures like orthonormalization (creating a set of vectors that are both statistically unrelated and normalized). It is common to *bucket* investments into categories, like sectors, and then use models to select a subset from within that bucket. By stratifying in this way, the model can ensure that each bucket is represented in the portfolio. As with all factors, portfolio sorts, transaction costs, alpha decay, rebalancing, and other implementation details affect whether a factor is predictive.

Momentum, Volatility, and Growth

Momentum, volatility, and growth are largely behavioral phenomena and tend toward, or cluster about, extreme values (fat tails or leptokurtosis). Momentum exists in behavioral forecasts and other herding activities including price patterns, earnings estimates, supply and demand, etc. Some activities, like technical analysis, can create self-fulfilling prophesies and cycles—both virtuous and vicious—that feed momentum and volatility. Other artifacts like short-selling rules, circuit breakers, and margin limits are designed to inhibit these feedback loops, but they may exacerbate large moves that cannot be stabilized by concentrating pent-up demand (i.e., the system is metastable).

Volatility is present in other styles and factors and is sometimes represented imprecisely as uncertainty. Common characteristics of this factor include serial correlation and clustering.

Mean reversion, sometimes referred to as “picking up dimes in front of bulldozers,” can also be behavioral. Mean reversion is also present whenever an observation is part of a prior distribution. When betting on mean reversion, a steady flow of small profits can be accumulated (selling insurance), only to be lost quickly in a crash or explosive rally.

Gary Shilling famously quipped that “markets can remain irrational a lot longer than you and I can remain solvent.”[14] Being correct is meaningless without the proper timing; many techniques, including Market Profile analysis, attempt to identify whether price patterns are part of a prior distribution or a new distribution in much the same way as Raleigh’s criterion attempts to determine if two dots are part of the same blur or distinct objects.

Carry

The interest rate *term premium* is one of the most popular carry trades. The term premium is the difference in yield between a long-dated asset and a short-dated asset. When the long-dated asset yields more than the short-dated asset, the pair is in contango, and when it is not, they are in backwardation

Structural forces, like storage costs and insurance, make most assets expensive to hold and make most term structures’ natural state contango. Simple instruments, like futures contracts, can contain embedded features that may affect carry in complicated ways. Markets are adaptive and may anticipate these outcomes, so they often price future investments at a lower price than more immediate ones to account for the carry costs. The expectation of a future event may cause investors and traders to price future delivery lower than spot (current). Currency carry, another popular trade, has been suggested to have a persistent carry anomaly.

“The road to hell is paved with positive carry.”[15] As with mean reversion, the risk of explosive reversals is significant for carry trades. It is difficult to exit a bad carry trade gracefully when markets panic.

Commodity carry has both natural and speculative purposes. They can be in backwardation for both manipulative reasons, like short squeezes, and structural reasons, like low oil demand causing producers to park their tankers offshore until prices rise. Commodities producers are influential participants in this market and may have nonpublic information about their operations. Credit carry can be traded using CDS.[16] Though common stock dividends are a form of income, they are not generally considered carry. Volatility carry is usually not considered separate from the volatility factor.

Asset and Market Features

There is no simple way to disentangle financial features from each other. Ultimately, investors and traders must buy and sell instruments, not factors. Asset and market features span factors, including:

- **Single-series** data like returns, price momentum, reversion, and valuation
- **Volume, liquidity, and leverage measures** like bid-offered spreads, credit spreads, and market impact
- **Covariance, autocorrelation, and multicollinearity**
- **Relationships and interactions** like term structure and carry
- **Behavioral indicators** like surveys and forecasts, flow, commitments (speculators versus hedgers), and breadth

Even investments in the same class can respond to different primary drivers. Consider fixed-income credit. Prepayments may be critical for MBS, while pension funded ratio and credit spreads may be important for corporate bonds. Supply and OAS may be common to both (cross-asset factors).

Domain specific features are subject to the same foibles as any other:

- **Stationarity.** Make the features stationary, such as by changing currency denominated factors into percent change.

- **Normalization.** Normalize features, such as by converting a number denominated in shares to currency and then computing percent change.
- **Survivorship.** Pay attention to survivorship bias, such as reassigned ticker symbols. Unique identifiers can solve this problem but can be proprietary.
- **Fidelity.** Note that most transformations reduce the information in the factor and should be provided in an additional feature. For instance, normalizing for currency removes size information, so a size feature should be explicitly added to the predictor set.

Assumptions. Many of these categories violate standard assumptions. For instance, volatility is well known to cluster. This can benefit predictions using special models, like Generalized Autoregressive Conditional Heteroskedasticity (GARCH), but also confound more standard models and joint distributions.

Revisions. It is important to use point-in-time data to separate the economic value implied by the factor from the announcement surprise (such as the earning response coefficient). The long-term economic value and the short-term announcement surprise are similar in effect to the persistent and temporary components of market impact.[17]

Alternative Features

The analysis of alternative data is the most promising opportunity for quantitative research. Traditional data is well structured and often accessible for older techniques. Rather than compete with brilliant minds that have tortured the same data set for generations, researchers can apply more flexible and powerful tools to multidimensional data with nonlinear or latent relationships.

Eventually, machines may outthink talented people head-to-head. Rather than compete directly, machines can currently manage complex and massive data sets. Some alternative data includes:

- **Crowding:** Ownership, put/call ratio

- **Insider and corporate actions:** Domicile, revenue source, risk source, supply chain
- **Sentiment:** Forecast skew, revisions, textual analysis
- **ESG:** Compensation, board age, and turnover
- **Alternative data:** Social media, satellite photographs

ESG investing exemplifies the many challenges common to alternative data. Gleaning signals from this data is difficult, and it is not clear that there is short-term causation between social investing and positive returns. Though studies have shown favorable comparative returns by ESG-oriented stocks, many papers espousing the return potential of “spurned stocks” contrast those that laud the benefits of social investing. Setting aside the economic and scientific arguments, compelling drivers of ESG returns are regulatory punishment and the application of fiscal policy to tax or stimulate businesses.

Competing frameworks have been created to organize and quantify social measures including the Global Reporting Initiative Offers, Sustainability Accounting Standards Board, and Taskforce on Climate-Related Financial Disclosures. ESG data is sparse, infrequent, inconsistent, and without a great deal of history. However, ESG-themed investments enjoy great demand, and investors have shown a tolerance for underperformance—to a point.[18]

Other forms of alternative data are less obscure and have been studied at length, but they are often inconvenient and difficult to use in predictive models. Unstructured data presents significant challenges in many industries. Researchers are innovating, both in how they can organize and consume data and in the assessment of complex relationships, making new categories of data accessible and applying data in ways that were unthinkable until recently.

Execution Features

Many features used to evaluate and predict execution performance are common to other models, such as the seasonality of closing time on Fridays or rolls for futures contracts (Figure 7-4). Other features are more relevant to determining optimal execution, such as *percent of volume* (POV) or *volume weighted average price* (VWAP).

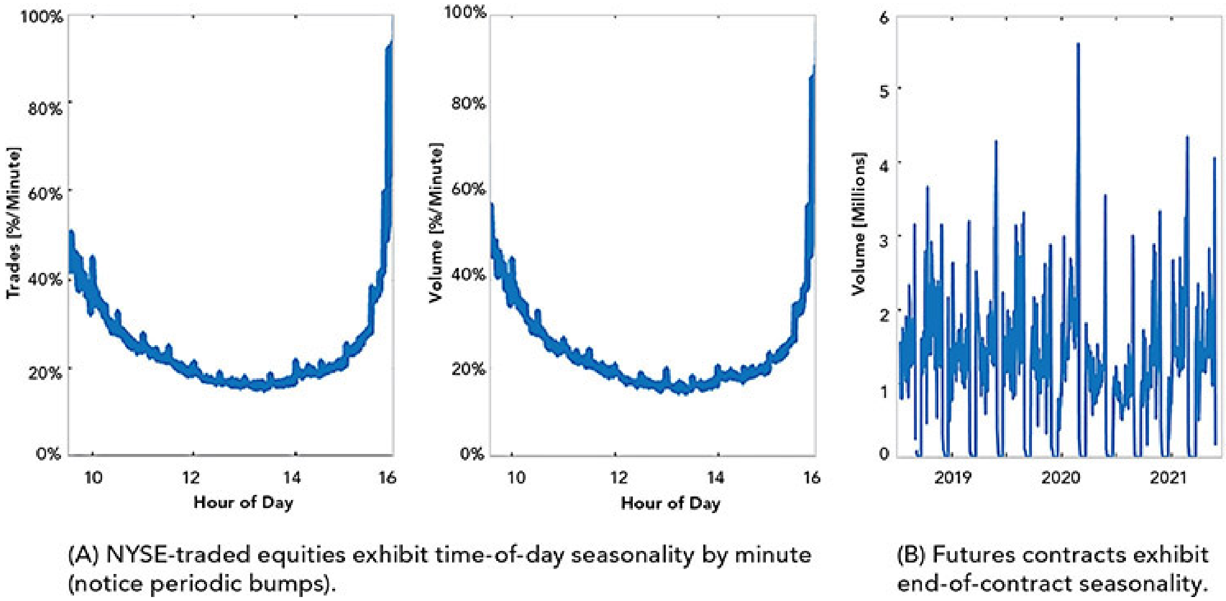


FIGURE 7-4 Seasonality in execution features.

Execution features tend to be more frequent and require manipulation of large data sets. Efficient programming skills and specialized tools may be necessary to pivot, group, join and otherwise manage data that would not be cumbersome in smaller quantities. Old command-line UNIX tools like Aho, Weinberger & Kernighan (AWK), Global Regular Expression Print (GREP), and Stream Editor (SED) can be surprisingly effective, while newer tools like Pandas can underperform.

Techniques for managing data that are too voluminous for one machine to hold, like MapReduce and compression formats like parquet, are often employed for rapid processing. Databases optimized for large time series, like kdb+, are also common.

Feature Conditioning and Timing

A variety of elements come into play in selecting and formulating features. The following are some of the key considerations.

Timing. All decisions involve timing, explicitly or implicitly. Investors are forced to time many aspects of their investment process. Some investors do not believe it is possible to time effectively; they focus on the time spent invested in a factor and try to minimize the effect of the entry and exit decision. However, timing effects are unavoidable. Implicit timers sometimes deny that they are making timing decisions because their buy and sell decisions are often the result of inaction or reaction. They may claim, “I just buy and hold.”

Investors must time their research cycle, asset allocation, security selection, and rebalancing frequency. Less skillful investors should diversify their trade timing and optimize for cost. This is not an endorsement for time diversification; more skillful timers will need more speed and will want to minimize alpha decay. Opportunistic timing, such as making investment decisions that coincide with flows or rebalancing triggers, may reduce costs. Position sizing is related to timing, and both should be managed together.

The relationship between timing and skill, and the necessity of timing, even when it is difficult, is a major theme of this book. Feature timers, who use regimes to time their decisions, sometimes consider themselves to be fundamentalists rather than timers. Even someone who thinks timing is impossible is likely to say something like “It is obviously good to invest in consumer staples now because the economy is sluggish.” These people may consider such decisions “common sense” but, of course, they are timing choices.

Interpretability. Investing is much easier when others understand and support our methods and trust our process (“buy-in”). Arguments against timing are fashionable and difficult to refute.

Regime prediction is difficult, but it is also difficult to avoid when making practical investment decisions. It may be challenging to predict the weather, but no reasonable person would be chided for glancing out the window or checking the thermometer before heading out. It would be reckless not to consider current conditions in relation to historical cycles.

In a counterintuitive way, fundamentalists may favor common sense over scientific study. This is a mistake; biases, like the availability bias or the gambler’s bias, may cause people to misjudge the probability of events.

A reasoned scientific study based on conditioning variables is a viable argument against ignoring the market state or guessing based on intuition.

Regime timing. Regime timing often takes the form of conditioning a predictor based on its state or that of another predictor. Economists use "dummy" indicators in regression models for this purpose; categories can be used in machine learning for the same effect. Regimes may be structural or seasonal. They may be applied to factors that anticipate returns or risk (regression), or to categorize and describe (classification).

Risk factors. It is usually easier to time risk factors than return factors. Risk factors tend not to trend and have larger variance than return factors. There are fewer ways to arbitrage them. Oftentimes, a risky event is scheduled, such as a major decision point like an election or a Fed meeting. The direction of the decision may be difficult to forecast, but the risk of price volatility may be more predictable. It is possible to use a risk factor to generate a return prediction.

Factor timing. Investors often use the market cycle, volatility, or some similar intuitive driver to time factors. These investors determine some factors or asset classes to be preferable or less so based on the perceived state of the driver.

Factor timing may be purposeful and worthwhile, but it is not easy. "Despite the rhetoric calling factor timing simple common sense, it isn't at all obvious that one should value-time factors [due to low frequency data and contrarians]."[19] Evaluating timing is complex and involves trade-offs with cost and decay, as well as reductions in diversification. Individual factor performance can vary with time (for instance, they can vary with economic cycles), and that time dependence may reduce the benefit of timing investments in the factor.

Minority data. It is particularly common and difficult to differentiate major changes from large outliers. Regime shift detection models attempt this, for example by using moving average crosses to identify the beginning or end of a trend. If a major shift is identified, it can be an opportunity to invest or a warning to hedge. Often, shifts that are identified too late, or when they do not exist, induce the worst possible investment choices. Deleveraging or hedging after extreme price movements can cause massive losses if the market "snaps back."

The trade-off between stability, timeliness, and accuracy. Different investors may prefer stability and accuracy (retail clients) or timeliness (high alpha traders). If the stability and accuracy are not reliable, it may be more important to identify *trust regimes*. For instance, we might prefer to use conditions when the factor is reliable (a "target rich" environment), rather than to find the opportunities themselves. When implementing style rotation or investing in multiple managers, identifying when to favor one over another is paramount.

Trust regimes can be identified in many ways including using measures of confidence (significance tests) or signals (strength and persistence). Once identified, these regimes can be used to adjust the frequency and size of investments. When doing so, it is important to discount the influence of chance. Many methods have been invented, including Michaud's resampling technique, to manage the uncertainty in estimation error. Fama-MacBeth regressions can provide standard errors that account for cross-sectional correlations by regressing returns versus the risk premium betas for each factor. Like many linear techniques, Fama-MacBeth regressions do not work with autocorrelated data or correlated factors; focusing on the residuals is a workaround.

Contrarianism. Contrarian timing is the converse of drawdown control. It involves selecting an opportunity to avoid or a bet against an outlier. This may be performed by comparing a factor to recent history and expecting a reversion to some trend. Simple contrarian methods may use a Z-score on the same factor (univariate). Other methods include building a par yield curve and comparing prices to an arbitrage-free investment, but these methods are not foolproof. A Z-score could be wide due to a structural shift, and the par curve could shift, making a poorly priced bond lose its mispricing.

Business Cycle

The business cycle (see Figure 6-10 in Chapter 6) is the quintessential example of a conditioning factor, yet it is notorious. It is both difficult to measure and difficult to use as a predictor. Nonetheless, it is almost universally respected as an intuitive and explainable predictor. Nowcasting has nearly eliminated the hurdle of infrequent economic releases. The remaining problem can be split into two steps—predicting the business cycle and predicting its effect.

Economists put great effort into predicting the current state of the business cycle and its near future. Their estimates can be used in place of a predictive model or a home-grown model using features like unemployment, wages, profits, sentiment, and housing. Some indicators are more timely, forward-looking, and resilient to large revisions than others, but all have their caveats and nuances. Predicting the business cycle is a difficult and well-worn problem.

By isolating these two elements, the stage of the cycle and its effect, the researcher can focus on predicting asset responses conditioned on known business cycles. Based on the current position in the business cycle, asset classes, sectors, and even assets themselves can be selected. Altering investments in response to stages of the business cycle is often called *rotation*, as in factor rotation, asset rotation, sector rotation, etc. Decision trees and layered models (see Figure 7-1) can combine many sub-models, both simple and sophisticated, to produce a comprehensive rotation scheme.

Secular Regimes

Like cycles and seasonalities, secular or structural regimes and trends are important conditions for the analysis process. International stocks diversified portfolios of US equities before the Asian financial crisis of 1997, but less so afterward. A significant decrease in the influence of systemic factors for emerging markets equities occurred without a clear regime shift, illustrating how difficult it can be to identify transitions from one state to another.

Portfolio Sorts

It is important to combine investments thoughtfully. Many techniques are effective; for example, it is common to group assets by their dominant driver using a significance test, like a Student’s t-test; then the assets within each group can be sorted based on their performance. Portfolios can be built with positive exposure to the highest returning assets of each group and negative exposure to the lowest of each group. There are many variations of this technique, including using Z-scores.

Time series require special techniques and involve cumbersome restrictions. If we can consider the time series to be stationary, we can remove the time dependency by transforming the response (Box 7-1).

Box 7-1 Time Dependence and the Decision Matrix

Rather than use a spot return (Figure 7-5A), we can use a periodic return, like a three-year return, and drop the time feature as shown in Figure 7-5B. The Date column is removed by replacing the single-period return for each row with a multiperiod return that begins one increment forward. Be careful to begin the response sample period after the predictors are gathered to avoid look-ahead bias.

Grouping similar assets can net out the variance of predictors and improve forecasting results by combining the rows of assets 1 and 2 and the assets of 3 and 4 for each date (Figure 7-5C). In the figure, the asset indices are replaced by portfolio names, and the operations used to combine the predictors are represented by an empty square, which may be different for different features. If performed improperly, sorting may employ in-sample information, creating look-ahead bias.

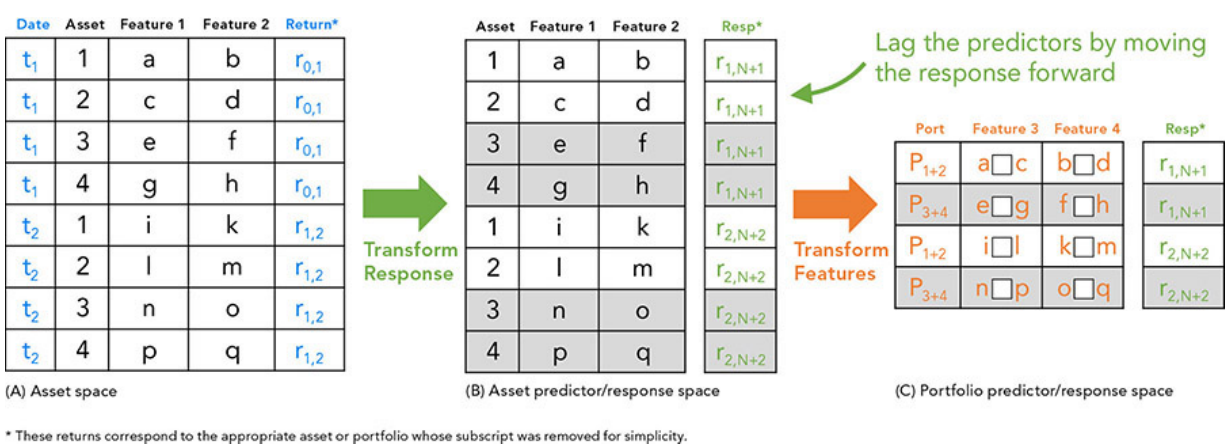


FIGURE 7-5 Mapping from asset space to predictor/response space

Factor Efficacy

As with most financial debates, academic papers argue that factors can be both effective and ineffective in isolation or when allocated, when diversified, and when timed. The many choices include how to:

- **Select** factors (indices, synthetics, dynamic, etc.)
- **Implement** them (factor tilting, pure factors, etc.)
- **Combine** them (naive, optimized, etc.)
- **Explain** and attribute risk and performance

All these choices make studies difficult to compare, and there is a great incentive for profitable strategies to be executed behind closed doors in funds rather than reported in papers. Traders exploit minute and fleeting opportunities, such as fund flows and contract nuances. Factors are often excluded from studies for the same reasons that make them profitable, such as low volume, which discourages competition.

Papers and blogs often overestimate their strategies by neglecting the drag of costs, fees, and other detractors. Naysayers deride active management by not acknowledging the effort and skill required to be profitable after paying those fees, salaries, rents, and other costs.

Mechanistic Transformations

Nearly all investments can be partially explained by some deterministic relationship or causal model. By this, we mean a set of relationships that have a bona fide process for one event to cause another—not a statistical tendency. Usually, these structural relationships can be temporarily interrupted by technical factors like supply and demand, but can be enforced by a pure arbitrage over time.

A simple example would be the return on a US Treasury bond; the return can be realized if the bond is held to maturity but is not guaranteed if the bond is sold prior to maturity. Many more complex examples exist, like the return on a US Treasury bond futures contract, which requires many relationships to be evaluated including carry and the five embedded options.[20]

Even the current pricing of a bond itself, prior to maturity, can be derived from a complex series of relationships using various rates (fed funds, swaps, forwards, bills, futures, notes, and bonds). The rates can be used to bootstrap an interest rate curve and then model the decomposed cash flows of the bond to determine fair value. While the bond may never "clean up" and be priced at fair value, it can be stripped and sold off as parts, or the cash flows can be combined from other sources to synthetically recreate the bond. In this way, the mechanistic relationships may be enforced through arbitrage. Of course, these relationships exist in most asset classes and most instruments.

Factors for illiquid assets as well as the methods required for resampling may need to be adjusted to make up for missing and minority data. There are a number of transformations that attempt to make these data more comparable by adjusting for things like reporting lag, time-weighted versus flow-weighted returns, survivorship bias, appraisal bias, and the effect of low frequency on returns.

■ ■ ■

Financial and economic features are a category of their own with nuances that are idiosyncratic and dynamic, defying generalization. Domain knowledge is valuable and expensive, including the labeling of these data sets for machine training. Good quantitative asset management requires careful study, transformation, and conditioning of financial features.

---

1. Just as modern portfolio theory and the fundamental law of active management emphasize both performance and covariance, features should be informative but different from each other.

2. In some disciplines, the data mining of features is encouraged. Investors have a healthy fear of overfitting and spurious relationships. A parsimonious, explainable, and interpretable feature set nurtures confidence in financial models.

3. Many investors who deride quantitative models that use fundamental factors choose investments in nearly the same way as quants but without the benefits of systematization. Designing and explaining models in a way that is familiar to fundamental analysis may increase the acceptance and adoption of quantitative methods.

4. A case study is available at the website for this book, www.QuantitativeAssetManagement.com.

5. We discussed a similar *survivorship bias* problem in Chapter 4 in relation to the Russell 3000.

6. This was discussed briefly in Chapter 2.

7. “One Ring to rule them all” in J.R.R. Tolkien, The Lord of the Rings, *The Fellowship of the Ring*, 1954.

8. A case study is available at the book’s website, www.QuantitativeAssetManagement.com.

9. bayes.cs.ucla.edu/jp_home.html.

10. Eugene F. Fama and Kenneth R. French, “Common Risk Factors in the Returns on Stocks and Bonds,” *Journal of Financial Economics* 33, no. 1 (1993), 3–56.

11. Fallen angels were discussed in Chapter 4.

12. Shorting is discussed in Chapter 17.

13. Harvey I. Houtkin with David Waldman, *Secrets of the Soes Bandit: Harvey Houtkin Reveals His Battle-Tested Electronic Trading Techniques* (McGraw-Hill, 1998).

14. Often attributed to John Maynard Keynes but more likely first spoken by A. Gary Shilling, quoted in Coke Ellington, “Economist Advises Change in Investment Strategy,” *The Advertiser*, December 3, 1986.

15. The market adage relates to the alure of income that is both ephemeral and can be swamped by depreciation.

16. As with all synthetic hedges and replications, great care should be taken to ensure that the agreement is effective for the intended purpose. For example, Credit Suisse’s additional Tier 1 bonds (known as AT1 bonds, contingent convertible bonds, or CoCos) were written down to worthless as part of their merger with UBS, which surprised many investors.

17. For more, see the case study on earnings surprise at the website for this book, www.QuantitativeAssetManagement.com.

18. Others, especially those with an economic interest in fossil fuels and other businesses that score low in ESG rankings, may pay a premium for non-ESG products or seek to discourage ESG investing.

19. Cliff Asness, “Factor Timing Is Hard,” AQR, March 15, 2017, https://www.aqr.com/Insights/Perspectives/Factor-Timing-is-Hard.

20. Galen Burghardt, *The Treasury Bond Basis: An In-Depth Analysis for Hedgers, Speculators, and Arbitrageurs*, 3rd ed. (McGraw-Hill, 2005).

8

Creating Factor Forecasts

Look Forward, Not Backward

Investment prices are adaptive, dynamic, and based on human behavior. Investors anticipate the future and adjust prices to reflect their forecasts, and opportunities result when an investor's forecast differs from the market's forecast in a way that can be monetized.

We can all agree that forecasting is difficult. Nevertheless, as George Cox famously said, "All models are wrong, but some are useful."[1] The goal is not to be correct, but rather to be useful. With sound money and risk management, even a slight edge can be lucrative when appropriately exploited.

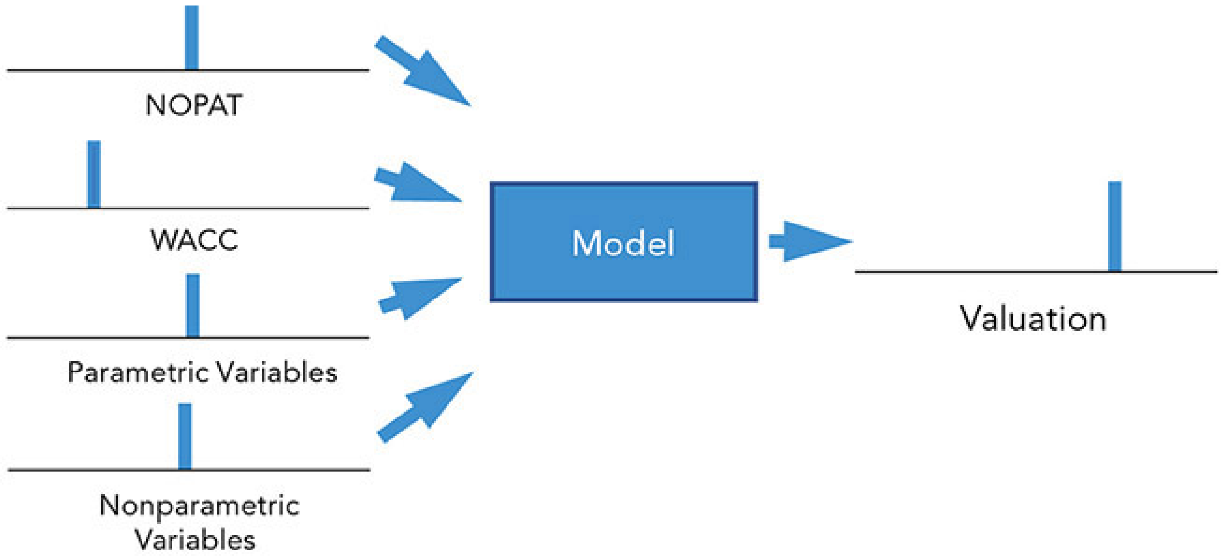
Past and near-current factor values can sometimes be used in a backward-looking fashion to determine the best way to invest. Backing out valuations requires history to be a reasonable estimate of future values. Although market implied estimates are rarely good enough, they can be a good starting point.[2]

Estimation error can overwhelm complex methods so that the most confident result may be overfitted. As a result, while forecasting methods can be advanced, they are often unsophisticated (though complicated) in practice, such as a rolling, exponentially weighted moving average using shrinkage with a three-year or five-day window, depending on liquidity. We previously discussed how simple and intuitive refinements can culminate in a complex and obtuse expression like the one in the previous sentence.

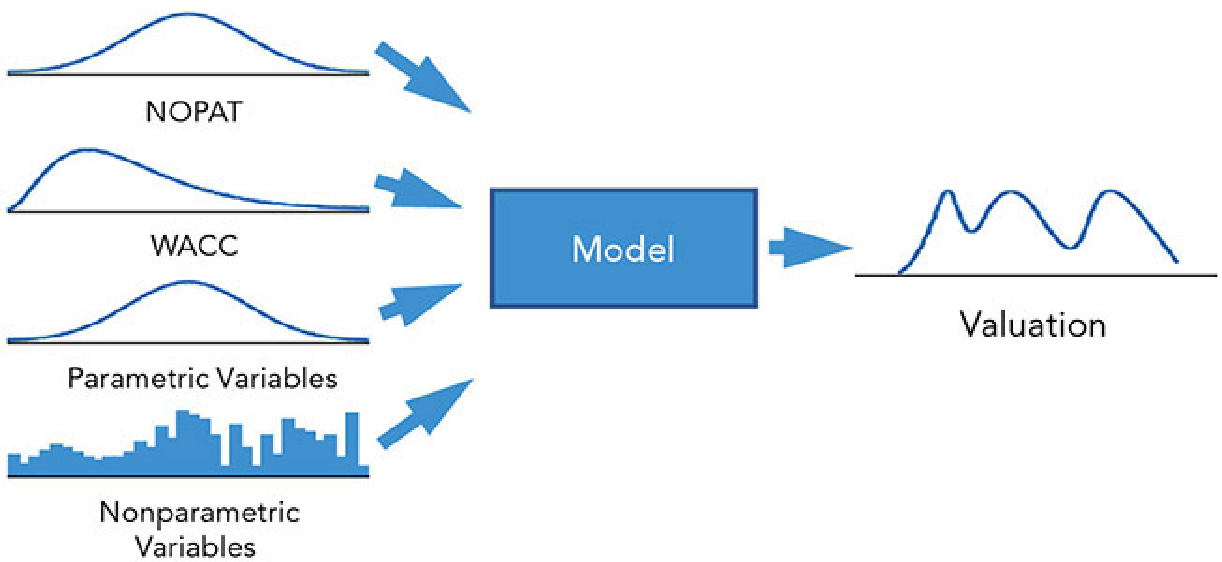
Researchers can be creative when trying to make historical data forward-looking. Sometimes that may mean blending time-series-based risk premia for individual factors with cross-sectional averages. In certain cases, indirect methods may be necessary, such as approximating volatility by simulating a strategy that shorts S&P 500 puts.

Though extrapolation (sometimes called momentum or "leadership" in marketing-speak) is the most common assumption when applying past data to future estimates, there are times when the results are antithetical. Mean reverting markets (and the tendency for managers' future performance to correlate negatively with past performance) are apparent exceptions to momentum. The influence of good and bad luck tends to diminish as time passes. High prices tend to result in poor performance.

It is better to anticipate future values than to rely on history, and better still to anticipate a distribution of future values for each factor. With factor distributions, an analyst can derive a forward-looking distribution of expected returns for the strategy (Figure 8-1).[3] There are many opportunities to adjust the model inputs to make them forward-looking, incorporate the company view, or test scenarios. Figure 8-2 highlights a few of these and places them in an investment framework.



(A) Point estimates producing a single output



(B) Distributions as inputs producing a distribution as an output

FIGURE 8-1 Point estimates of inputs carry little information; distributions contain much more.

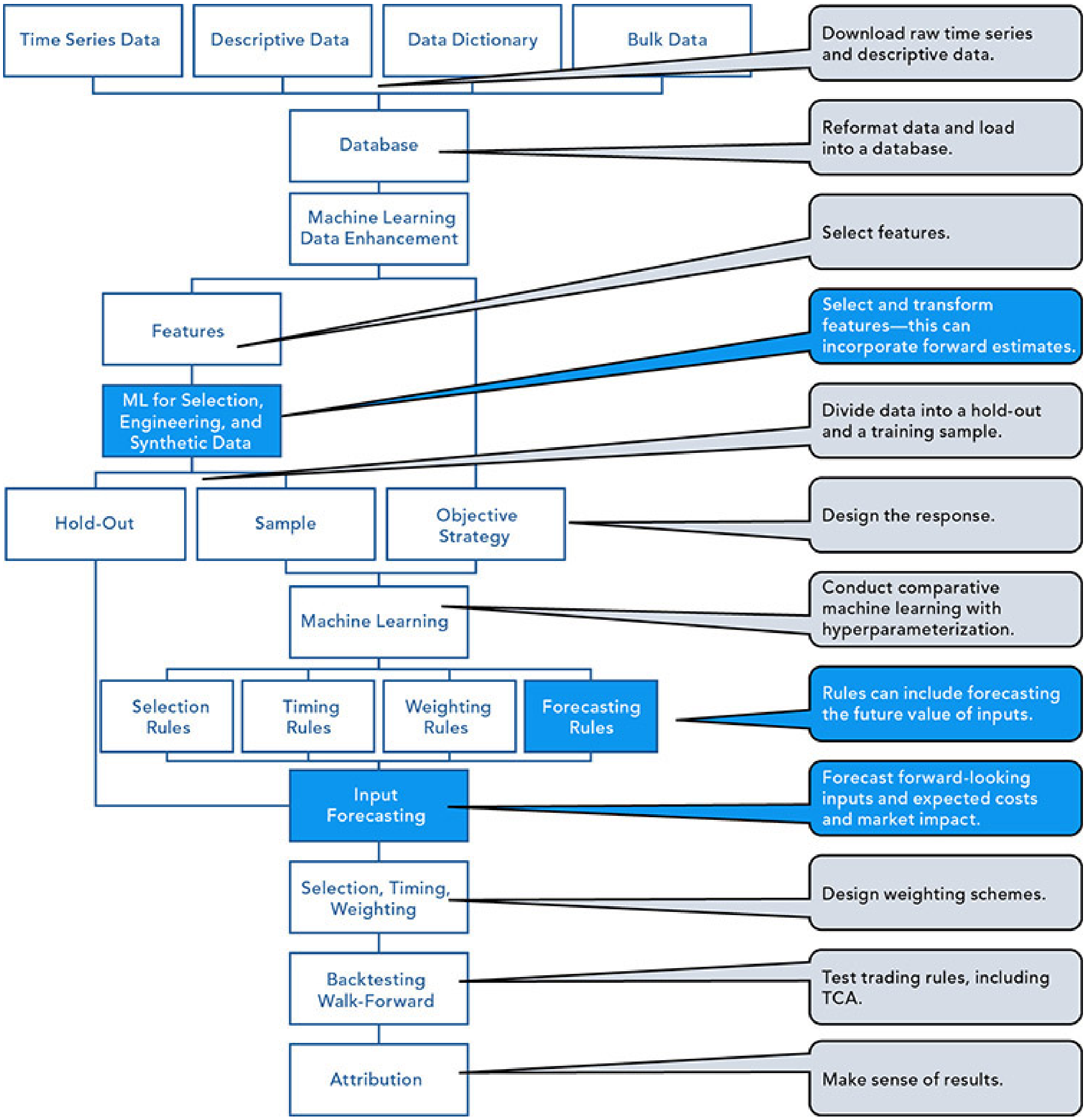


FIGURE 8-2 Factor forecasts can be employed in many stages of the investment process.

Systematic forecasting can be part of the feature engineering process. If it is, be sure to include it in the sample as well as the hold-out.[4]

Ad hoc, subjective, unanticipated, or manual forecasts can only be applied later by an analyst, the portfolio manager, economists, and investment committees to improve the information used by the model.

Many processes require subjective inputs that are difficult to predict, even in hindsight, such as the house view of the parent's investment committee.[5]

If the input is part of the process and can be estimated with some accuracy—with some foresight or behavioral heuristics—then it is better to model the ad hoc input imperfectly than to ignore it. Although going "off script" can be irresistible, meddling has caused the failure of many systematic funds, so modeling with and without deviance can be instructive. Sensitivity to subjective inputs (board mandates) or downstream modification (investment advisor adjustments) can create guardrails and warning flags for the implementation team.

It may be possible to incorporate prior views as a factor (such as by using historical capital markets assumptions as a data series), but it may be impossible to know how future views will differ from systematic forecasts. A good example of unanticipated subjective views that surprised investors occurred when Fidelity Magellan changed management (discussed in Chapter 2 and Figure 2-4). While forecasts and views are essential as inputs for valuation, alpha, and risk forecasting, they can also be used to influence asset allocation calculations by adjusting inputs to naive models or by using view-aware models like those based on Bayesian inference.

A useful paradigm in investing is to simplify the analysis. We may reduce the set of accessible opportunities and then isolate interesting and actionable relationships. If it is prudent and affordable, we can hedge or diversify the unwanted risk. These components are usually discussed as the impetus for investment decisions but can also be determined by reverse engineering asset allocations or investment performance. One way to do this is by calculating implied values and reverse optimization, as we will discuss later in the book.

In this section, we covered:

- **Assets**—nuances of our data, the sources of data.
- **Features**—how to make data more salient.
- **Financial and economic factors**—how to identify some drivers of risk and performance. We can accomplish this statistically through dimension reduction, classification, and regression. We also identify drivers structurally with economic concepts and categories.

In the pages that follow, we will discuss how to develop our factors further to make them forward-looking.

Capital Markets Assumptions

In making assumptions regarding the capital markets, here are three key concepts to keep in mind:

- Leverage resources
- Do not “reinvent the wheel”
- Pay attention to competitive advantage.

Economists make forecasts and periodically publish their capital markets assumptions, outlooks, and surveys. The skill and cost required for these projects are substantial, and most investors can benefit from using forecasts produced by more well-capitalized institutions.

Economists, like most financial professionals, may appear to prioritize accuracy, but often their incentive may actually be popularity.[6] Accuracy and precision can be ranked for external or internal forecasts. Internal ranking is sometimes called *alpha capture*.

Third-party forecasts can ward against being “alone and wrong.”[7] But relying on external forecasts makes it more difficult to find opportunities because competitors have access to the same information. It is often better to be different than smarter, especially when sales are more important than repeatable and reliable performance and memories are short.

Information may be insufficient or prohibitively expensive. Both retail and large investors often find it necessary to produce their own research. Our experience is that once a retail investment company has about a billion dollars under management, they become large enough to institutionalize themselves and hire a proper chief investment officer. Institutions usually have a house view, and some impose the view on their managers. In other cases, manager views and proprietary trading research are distinct from treasury, institutional, and retail research.

For business reasons, most client-focused companies feel compelled to offer their own research to their clients to differentiate themselves from competitors, even if it detracts from the investment value of the research. Performance-driven companies with a strong culture and a rigorous scientific process are self-evident and do not require this “window dressing.”

Economists’ forecasts are usually infrequent. Many methods have been devised to update forecasts between reporting dates. Bayesian methods of updating prior beliefs with new information are effective, and causal models are coming into their own.

It is possible to create valuable forecasts from seemingly irrelevant or insufficient information using methods similar to bootstrapping a yield curve, solving simultaneous equations, copulas, or relative constraints for optimization. Simply rank-ordering forecast relationships in a pairwise fashion to create an incomplete distance matrix can produce impressive results.[8]

It is usually easier to estimate a spread or change than an absolute level. For example, it is easy to answer the question, “Is Los Angeles farther from New York than Chicago?” But it can be more difficult to answer, “How many miles between Los Angeles and New York?” Expert systems, alpha capture, and meta-prediction can be used to assess and combine many predictions from different sources. Alternative data, including web postings, online retail prices, ship tracking, satellite images of parking lots, and the shadows of petroleum storage tanks are proven ways to enhance and update forecasts. The Billion Prices Project, begun in 2008, is an excellent example.

An effective forecasting methodology requires diligent measurement and assessment. There are many appraisal methods, like the Diebold-Mariano test. Parameter uncertainty, dispersion, and other population measures also provide valuable insights into predictive skill and reliability.

Strategic Forecasting

Not surprisingly, strategic forecasts influence strategic allocation decisions. They follow long secular trends (or supercycles), such as globalization, aging populations, and the influence of technology. These themes do not need to be exogenous; life planning, such as the need for large purchases like a home or education, can influence strategic plans.

Strategic decisions can relate to the inherent nature of an asset, such as structural covariance, mechanistic relationships, volatility, or countercyclical trends. They should reflect core beliefs and through-the-cycle estimates. Emphasize characteristics that are invariant when averaged over the strategic horizon. Intra-horizon relationships, like covariance, should be the focus of strategic forecasting.

Long sample periods can be misleading or inappropriate. For instance, regime changes and assessments of evaluation horizons are commonly miscalculated. It is unclear whether the nature of the observations has changed over this long period. Long histories invariably encompass many disqualifying flaws, including:

- **Regime change.** An example of this would be whether currently developed markets had evolved from emerging markets at some point during the period.
- **Estimation bias.** This refers to a variation in quality, such as estimated or sparse data for some portion and highly liquid data for others.
- **Survivorship bias or a compromise to control it.**
- **Weighting scheme.** This can entail either ignoring imbalances or imposing a bias, such as exponentially weighting by time, to minimize discrepancies.
- **Framing.** The inception date can have a significant influence on conclusions.
- **Government intervention.** Fiscal or monetary policy can affect results.
- **The dominant trend.** Is it linear or bottoming? Interest rates rarely descend below zero, recent history notwithstanding.
- **The relevant trend.** Certainly a 700-year trend does not apply to a practical investment portfolio, and even a 60-year average is far too long for an investment manager’s review period.

For shorter strategic periods, long cyclical effects may be included in the strategic forecast. Broad economic variables, like inflation, may be considered cyclical, but their variation may be slow enough to influence strategic decisions. For inflation, the mid-1960s through the early 1980s represented one regime with hourly earnings increases ranging between 5½ percent and 8½ percent. The mid-1980s through 2020 experienced a range of about 1½ percent to 4½ percent. Within those longer trends, smaller cycles can be examined for tactical decision-making. Minor influences can become significant over long periods. One-way drags like inflation, fees, and taxes can have a devastating effect on returns and may vary greatly over time.

Economic variables are not the only features that are used in strategic analysis. Market variables, like commissions and bid-ask spreads, and data categories that are typically considered relatively stable, such as geography or credit rating, can and do change over strategic periods.

Relationships can be resilient, as with bond autocorrelation, or they can be difficult to forecast, as with stock autocorrelation. Even relationships that are taken as a rule of thumb, like default recovery rates, need to be considered variable for strategic forecasting purposes.

Some relationships, such as correlations, are related to other variables, like volatility and liquidity. Such indicators often respond to unpredictable events. Rather than attempt difficult forecasts, transformations like shrinkage are sometimes used to make models less sensitive to estimation errors and large variances.

Tactical Forecasting

Forecasting the future direction of the current state of events is an irresistible topic for most investors—even those who deride quantitative prediction and attempts at timing. As opposed to trading, tactical investing is often founded in the prediction of external cycles and themes. The business and profit cycles are two of the most salient drivers of cocktail party conversations and tactical decisions. Like most qualitative approaches, fundamentalists' typical "weight of evidence" argument is just a less precise version of Bayesian analysis. These forecasts (quantity, variance, and confidence) are then used to rotate sectors and asset classes, select investments, or tilt policy allocations within the allowed bands or ranges.

Risk premia are merely the marginal return (premium) that investors are expected to require for a specific contribution of an investment's risk. Some examples of risk premia driving allocations across asset classes include:

- The impact that economic growth may have on earnings and valuations. Fiscal and monetary policy and other economic indicators may induce a manager to alter a company's exposure. Inflation may make tangible assets more attractive.
- The impact of changing interest rates on discount factor estimates. The shape of the yield curve may influence fixed-income decisions.
- The influence that cash balances, credit spreads, and carry may have in making credit investments more or less attractive.
- The effect that political and natural events may have on the supply and demand of commodities.

Forecasting these states is only the first step in crafting a tactical investment method. The investment market's incessant adjustment of prices in anticipation of future events is one of the reasons that applying feature forecasts can be challenging. It is not enough to forecast an event; we must also forecast the reaction to that event.[9]

Sometimes the target allocation will be a strategic goal, and bands allow deviation for tactical bets or implementation realities. Policy portfolios and their band sizes are often determined through a formal process and may require approval by a board; frequently they are written into the Investment Policy Statement (IPS).

Risk Premia

Risk premia exploit the assumption that rational investors will not increase their risk without compensatory incentives.

Reckoning. A variety of forces may determine risk premia. Broadly speaking, there are five ways to determine value to offset a marginal increase in risk:

- **Surveys** of market participants, academics, or the general public.

- **Historical returns** and economic indicators, which are ready sources of data, as in corporate bond yields minus Treasury bond yields. They can be used in statistical models without structural influences. The literature tends to conflate these ex post risk premia with ex ante risk premia. While this simplistic view has its uses, we are interested more in identifying the forward-looking expectations implied by the market without the noise and pollution of market forces.
- **Supply-side models** based on variables like inflation, income (dividends), and growth (earnings).
- **Demand-side models** that use investor appetite to determine the required return.
- **Hybrids** that combine several of these methods.

Challenges. The methods may overlap. Returns and economic data are used to determine both supply and demand-side models. All these models share challenges that are common to financial and economic time series. The theory and the supposition that these incentives are additive are debatable, but the approximation is useful. Many other challenges remain unsolved. For instance, nonstationarity can make long histories confounding rather than supportive. The usual confounders apply, such as:

- Seasonality
- Trends
- Publication bias
- Autoregression
- Multicollinearity
- Irregular distributions
- Nonstationarity

As a result, estimates can be scattershot. Academics continue to seek a unified theory, such as linking the *equity risk premium* (ERP), bond spreads, and real estate cap rates, making a challenging problem even more difficult by trying to integrate all possible cases. By leveraging the additive nature of risk premia, Ibbotson and Seigel[10] devised a convenient and intuitive "building block" methodology. They describe most assets' expected returns; for instance, the risk premia of a cash investment is comprised of inflation and the real riskless rate of return. Stocks and bonds require an additional “bond horizon premium.” Stocks enjoy an “equity risk premium,” as well.

A rational investor requires more compensation for premia that are wrong-way (like illiquidity) or are particularly unpleasant (like taxes). Behavioral influences like risk aversion, economic conditions, and investor wealth add to significant data challenges. Most data comes from survivor populations (like the United States[11]). Long periods include vastly different characteristics, such as emerging market factors for the US in their infancy, before survivorship is assured, and develop market characteristics after that transformation.

The risk premium puzzle. When academia accepts a theory, like risk premia, and reality does not cooperate, it is audaciously termed a puzzle or a paradox. The shortcomings of these models and estimates form this puzzle, and there is no good answer. However, when pondering the changes in risk premia, these models provide a framework and common language for prediction.

Types of premia. As with behavioral heuristics, a risk premium has been named for nearly every source of return variation. We will discuss a few of the more common ones using the building block framework for convenience. As with all things practical, the hierarchy shown in Figure 8-3 quickly becomes fractured and intertwined with crossover products, like convertible bonds, and interaction effects. Few of these premia are independent, and their effects are likely nonlinear. Nonetheless, the building block method is a useful organizational tool and framework for estimating influences when making uncertain predictions regarding risk and return.

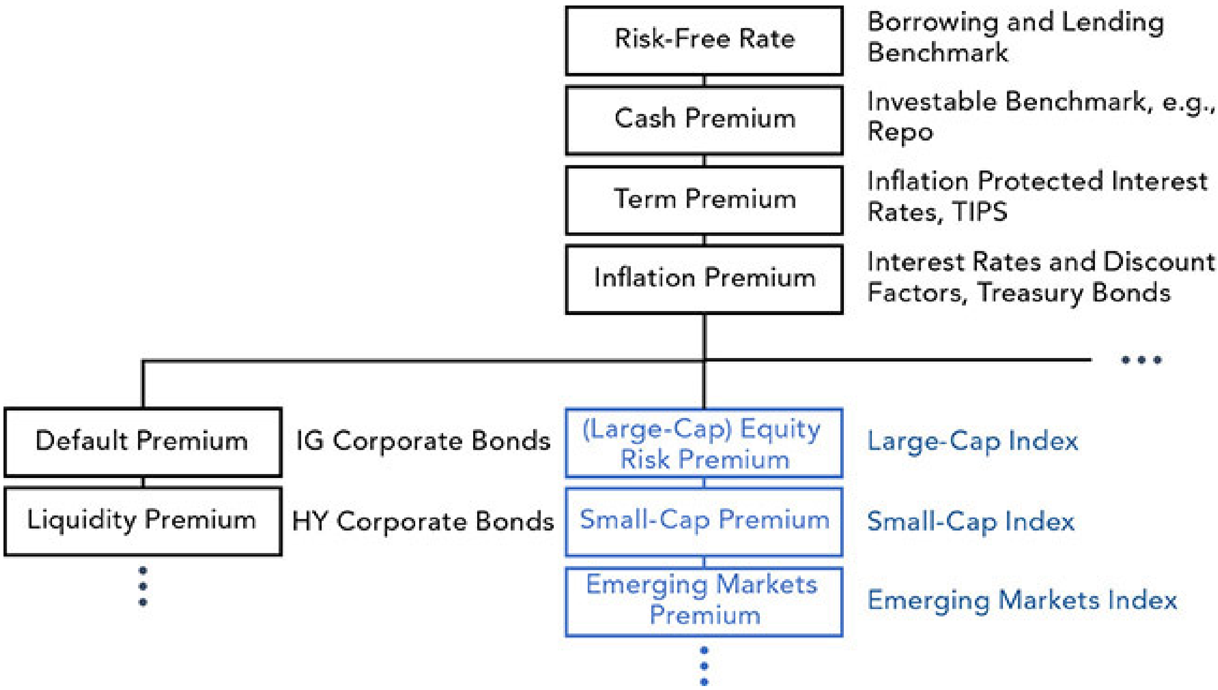


FIGURE 8-3 The beginning of a risk premium hierarchy tree

Fixed-Income Premia

Several factors determine fixed income premia. Following are some of the key elements that should be considered.

Risk-free rate. Because discounting is intrinsic to all valuations, the foundation of all valuations includes the risk-free rate, which is highly variable. As fundamental as it is, the risk-free rate is a source of three controversies:

- Some argue over the *validity* of its riskless assumption, particularly when political brinkmanship threatens default.
- Another contention is the *tenor* of the rate (which should match the investment horizon), since the risk-free rate should have no reinvestment risk.
- *Deflating* real rates also creates choices and confusion.

Term premium (interest rates). When discounting cash flows for valuation, we need to know the rate for different borrowing periods.[12] To expand the risk-free rate to a risk-free curve, we estimate the *term premium*. If the method for determining the risk-free rate is well defined, such as using US sovereign debt with formulaic adjustments, the term premium may be less contentious. It is not trivial, however, because actual investment vehicles are polluted with a multitude of idiosyncrasies like:

- **Supply** (issuance and stripping)
- **Demand** (on-the-run versus off, special dates like quarter end)
- **Day-count** bases and holidays
- **Specialness** (idiosyncratic borrowing rates)

Some of these concerns and others were addressed in our discussions on valuing the roll (Chapter 5). Using a method called bootstrapping, we can generate term premia by combining disparate investment vehicles, removing biases, and decomposing the rates.[13]

Unlike equities, interest rates have defined payment schedules. As a result, their variance is explained mostly with three term structure or yield curve parameters:

- **Level**, yield, or returns (PC1)
- **Slope** or term (PC2)
- **Curvature** (PC3)

Duration, similar to term, is a primary risk factor for rates. Notes of different durations can vary dramatically in risk. Investment advisors frequently oversimplify portfolio risk measures by describing them in terms of percent equity.

Though large-cap developed market equity indexes have similar factor categories, fixed-income indexes do not. Fixed-income indexes vary significantly with duration and credit spread. These attributes also vary with time. A percentage allocation, like 60 percent equities / 40 percent bonds (60/40), cannot adequately explain the risk of a portfolio, and unwitting or unscrupulous investors can dial their risk up and down without alerting their clients, e.g., by changing their fixed-income benchmark.

Duration is a weighted average of cash flows, which may include income and principal. When policy rates were at historically low levels, most interest rate products paid little income and did not have the cushion to rising rates that these instruments relied on previously.

Credit premium. A top-down analysis of rates is an integral component of any fixed-income analysis, but credit influences also require a bottom-up valuation. Macro models might avoid accounting for intricate security-level details, but a specific analysis of the issuer, and often the offering itself, are crucial to investing in credit. Success in credit investing can hinge on the legal interpretation of nuanced details, such as the definition of a credit event.[14]

Default premium. Typically, credit is determined by a series of technical directed cyclic credit events that may culminate in default. Default is the most definitive state and is the primary target for modeling and valuation. *Expected loss* (EL) is often modeled in three parts:

- **Exposure at default** (EAD)
- **Probability of default** (PD), often expressed as a transition matrix
- **Loss given default** (LGD), or 1 minus the *recovery rate* (RR)

Note structure. Note structures may involve *credit enhancement* or some other contingent financial engineering construct. These structures may collapse suddenly and unexpectedly creating a *jump to default*, which must also be modeled.

The performance of an individual investment depends on the features of its structure. Coupons, embedded options, covenant strength, and a multitude of other features make valuation nonlinear and often counterintuitive.

Many models have been devised to link default intensity to credit spreads, including *distance-to-default* models that focus on book value (because it is owed to creditors) rather than market value, *market-based* models, and *hybrid* models that join accounting-based approaches with market-based models.

Ubiquitous embedded options make volatility an important factor in valuing credit. *Option-adjusted spread* (OAS) is a common measure of ex-post premium. Credit is a wrong-way risk, and so credit-risky investments are expected to outperform realized defaults during normal periods and take painful losses during crises. Though spreads usually overestimate the loss due to default, profits can only be realized if spreads collapse or the bonds are held to maturity. Market segmentation (i.e., fallen angels) gives BB credits the greatest edge, while long-dated corporates can fall short.

Duration times spread (DTS) is a predictive measure. Credit default swaps (CDS) can be a proxy for credit spreads, although there are issues with applying spreads to specific issuances and evaluating defaults from causes other than bankruptcy without access to proprietary data. Providers like Moody's offer such data in a point-in-time format.

Modifications also complicate credit spread valuations. Lenders can be burdened even though the default process can be averted through modifications.

Style factors are used by some investors to forecast the credit premium:

- **Value**, such as option adjusted spread (OAS), yield to maturity (YTM), yield-to worst (YTW), and zero-volatility spread (Z-spread)
- **Quality**, such as leverage, free cash flow (FCF), credit rating, probability of default (PD), loss given default (LGD)
- **Size**, such as debt outstanding for an issuer or for a particular issuance
- **Volatility**, such as OAS volatility, DTS, yield volatility, duration times yield
- **Momentum**, such as transformations on price or yield changes

Equity Premia

The equity risk premium (ERP) is the marginal reward required by investors to bear the marginal risk of buying the market portfolio of stocks. The market portfolio is challenging to define, as are the other premia that make up common equity risk.

Uses of the ERP. The ERP is important because it may be used for:

- **Cross-sectional** analysis, such as comparing the compensation for risk for asset allocation

- **Time series** analysis, such as timing investment entry and exits of equity class investments
- **Business decisions and corporate actions**, such as those that require calculating a company's weighted average cost of capital (WACC) to evaluate real investment opportunities and dividend and buyback policies
- **Institutional planning**, such as asset/liability studies for endowments and pensions
- **Life planning**, such as such as evaluating the relative efficacy of Roth versus traditional Individual Retirement Arrangements (IRA)[15]
- **Government policy**, such as such as pricing for utilities

Economic foundations. It is well accepted that economic growth, specifically gross domestic product (GDP), directly influences common equity returns and the ERP, although GDP is a lagging and unreliable indicator. With some adjustments, we can make the relationship clearer.

Other fundamental influences include liquidity and flows; policy and policy uncertainty; behavior and risk aversion; inflation; everyday risk, drawdown risk, and wrong-way risk; and information flow and accessibility.

Confounders. As important and useful as the ERP is, it is elusive. Confounding influences include:

- **Dilution.** New shares increase the denominator for earnings and dividends, driving a wedge between profits and GDP.
- **Retention.** The percentage of profit used to pay labor, management, and investors (dividends) alters the relationship between revenue and GDP.
- **Irrelevancy.** Multinational companies tend to be the largest by market cap, which can diminish the specificity of even the most seemingly diversified investments.
- **Hidden.** Private companies contribute to GDP but not the market portfolio. "Unicorns" that grow from minuscule efforts to giant corporations have a significant effect.
- **Adaptive expectations.** While GDP is backward-looking, equity prices anticipate growth, making them a moving target.

- **Regime change.**[16] It is easy to dismiss long histories encompassing the emerging United States, the War of 1812, the Civil War, and the Great Depression.
- **Survivorship.** ERP estimates are usually based on the most liquid markets that have avoided the worst problems. Long histories reveal devastating market losses in Argentina, China, Egypt, Germany, and Japan.
- **Price/earnings (P/E) expansion.** Many investors formed their ERP assumptions during the last 20 years of the twentieth century, during an unusual period of rising P/E ratios (from around 7x in 1979 to over 35x in 2002) and falling rates (from around 17 percent in 1981 to about 2 percent in 2003). While innovation in technology (both traditional and financial) is often used to explain why P/Es would naturally rise, it is useful to recall that, while advancement was much faster and much more revolutionary in the beginning of the twentieth century than at the end, P/Es remained rangebound for most of that century.
- **Agency and momentum.** Estimates tend to rise quickly and fall reluctantly, even when they exceed GDP, which is likely a ceiling for the ERP during normal periods. Even when impartial,[17] most estimates heavily weight recent past values rather than mean reversion, which may be more appropriate for a long-term market portfolio.
- **Ad hoc adjustments.** Many developments are ad hoc. For instance, Ibbotson ignores historical P/E growth because they do not anticipate continued growth.
- **Artifacts and nuances.** Sensitivity to the sample period (especially the start date), the effect of survey design and population, statistical artifacts, the effect of sample frequency on error rate, quirks in behavior like comfort-seeking, and complexity like taxation rates are just some of the details that make ERP estimation difficult.

Other Premia

While many researchers try to unify theories on premia, others study more detailed and nuanced premia. Niche opportunities are the mainstay of many investment strategies. Academics, in search of primary or “true” factors, sometimes refer to the abundance of mostly interrelated or compound factors using Professor Cochrane’s term “zoo of factors”[18] or, more commonly, “factor zoo.”

Disjoints. Factors that permeate many types of risk, like leverage, affect some assets differently than others. Private equity analysts are concerned with whether the financing is venture capital (VC) or leveraged buyout (LBO). Real estate investors may be concerned with construction risk, maturity, commercial/multifamily/single detached, etc. Oftentimes, and especially in private investments, funding risk is more important than liquidity risk since it cannot be endured.

Interactions. Inevitably, premia share is determined by drivers and other influences with their basis in a common economy and behavioral tendencies. Interaction effects are inevitable, and only unintuitive statistical factors can ensure independence.

■ ■ ■

After overcoming the challenges of financial data, teasing out the salient features, and isolating the independent factors, these drivers must be forecast to account for the anticipation of the market participants who determine investment valuations. Skilled economists freely offer many estimates, but it can be difficult to adapt estimates to the multiple time frames and objectives of the investment policy.

Analytical frameworks, like the concept of risk premia, provide a basis for discourse and analysis. Unfortunately, they also present significant hurdles if taken too literally. Risk premia are difficult to agree on, both theoretically and practically. Expert estimates of their values are almost humorous in their dispersion. The implied overconfidence of forecasters[19] may often be attributed to their professional need for attention and credibility, but consumers of their advice need to be wary of putting too much faith in them. As always, it is critical to remember that the primary task is finding a useful approximation of the market and applying it effectively within its limitations.

---

1. George Cox, *The Journal of the American Statistical Association*, 1976.

2. The "wisdom of crowds" is often cited as a reason for considering implied values as an expected outcome. Even if this wisdom were predictive, is not always reflected in the market. Valuations may not be anyone's estimate, but rather a muddling of the conflicting motivations of many diverse market participants.

3. Oracle's Crystal Ball is a brilliant and simple tool that allows an analyst to define distributions as inputs to an ordinary spreadsheet, sample those distributions for use in a Monte Carlo analysis, and collect the results to produce a distribution of outcomes.

4. Recall our discussion of adjusted versus unadjusted prices from Chapter 5. If the feature forecasting process is included in the backtest as part of the system, rather than including forecasted features as inputs, the model can be trained on raw data and the resulting performance will be more reproduceable.

5. Adding colored noise can simulate ad hoc subjective changes, skewing the result to helpful or disruptive.

6. A chief economist at a bulge bracket bank once told us that his compensation formula explicitly included the number of unsolicited phone calls he received (as a measure of his popularity and contribution to brand equity).

7. In a canonic principal-agent problem, analysts and investors often prefer to risk being wrong along with many others rather than make bold choices and risk being criticized for the decision. Echoing this sentiment, the saying "Nobody ever got fired for buying IBM" has been familiar to IT professionals since the 1970s. Interestingly, the opposite is true of some forecasters who rely on the salience of outlandish and correct predictions and the forgettable nature of incorrect ones. A forecaster who correctly predicts an important event may be remembered for it despite many incorrect predictions.

8. A canonical example of this is to use a table of pairwise distances to produce a surprisingly accurate geographical map, such as the *Tabula Peutingeriana* constructed from *itineraria* circa 1200 AD. More modern examples include using multidimensional scaling (MDS) to turn lists of pairwise distances into maps.

9. Matt Levine, "Knowing the Future Isn't That Helpful," Bloomberg Opinion, November 26th, 2019.

10. Roger G. Ibbotson and Laurence B. Siegel, "How to Forecast Long-Run Asset Returns," *Investment Management Review*, September/October 1988.

11. US capital markets were not always developed. An analysis that ignores survivorship bias may lead a researcher to overvalue risk premia.

12. Practically speaking, different curves are required for borrowers, and lenders and for borrowers and lenders with imperfect credit. Multiple curves for different reference rates like the London Inter-Bank Offered Rate (LIBOR), overnight index swap (OIS), and Secured Overnight Financing Rate (SOFR) complicate matters, as do various adjustments like credit value adjustments (CVAs) and debt value adjustments (DVAs)—and that's just for US dollar rates!

13. A case study on valuing a par yield curve is available at the website for this book, www.QuantitativeAssetManagement.com.

14. Recall our previous discussion regarding the seniority of Credit Suisse's CoCos when Credit Suisse merged with UBS.

15. Often called individual retirement accounts but technically called individual retirement arrangements by the United States Internal Revenue Service (IRS).

16. We are referring to market regimes, not necessarily political ones.

17. Explicit and deliberate overweighting is also common and may be accomplished using methods such as exponentially weighted moving averages (EWMAs).

18. John H. Cochrane, "Presidential Address: Discount Rates," *Journal of Finance* 66, no. 4 (August 2011):1047–1108.

19. Some people argue that professional forecasters are incentivized to predict spectacular outcomes because their success will garner accolades while their failures will be forgotten. Many investors read commentary and outlooks for their interesting observations and ideas, and not their forecasting accuracy. Modest and uninteresting forecasts would not attract much of a following from these investors.

9

Strategy, Objective, and Conditions

What Are We Trying to Achieve?

Factors and their transformations are necessary to form the predictor variables for our analysis, but they are not always sufficient to inform our model. Modeling a strategy may require us to apply complex transformations and conditions to our response variable and to add structure to our objective function. This chapter focuses on predicting the result of the strategy rather than forecasting future values of the predictors.

Forecasting predictors versus response. Analysis involves representation, prediction, evaluation, and optimization. Machine learning can be vital in forecasting predictor values (forward-looking independent variables), and often involves complicated transformations that clarify the response. Adjustments to the strategy response are designed to increase accuracy when compared to live trading, but may obscure the signal with excessive noise. It is important not to confuse the models with overly complex responses. While predictors are used to simulate the state of the world at the time of a forecast, the response (dependent variable) is used to describe the successful forecast for training purposes. So, we must carefully guard against *lookahead bias* when constructing our predictors and use future data when building our response variable.

The response. A critical element of prediction, evaluation, and optimization is defining the response function, also referred to as the *objective* or *scoring* function. The purpose in designing a response is to offer a parsimonious and tractable goal to the algorithm to enable it to learn. The objective condenses the strategy (which may include conditions describing complex sociological rules like herding or panics) into a formula and allows the model to combine predictors to produce the result.

Responses are the examples that supervised models learn from. The analyst must manage the trade-off between realism and tractability. If the response is too nuanced, the model will become confused and fail to learn. If it is too simple, the response will not match the outcome of actual trading and the model will fit the wrong function. We can and should use a backtest and forward walk to simulate the realism that is impractical to implement in the response. This allows us to use a tractable response without sacrificing accuracy.

Naivete may produce good results. While the models predict the result using various techniques—including arbitrage, factor models, events, and microstructure analysis—the response merely calculates the result in a convenient form, preferably a single statistic like a Sharpe ratio or rolling return. The choice of which expression to use for the response may seem trivial until we consider the multitude of options. For instance, it is common to choose a 1-day, 5-day, or 30-day return for a response; but how? We must strike a balance between being quick to calculate and predict, being realistic enough, and being simple enough not to confound the learning algorithm. Oversimplification can invalidate a model, but because of the time and effort required, a faithful response is often impractical.

Realism. Identifying an investment strategy that works on paper and not in real life is the quintessential problem of backtesting. For instance, it is not enough to predict returns if our portfolio is benchmarked. Without a response that incorporates tracking error or some other relationship with the reference rate, an inconveniently timed loss could leave us "wrong and alone,"[1] which is much worse in the eyes of our investors than experiencing a drawdown along with the rest of the market.

Complexity. There are many ways to construct the same response. For example, a response can be more or less *continuous* (returns for regression); it can be *discrete* (categories for classification, like profit and loss);[2] it can be *binary*, buffered, or manifold; it can be *European* (with a single ending value), *American* (path-dependent), *Bermudan*, or a more exotic structure involving a series of decision dates.

Modeling a live strategy subject to human overrides is an involved process, but simplification of a successful strategy can result in a reduction in predictive power. Unlike many modeling exercises, the goal is not to predict the power of a factor, but the performance of a “living, breathing” investment vehicle. Tracking error, risk tolerance, reaction time, and many other elements are important. Even the sampling frequency (time of day or month of year) can introduce nonstationary characteristics, such as “window dressing” and other seasonalities. Yet, ignoring high-frequency performance can be unrealistic as well. Spikes in data (including prices, volatility, and wrong-way risk) can lead to the untimely end of some investments due to breaches in risk tolerance, margin calls, and other realities.

Integrating the factors. Some analysts will focus on factors in the modeling phase, while delaying behavioral simulation and complex strategies until after the factors’ efficacy has been determined by the model. When this is the modeling strategy, constructing the response is a much simpler and cleaner task. Separating modeling from backtesting adds *scientific rigor* and reduces the possibility of overfitting.

Nonetheless, separating the factor and response model is not ideal. In the same way that separate and distinct risk and alpha models can introduce discordant and incompatible recommendations, separating the factor models from their response will train factors that are suboptimal for the ultimate purpose of operating an effective fund. The analysts may become uncomfortable and force the manager to deviate from the model. Ultimately, the manager may be faced to decision to “pull the plug” on the model. This happens during periods of intense emotional stress and uncertainty, which Errol Morris termed “the fog of war.”[3] It is often more practical to design a system to incorporate these decisions, along with a game plan for as many foreseeable situations as possible.

Tractability. For effective forecasting, the response should be unbiased and well behaved (not jagged or discontinuous). If the response does not violate many assumptions, it may allow a broader arsenal of algorithms to be applied. The response can be designed for more interesting, nuanced, and conditioned prediction at the risk of confounding the algorithm. A well-designed response can also reduce processing time, enabling a more comprehensive and thorough analysis.

Experimentation. Different models use their responses differently. The best combination of factors to use in producing our response may not be intuitive; try several formulations to see what works best. Sometimes the simplicity of a machine learning algorithm causes it to respond differently to similar transformations.

Backtest and forward walk. Portfolio optimization typically employs a utility curve to help define the objective. In backtesting, the problem is more nuanced. Thankfully, backtesting does not require a simple function to represent the objective. Backtesting software can be arbitrarily complex; the only limitations are modeling skill, processing power, execution time, and interpretability. Modelers should incorporate intricacies like multiperiod execution, transaction costs, and complex hedging strategies in the backtest.[4] The responses for this analysis must be simple and smooth enough for the algorithms to learn from, but the backtest is not restricted in this way.

Moving target. The disconnect between the machine learning response and the backtesting result is what sets quantitative asset management apart from many other machine learning exercises. The model is predicted using a simplified response and may not be trained for a result that will produce effective backtest performance. This is why the backtest and the subsequent attribution analysis is necessary; backtesting serves a similar purpose as animal and human trials in drug design.

Ideally, we will predict:

- **Selection** (which investments to make)
- **Direction** (whether to buy or sell)
- **Size** (how much of each investment to buy and sell)

Determining direction may not be practical, and shorting may be difficult to identify and implement. Shorting may also be expensive at the worst times. In addition, a long-only portfolio, or one with limited shorting, may be a constraint imposed on or by the investor or manager. Size determination may be constrained or confounded due to risk (both everyday risk and catastrophic risk), uncertainty, confidence in the predictions, momentum, liquidity, transaction costs, and other details, like an explicit policy mandate. Figure 8-2 (see Chapter 8) outlines the various stages in modeling a strategy.

Removing Time Dependence

Time series present problems for general purpose models, including most machine learning models. Financial time series are problematic because they usually violate assumptions like stationarity and autocorrelation. By stripping the time element from our data, we can broaden our palette of models and increase the amount of data available to train them.

In Figure 7-5, we demonstrated how to strip the time dimension from signal prediction by making the response encompass intervals. We showed this for individual assets and groups of assets (portfolios). Figure 9-1 expands on this to illustrate a technique for avoiding look-ahead bias. One way to do this is to avoid using data (embargo, hold-out, gap, etc.) between the feature set and the response set. This will help avoid autocorrelated data from affecting the prediction (bleed, contamination, diffusion, etc.).

Figure 9-1A shows a timeline representation of how a lag, L, can be used to insulate the data used to determine the features from the data used to determine the response, including an observation period, $t_{0,1}$. The observation period represents the naive response, say, a one-period return, $r_{0,1}$.

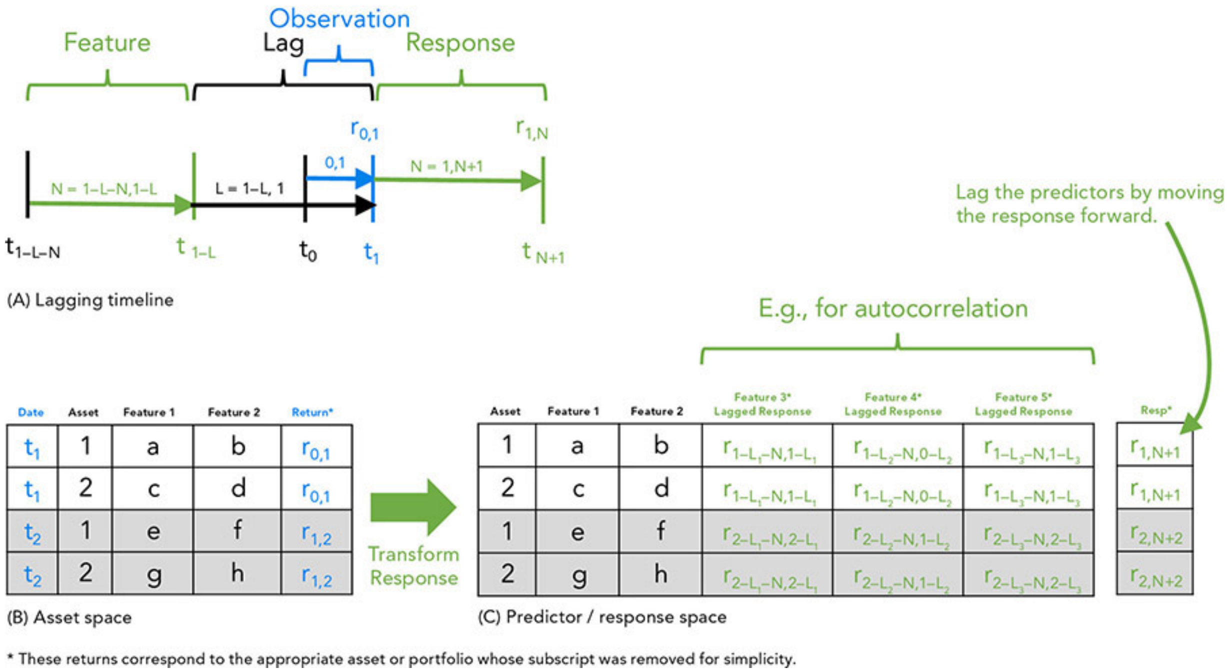


FIGURE 9-1 Transforming the response to remove time stamps and avoid look-ahead bias

Figures 9-1B and 9-1C represent this example as a table, both before transformation (B) and afterward (C). In this case we could build a feature set based on, let's say, three month correlations beginning a month ago, $\rho(r_{\text{Sep-1 to Nov-30}})$, to predict one-month returns starting tomorrow, $r_{\text{Feb-1}}/r_{\text{Jan-2}} - 1$. In this example, we would not use the data from December 1 through January 1 for this row; the lag would be one month prior to the observation date, t_0, and it would continue for one day after the observation date.

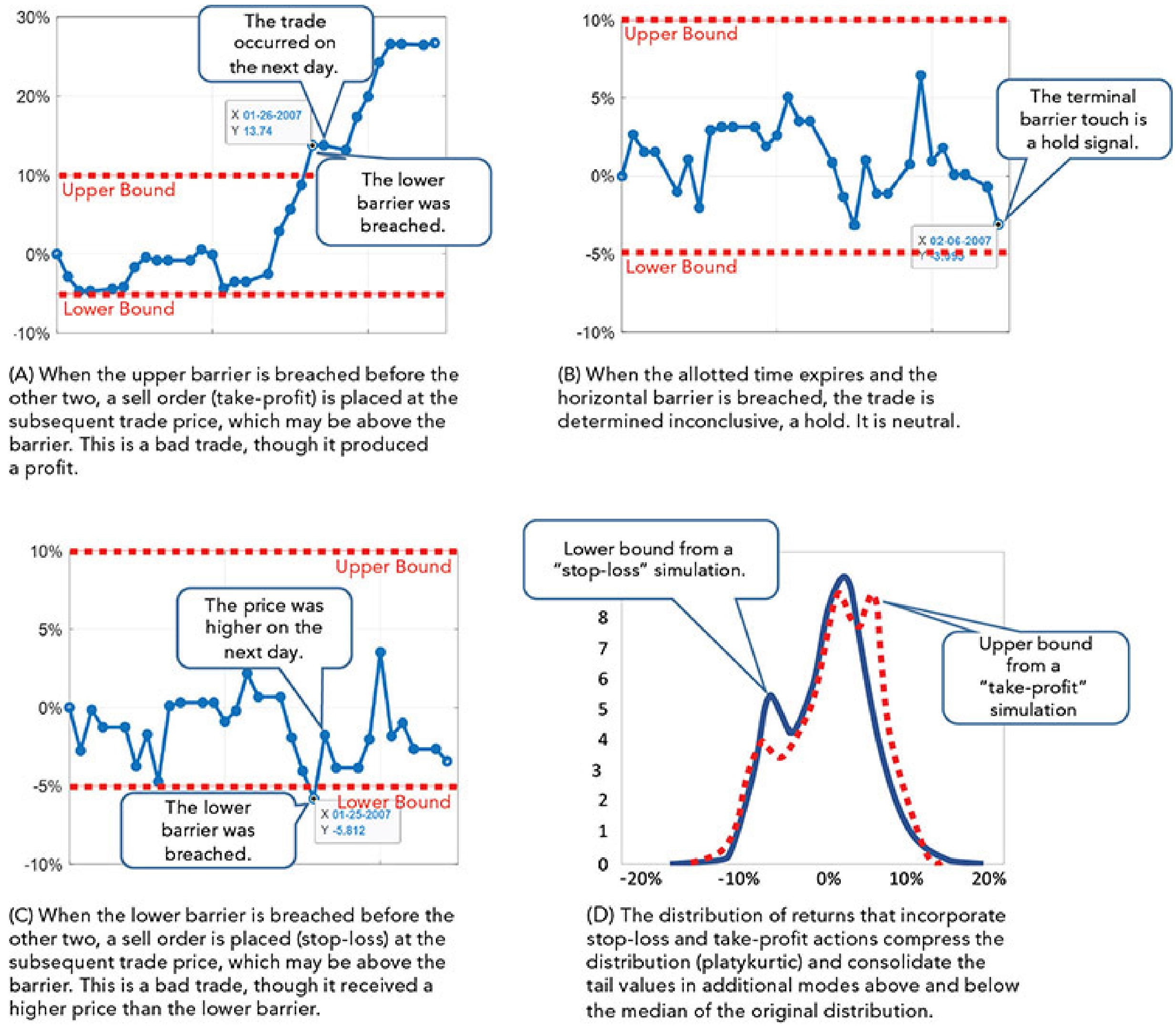
Fixed Horizon Versus Path Dependency

Responses that evaluate a single statistic are unrealistic. Common examples include the return after a fixed horizon (such as a three-day return) or a point statistic like the high or low return in a period. It is unlikely that a fund would wait to achieve a small expected return in an asset if it experiences a much larger loss before the end of the intended holding period. Even if the fund is so systematic that it would wait for the gain to materialize, it could lose its investors or suffer externalities like credit calls or lending failures.

Triple barrier method. If it is possible, the response should incorporate some basic money management features. It is difficult to design a good response that is not path-dependent, but it does not need to be complex. A simple example of this is the *triple-barrier*[5] method. Figure 9-2 shows how a triple-barrier can be used to create a continuous, discrete, or categorical response. Begin by setting the three barriers: a maximum profit, a maximum loss, and a maximum time.

These three categories can be used for classification:[6]

- **Top barrier.** If the return of the investment exceeds the maximum profit before the maximum loss or the time period times out, the response is a "buy" (Figure 9-2A).
- **Bottom barrier.** If the return exceeds the minimum loss first, it is a loss or "sell" (Figure 9-2C).
- **Right barrier.** If the maximum time period passes before either the upper or lower limit is breached, the trade is neutral or "hold" (Figure 9-2B).



(A) When the upper barrier is breached before the other two, a sell order (take-profit) is placed at the subsequent trade price, which may be above the barrier. This is a bad trade, though it produced a profit.

(B) When the allotted time expires and the horizontal barrier is breached, the trade is determined inconclusive, a hold. It is neutral.

(C) When the lower barrier is breached before the other two, a sell order is placed (stop-loss) at the subsequent trade price, which may be above the barrier. This is a bad trade, though it received a higher price than the lower barrier.

(D) The distribution of returns that incorporate stop-loss and take-profit actions compress the distribution (platykurtic) and consolidate the tail values in additional modes above and below the median of the original distribution.

FIGURE 9-2 It is easy to make the response simultaneously path dependent and tractable.

Execution price. If a continuous response is needed for regression, use the return that follows the breech to allow for an execution delay and to avoid look-ahead bias. For a more realistic execution price, we can use the trade weighted average price (TWAP) or a similar statistic for the subsequent period.[7]

It is tempting to use an execution price based on the trade size as a percent of volume (POV). The POV method is cumbersome because the intraday data required to calculate the cumulative volume can be overwhelming. In Chapter 15, we will discuss a simple transformation that can greatly reduce the data size and computation time required to calculate and use POV. Regardless of the method, attempts at realism in a continuous response may confuse the algorithm because the ultimate return value used in the response may be above or below the barrier that triggered the trade.[8]

Return distribution. To determine the barrier levels, an analyst might try several combinations and collect the results in a distribution (Figure 9-2D). The distribution of a triple-barrier response is often multi-modal since the stop-loss and take-profit barriers will dampen the extreme values and accumulate them in the upper and lower modes. The upper and lower modes should be centered around the upper and lower barriers and the variance around those modes is due to the delay between the exit signal and subsequent execution.[9]

Complications. Adding simple transaction costs to the response may seem like a good idea, but it may blunt the signal. The response can be made arbitrarily complex, but the algorithm may become confused. Even the inclusion of transaction costs may confound an algorithm, and most of the realism might have to wait for the backtesting stage.

It is good to check several features of our response independently and in combination. The predictability of our model may be unreliable using some complications but may surprise us in its capability with others, and adjustments that are expected to help a model may be detrimental.

Not just for performance. Not all responses are based on price performance. Since markets are driven by the behavior of crowds, prices add noise to the theoretical value of an investment. It may be easier to use a primary driver of theoretical value as a response, especially if the valuation method is substantially deterministic, as it is in many fixed-income instruments.

Machine learning models have been effective in credit scoring, especially if the algorithm has access to large amounts of proprietary data, such as bank or insurance company data. It may be more accurate to predict credit first and then—based on those credit predictions—predict portfolio performance. Volatility is easier to predict than price movement and is a primary driver of options prices. Predictions can benefit from path dependence. Credit transition probabilities and volatility clustering are integral to credit and volatility models. Causal models understand deterministic relationships and can determine value from predicted drivers. They can also identify when deterministic relationships fail and even how to identify new relationships that can replace the broken ones (*self-healing*).

Allocation, Selection, Direction, Timing, and Quantity

The execution model requires five parameters: allocation, selection, direction, timing, and quantity. It is common to use a strategy model to address only the first two (allocation and selection) and to use a naive default for the others, such as 1/N weighting, risk parity, or next-day execution. Any arbitrary assignment is an uninformed decision with the same hazards (intended and unintended) as a well-thought-out analysis but without the benefit of planning (or the unintended consequences). If the skill required to make an informed decision is unavailable, a neutral choice may be best. Even when a wise decision is possible, the cost or overfitting involved may outweigh the benefit. The strategy may determine all or none of these five parameters directly; it may be combined with other models (layered); or conditioned later (overlayed). Strategies that predict dubious performance (such as factors that do not translate neatly into investment choices) must later be mapped into one or all of these parameters, which involves basis risk.

Allocation. Identifying which factor or asset class to buy and sell is probably the most useful goal for the strategy response for most investors, since there is less efficiency in macro factors than in many securities transactions (and, therefore, more opportunity to benefit from uninformed competitors or competitors with different objectives and time frames). Institutions are often interested in deviations from their policy portfolio, such as overweights and underweights, rather than absolute allocations, though there are usually common-sense absolute limits (such as cardinality) even if they are not explicitly defined.

Selection. Some investment teams focus on choosing which investments to buy and sell, rather than which class of assets. Selection models can be entirely different from allocation models. When addressing both allocation and selection, a layered model (e.g., a top-down or bottom-up model) allows the analyst to choose appropriate factors for each without overburdening the algorithm with conflicting mandates and interactions. More focused investors (such as stock pickers and arbitrageurs) care only about the instrument choice and either restrict themselves to a single asset class or ignore classes altogether.

Direction. Funds can be long-only, long-short, or short sellers. This choice influences whether to buy, sell, or hold.

Initiating or increasing a short is more involved than reducing a long, since it involves finding and borrowing the instrument as well as paying an extra cost. Shorting requires a more involved direction model than purchasing, and the nuances of shorting often reveal themselves at the worst time. Moreover, the desired shorts may be prohibitively expensive or impossible to find when the model is live. Complicating matters, assuming that the absence of a buy signal is a sell signal may inject significant wrong-way risk into the investment process.

Different skills or a different model may be required to identify shorts. Many stock pickers are poor short sellers.

Leverage sometimes magnifies direction. An oversimplifying assumption in some models (such as naive portfolio optimization) assumes that levering an optimal portfolio up and down is superior to fine-tuning its components. This assumption ignores many realities like access to capital, concentration risk, and model sensitivity.

Timing. Investments are often bought and sold during the period following a the signal. More complex models can balance alpha decay and transaction costs (including market impact). Allocation and selection models may be conditioned on regime identification (such as hidden Markov models) or favored using another layer of models (such as meta-labeling) to improve timing. Some models attempt neither allocation nor selection; they ignore selection and only try to forecast timing. These models are appropriate for investments that are predetermined by traditional fundamental analysis, or benchmarked to an index like the S&P 500 or MSCI All Country World Index (ACWI).

Risk control overlays and portfolio insurance strategies combine timing with quantity and direction in an attempt to mitigate the risk of the portfolio or become more aggressive in its investment selection. Other cash management strategies, like variations on the Kelly criterion (wealth maximization), are also popular.

Quantity. Determining how much to buy and sell can be a vexing decision. The confidence level of the allocation or selection choice can be used to determine quantity. This has the advantage of integrating the quantity decision with the allocation and selection model. Layered models, like meta-labeling, can judge the confidence in the model itself. By separating the direction decision from the sizing decision, the algorithm may benefit from an easier task. Moreover, decision alignment (e.g., a strong signal and a large size) or misalignment (e.g., a strong signal and a small size) may be informative.

■ ■ ■

Many models are used in the investment process. Some are used to help clean the data; other are used to forecast factors, combine factors, identify factors, predict factors, predict transaction costs, or a range of other objectives. When modeling the investment strategy to combine factors, the response must be simplified to help the learning algorithm identify opportunities. Adding too much realism to the model can hinder the algorithm's capacity to learn. Choosing what to leave in and what to take out of the response is part art and part chance, and may make a significant difference in realizing profits.

To better evaluate the strategy and add realism requires backtesting with a more sophisticated model in the same way that drug modeling requires clinical trials. Backtests do not ensure that the strategy will trade well, but they are an important step in improving our understanding of how the strategy will behave.

---

1. Although avoiding original and inventive investment strategies was discussed derogatorily in the previous chapter, the risk of outstanding losses is nonetheless real and perilous. It is a *wrong-way risk*, while being wrong and unoriginal can be a *right-way risk*.

2. Classification of discrete categories like “buy,” “hold,” and “sell” tend to work better than regressions that produce a numerical value like a return. Classes can be made more granular, e.g., by adding “rout” and “windfall” to “buy,” “sell,” and “hold,” but even three categories is sometimes too many.

3. Sir Lonsdale Augustus Hale’s 1896 “The Fog of War” is the correct reference, though it is often attributed to Carl von Clausewitz. In his 1832 book *Vom Kriege*, or *On War*, General Clausewitz wrote: “War is the realm of uncertainty; three quarters of the factors on which action in war is based are wrapped in a fog of greater or lesser uncertainty”; however, Clausewitz does not seem to have used the phrase “fog of war.”

4. Refer to Chapter 14 for a discussion of the EMSC execution management object we use to manage these details in our backtests.

5. Marcos Lopez De Prado, *Advances in Financial Machine Learning* (John Wiley & Sons, 2018).

6. The following description of the three classes is not entirely correct. Since the price received after a breach is triggered occurs after the breach, it is possible for a breach of the top barrier to produce a loss, or for a breach of the bottom barrier to produce a profit. For instance, if the top barrier is broken, an order will be generated to sell an investment but the transaction cannot occur instantly and the price received for the transaction may be lower than the upper barrier level, resulting in a loss when a profit was expected.

7. If the strategy was used in actual trading in the past, we might use the realized buy and sell prices.

8. It is important to scale the response to the investment. For instance, a 5 percent return may be a “buy” for an investment with a low variance and a “hold” for a volatile investment.

9. While the left and right modes are often centered around the lower and upper barriers, respectively, the dispersion around those modes represent the difference between the barrier level and the realized execution price.

10

Time Series and Cross-Sectional Analysis for Financial Markets for Part II

What Works

The purpose of this chapter (and later in Chapter 17) is to lay out a sampling of the many different ways we use math to help us invest. A key objective is to entice readers to explore novel applications that can provide an edge in a stiflingly competitive and crowded field rather than relying on threadbare techniques and overanalyzed assets.

Traditionally, a discussion of quantitative methods would begin with a taxonomy: statistical versus machine learning, supervised versus unsupervised versus reinforcement, and so on. Our goal is neither to survey all methods nor to focus on the theory and technology of the methods. Instead, we illuminate the challenges and some past attempts at solving them.

Quantitative analysis and investing have always been intertwined. Qualitative analysis can be powerful—many brilliant analysts work qualitatively. But qualitative analysis can also be used as a rationalization for undisciplined thinking. The difference between quantitative and qualitative analysis is in the:

- **Degree.** The analyst's explicit and clearly defined understanding of the investment concept,
- **Expression** of that understanding as a precise actionable (and preferably extrapolatable) idea, and
- **Proficiency** in describing it to himself and others with precision, such as with a repeatable, learnable, and scalable process or formula.

The stereotype of the technologist is a quant with a poor understanding of economic intuition and an excellent skill at manipulating numbers. The genius-cowboy stereotype is of a confident analyst with an intuitive, unteachable talent for investing that is best gleaned by apprenticeship. The truth in these stereotypes is that many technicians are uninterested in domain knowledge, and some investors have not taken the time to learn the clarity and precision of mathematics.

Qualitative investing may be seen as being more focused on specific events and exploiting critical opportunities, while quantitative investing relies on applying rules to large samples. Qualitative analysts are often limited by their human inability to focus on more than a few opportunities at one time, and quantitative analysts can be limited by their inability to describe opportunities in sufficient detail without overwhelming their technology or financial understanding.

The evolution of computation has allowed machines to work so quickly and relentlessly that they can combine and repeat many simple ideas to exceed some of the most ingenious human introspection. Quantitative financial analysis is burdened by many considerations that invalidate more general methods and require:

- **Novel** and specific solutions to problems.
- **Adaptations** of better-understood solutions, often employing compromises and imperfections.
- **Standard** solutions with caveats[1] or willful ignorance, like ordinary least squares (OLS). These methods are often well understood and perform well considering estimation error, but may be underpowered.

Rather than organize methods generally by their history or technical similarity, we categorize quantitative methods in the way they are used. This chapter will discuss:

- **Data gathering and manipulation,** including finding the data: economic time series, market data, unstructured data, behavioral data, mechanistic and structural adjustments to data, and creating synthetic data
- **Strategy, objectives, and conditions,** which are primarily concerned with the response

The spectrum of models includes:

- **Rules** like moving averages and trailing stop-losses
- **Statistical models** like ordinary least squares (OLS)
- **Econometric models** that may impose a theoretical bias like some vector autoregressive (VAR) models
- **Classic machine learning** (ML) models including *supervised* models, like least absolute shrinkage and selection operator (LASSO), that generally predict values, and *unsupervised* models, like spectral clustering, that can help uncover and explain more unanticipated relationships
- **Complex ML models** like *deep learning*, including *long short-term memory* (LSTM), and *reinforcement learning* which can handle more complex or unanticipated relationships,
- **No free lunch.** The *no free lunch theorem* (NFL),[2] which expresses a lack of dominance of any one quantitative method

The Problem with Models

Generally, models suffer from many compromises and imperfections. For instance, one common problem is overfitting, which can result from regressions, gradient boosting, artificial neural networks (ANNs), and linear discriminant analysis (LDA). Support vector machines (SVM), trees, and k-nearest neighbors (KNN) are more reliable out-of-sample. An advantage of ML over many traditional statistics is the systematized effort to detect and control overfitting. Some examples:

- **Regressions** can be made more parsimonious.
- **Trees** can limit depth and number of leaves as well as impose a minimum number of leaves per branch. *Random forests* are designed to fight overfitting. *Boosted trees* can limit the number of trees, learning rate, number of leaves, penalization, and other constraints.
- **Neural networks** can include penalization and dropouts, decrease learning rate, and impose other constraints.

Overfitting is a critical concern for nearly all models, and ML models are more prone than most due to their ability to find patterns in noise. Many other issues also lead to compromise and imperfections. Good data scientists are scrupulous about pitfalls and balance performance with the limitations of data and algorithms. Some of the more common compromises involve the following.

Bias (stability) versus variance (fit). Bias, including the bias created by regularization, and variance are inversely related. Because economists are often more interested in the economic mechanism and exploring counterfactuals than the fit of the model, they often build *structural* models like dynamic stochastic general equilibrium models (DSGEs). Well-known models that conceal sophisticated reasoning and generally good performance in a simple formula are influential despite their limitations, such as the *Taylor Rule* for the federal funds rate.

Cross-validation (CV) is an integral element of ML but is at odds with the statistician's concept of causality because it draws repeatedly from a single sample to establish out-of-sample performance. CV assumes independent and identically distributed (IID) random variables. Information leakage due to the serial correlation of time series can also invalidate CV.

Big data and many features. Many methods, both high-variance and high-bias techniques, can manage large date sets including ANN, trees, gradient boosting, classification and regression trees (CART), and LDA. Some regressions, SVM, and KNN have difficulty with too many features.

Interpretability, explainability, and fairness. Linear models, decision trees, many types of regressions, CART, KNN, LDA, and generalized additive models (GAMs) are relatively straightforward techniques. Other models, like ANN and SVM, can benefit from ex ante analysis using methods like partial dependence plots, Shapley additive explanations (SHAP; uses game theory to evaluate marginal contributions of models near an operating point), local interpretable model-agnostic explanation (LIME; uses linear approximations for complicated models), individual conditional expectation, occlusion sensitivity, or grad-CAM. For most quantitative investment processes, applying common-sense economic intuition and concern for estimation error is a *right-way risk* for high-stakes decisions (that are not always recognizable as high-stakes when made) that builds goodwill and resilience with stakeholders like managers and investors.

Uncertainty is an integral element of many stochastic methods, like ANNs, which use stochastic gradient descent (SGD) to optimize weights. These methods can be sensitive to initialization weights. The randomness and uncertainty regarding these weights can be disconcerting, so methods like prediction intervals (PIs), stochastic differential equation neural network (SDE-net), DeepEnsemble, and MC Dropout may be used.

Inefficiency can make an otherwise good algorithm too burdensome to use. The slow convergence of some models like sequence learning networks is an example. Some techniques can be parallelized, but others cannot. Technology can reinvigorate old techniques and make them faster. The perseverance of Aho, Weinberger & Kernighan (AWK), Global Regular Expression Print (GREP), and Stream Editor (SED) is a testament to the persistence of well-designed solutions devoid of the bloating of many modern implementations.

Static models, retraining, continuous learning, and transfer learning. The frequency of retraining and the ability to systematize and backtest models without bias is a significant source of performance uncertainty. Traditional static models are retrained on a schedule, triggered, or retrained ad hoc, but continuous learning models retrain automatically based on feedback regarding changes in the environment and the effectiveness of the response.

Interaction effects manifest where additional, unspecified variables may influence the prediction. An interaction plot's telltale cross pattern is one method of identifying interactions, but correlations between variables are not. Partial dependence plots and accumulated local effects plots are common tools. We discussed effective methods to deal with interactions in Chapter 6. Search algorithms and other less well-known additive models that can be applied to data interactions include multivariate adaptive regression splines (MARS) for regression and flexible discriminant analysis (FDA) for classifications. Local interpretable model-agnostic explanations (LIME) and tree techniques are used. A common technique is to perturbate predictors and observe the marginal change in forecast. Adversarial techniques try to identify predictors that produce false forecasts. Cluster analysis can be used to separate prototypes, which represent the data set, from criticisms, which represent minority data.

Loser's game. Quantitative investing is both a "winner's game" and a "loser's game."[3] In addition to the difficult task of finding an investment strategy that beats the market regularly after fees (skill and luck creates winners), there are a plethora of tiny errors that are nearly impossible to completely sidestep (competence and diligence prevents losers), any of which can make a good strategy fail.

Training burden can make a model intractable or too slow to train frequently and unsuitable for *online* use. Pretrained, transfer learning, and simplified models can be used for time-dependent applications (like high frequency trading, where piecewise linear approximations can be used instead of complex curves). The less complicated regression models, KNN, LDA, and CART, may be quick to train. Complex models generally take more time. SVM, though relatively simple, is slow to train.

Fragility and sensitivity. High variance models can be fragile due to overfitting, while high bias models may be fragile when they are applied to the wrong regime. *Perturbation and importance sampling* can be helpful when determining the extent and importance of factors and the model's sensitivity to small changes in general or around an operating point. Interpretability, dependence, and importance methods can be used as well as simple *sensitivity* studies, like using key-rate duration to test the sensitivity of a fixed-income investment to various zero rates.

Scarcity of observations. It would be useful to be able to identify conditions that cause windfall profits or catastrophic failures, but the sad truth is that these events are rare and are thus difficult to train on. When seeking to identify a minority category, algorithms are usually confounded by the plethora of noise and fail to find the signal. Without proper training, algorithms with minority categories typically label all events as the majority class, ignoring the extreme labels. The opposite problem (outliers skewing the population statistics) is common when fitting a trend.

Insidious familiarity. While domain knowledge is possibly the most important factor in designing an effective algorithm, it can also cripple a researcher's ability to build a model that will survive out-of-sample. As researchers become intimate with a historical period—or unduly influenced by one, such as the Great Financial Crisis of 2007–2008—they can be unconsciously drawn into testing for resiliency to those circumstances. Vigilance, culture, and well-thought-out processes can help guard against myopic look-back. It is common for "Chinese walls"[4] to divide research and strategy evaluation groups in an attempt to minimize this bias.

Hyperparameters can make a good algorithm ineffective. That is why it is important to tune hyperparameters before comparing techniques.[5] A random search for optimal hyperparameters is generally more appropriate than a grid search, especially when many hyperparameters create a highly dimensional search space. Resampling random searches are effective for models like ANNs and gradient boosting machines (GBMs). Bayesian optimization is usually preferred but is affected by the starting point and may get stuck on a local minimum. Gradient descent methods add a third level of abstraction and have their own parameters, like batch size, learning rate, decay rate, and the root mean squared (RMS) scaling factor. Other searches are also used, like genetic optimization, simulated annealing, and Nelder-Mead simplex.

Predictive techniques may neither employ nor imply causality. Predictive techniques may only measure tendency and require statisticians and data scientists to determine causality. Ex post narratives create the illusion of causality in an attempt to legitimize a potentially spurious or fragile prediction. Canonically, statisticians and economists prefer high-bias models for causal mechanisms, such as global macro predictions. Purely statistical solutions are common but lack the same perception of robustness and respectability. Causality theory (structural causal models or SCMs) offers an attractive alternative to statistical inference but are currently limited to solving problems that can be represented by directed acyclic graphs (DAGs).

■ ■ ■

As we can see, modeling can introduce many complications that the power of machine learning cannot simply correct. It is important to be aware of these potential shortcomings when designing our model and its response.

Causal Model Concerns

Causal models may be useful in addressing a pervasive problem in machine learning. Algorithms are generally designed to fit a function to data for prediction, rather than to answer questions and form an *unbiased estimate* of *ground truth*. Econometricians tend to prioritize the effect rather than the goodness of fit. Although they may arrive at a model that adequately describes the mechanism, the model may still not fit well. This can be a crucial shortcoming in *structural* models, but it is not as important in models that are frequently retrained. Models that adapt general knowledge, such as *transfer learning* techniques, can adjust predictions to more subtle logic.

While ML techniques like bootstrapping can help uncover spurious relationships, ML techniques generally split the same data for training and testing and cannot overcome some problems like estimation error, which would affect both partitions. While most ML models risk overfitting by using more elaborate methods and parameters, causal models may describe the mechanism with more detail and be more robust.

One method to address issues like estimation error is to alter the *objective function* to compensate for the homogeneity of the data and to use high bias models and statistical characteristics of the estimators and coefficients—the marginal effects rather than the gross prediction. Less comprehensive solutions can settle for half-measures such as *dimension reduction* and *importance* techniques to replace nonparametric models with *semiparametric* ones. For instance, *maximum likelihood* can be used to filter out noise. Many other methods have promise, such as analyzing *residuals-of-residuals*, *difference-in-difference* (DID), and *matrix completion*.

Similarly, ML techniques usually focus on risk and eliminating covariates, rather than identifying the nature of the inference in the way that statisticians would with techniques like *randomized experiments* and *controlling for covariates.* In many cases, causal modeling differs from ML in the thoroughness of its process. The ability to contemplate all possible assumptions helps ensure that they are substantive and unconfounded.

While statistical models can require unreasonable assumptions, assumptions can be relaxed by a less orthodox ML model. The difficulty is that this often requires experiments, and it may be impossible to *design* experiments to affect change but only observe *natural experiments* or *instrument variables*. It is useful, for instance, to perturbate a covariate in isolation or to perform *supplementary analysis*, like using placebos in drug studies.

Once causal relationships are determined, *interventions*, like mind experiments, can be performed using only the model and do not require collecting more data. Interventions allow easy and riskless virtual experimentation to help overcome the human bias of overweighting outcomes rather than relying on cause and effect.

Invariant causal prediction (ICP) leverages the property that causal relationships should be stable when conditions change. *Do-calculus*,[6] and to a lesser extent *reinforcement learning*, *genetic programming*, and *multi-armed bandits*, can be used to explore assumptions in the pursuit of *counterfactuals*. These types of algorithms must balance the *explore-exploit trade-off* (by finding the best choice, not the most common outcome), and the *credit assignment problem* (identifying if a prior decision deserves a reward in an enlightened way), and not just the most recent action.

These analyses can benefit from winnowing misleading data and relationships from subsequent iterations. Sophisticated ML solutions may try to "predict" counterfactuals (if they are related to error) and *instrumentals* (if they are not) but are limited by dimensionality. LASSO is an effective but limited tool, though deep learning has shown promise in overcoming LASSO's limitations in predicting instrumental variables.

Determining structural equations with SCMs, also called *functional causal models*, can analyze root causes and detect faults in real time (*self-healing* models). *Causal additive models* (CAMs) can be used to generate flow charts (self-drawing graphs) and for asset allocation, supply chain modeling, and capital stack analysis.

Data Gathering and Manipulation

We have spent a great deal of time discussing the selection, conditioning, and use of data as an input to our process. Now we will dive into how data affects our model choices and uses, and how the models behave.

Let's start with skill and specialization. Specific tasks like data cleaning, feature identification, and signal identification may be performed by the portfolio management teams or lifted out and assigned to data specialist who can combine the tasks to create *data factories*.

While domain knowledge may suffer by institutionalizing data and feature processing, it is worth the trade-off because of the improved scalability, benefit of skill, and *separation of functions*. Separating data teams, feature teams, and strategy teams can prevent common mistakes like commingling feature engineering (signals like high return on equity, or ROE) with investment strategy (systematic biases like behavioral heuristics). Critically, separating data research from strategy research can help isolate the data phase (that takes great care to avoid overfitting with techniques like cross-validation) from the strategy phase backtest (which is often abused, much like p-hacking).

The *source* of the data can offer an edge to astute analysts who recognize difficult data and know how to deal with it. Data that is so frustrating to work with that it seems to be actively working against the analyst is sometimes referred as "pathological" data. Experienced analysts who are able to assess gaps and outliers in the data may reveal helpful patterns of information, such as using *self-reported* data as a binary indicator of returns due to *adverse selection*.

Frustrations. Sources of data can frustrate less experienced analysts while giving astute analysts a valuable advantage.

Hand-typed, scanned, or handwritten data is common and often full of errors, especially self-reported data and data with *selection bias*. Mortgage applications, news stories, social media, product and service reviews, and search terms are examples of human-entered data that may have spelling mistakes, inconsistent labeling, poor grammar, transposed digits, and other spoilers. Often, labels used to train algorithms are generated "by hand" even if the features are not, which is why supervised learning can be more expensive and difficult than unsupervised learning. *Semi-supervised learning* can be used to train models to add missing labels.

Institutional data, like regulatory filings, economic data, and many legal documents, is often standardized but sparse. Mistakes are less common, but estimation error may create revisions that add a second time dimension.

Frequency mismatches may occur when combining low frequency data with high frequency data, such as when economic data (low frequency) is combined with market data (high frequency). This mismatch can be addressed with method like a Gibbs sampler, which can impute incomplete data by applying Bayesian techniques to contemporaneous time series. This method can produce *nowcasts* using data that is gathered between reporting dates to update the prediction of the next economic release. When the Gibbs sampler is too difficult to apply, techniques like Metropolis-Hastings may sometimes be a useful alternative.

Static factor models, like principal component analysis (PCA) with a bridge regression, can be used for contemporaneous features, dynamic factor models for lags, exact models with a diagonal correlation matrix for correlated features, or approximate factor models for less correlated features.

Bayesian vector autoregressions and mixed data sampling regressions (MIDAS) are popular for nowcasting. *State-space* models like those using a Kalman filter are also common.

Machine-generated data includes most market and transaction data,[7] and a great deal of unstructured or oddly structured alternative data, like data from the *internet of things* (IoT), satellite imagery, mobility data, and metadata. The volume of the data favors ML strategies, but the speed can inundate algorithms—even those designed to be used *online*. We may be able to avoid overloading algorithms by simplifying and approximating functional forms, like piecewise linear representations of market impact curves, and through transfer learning techniques that build on previous learning. Old techniques, like periodic model training (such as overnight batches) and updating models intraday using approximate relationships, like bond and options Greeks, are also popular. While machine-generated data may have lower latency than other kinds, the speed of data generation may create an undeserved comfort. For instance, time stamps are often inaccurately recorded as the time data arrives at the recorder rather than at the time of the event being recorded.[8] Other details, like time zones, may be obvious in more curated data but overlooked in high-velocity and unstructured data.

Refining data can also distort the data because the process almost always destroys information. *Raw data*, such as scans of regulatory filings in image form, contain hints and interactions that are filtered out when databases are cleaned and processed, and muddled as multiple sources are aggregated. As data is further manipulated into *signals*, like sentiment, and other results, more information is lost.

Best practices dictate keeping a *data history* and a *chain of custody* to allow the recovery of lost information. This should include data collection techniques, *data dictionaries*, and other supporting documents. Databases often have epochs with different volumes and richness of data. For instance, old SEC filings, patent filings, and loan documentation may only be available in a format scanned from printed paper, while newer fillings may be highly structured, such as those that are recorded in *eXtensible Business Reporting Language* (XBRL).

Machine-generated data may be too structured and esoteric for unrefined use, like much data generated by the IoT. The *build versus buy* decision applies here. *Sweat equity* can produce a competitive advantage if the resources are available. Effectively applying canned technology to data that is available to anyone with a checkbook requires more skill and luck. As ML techniques become more assimilated into mainstream investing, novelty will no longer attract capital and results will be necessary to justify effort.

Availability. Another area of concern that can spoil the outcome pertains to data *availability*, *sparseness, and imbalances* (Chapters 4 and 6). Analysis problems can crop up when dealing with *imbalanced data*, *economic data*, and *synthetic data*.

For instance, imbalanced data is a significant problem in financial machine learning when outliers, such as large profits and losses, are of interest for predicting profits and managing risk. For algorithms that are not specifically designed for imbalanced data, the data or features must be adjusted to avoid learning bias. Care must also be exercised when evaluating models since measures like accuracy will obscure majority data biases. Other measures may work better, like receiver operating characteristics (ROC), F-measure, and G-mean.

Solutions to imbalances can be sorted into several broad categories: *over- and undersampling*, *feature selection*, and *dimension reduction* and *importance*. Oversampling, undersampling, and combinations of the two are the easiest solutions to employ because they can be independent of the learning algorithm. These solutions include clustering and distance methods (such as synthetic minority oversampling technique [SMOTE]), as well as genetic algorithms, SVMs, kernel bootstrap, self-organizing maps, and Bayesian methods like the *Gibbs sampler*. Combined methods can use different techniques together. Overindexed and underindexed data presents a similar problem, but instead of the data being more or less numerous, it produces an oversized or diminished signal.

Feature selection methods, including *filters*, *wrappers*, and *embeddings*, reduce the set of possible features and may involve misclassification penalties to emphasize some types of errors over others, such as false positives versus false negatives. *Filter* feature selection, like sampling methods, is the easiest to use. Filter methods measure feature characteristics like variance, probability density, correlation, or other distance measures and can be used before machine learning.

Dimension reduction and importance methods include PCA, independent component analysis (ICA), singular value decomposition (SVD), chi-square tests, Laplacian scores, minimum-redundancy maximum-relevancy (MRMR), and nonnegative matrix factorization (NMF).

Classification methods can be helpful, including consensus clustering or ensemble clustering. A common distinction is made between *wrapper* and *embedded* methods. Wrapper methods add or subtract features as they train learning models. Wrapper methods are better integrated with the learning method but can also be more difficult to use and typically involve a search that is *sequential* (like a heuristic or gradient descent method) or *random* (like random selection or genetic methods). Embedded methods are part of the learning algorithm and are the most closely integrated of the three methods. Examples include Gaussian process regression (GPR), stepwise generalized linear regressions (GLM) and regularization, such as in LASSO, elastic net models, and ensembles and boosting, generally.

Economic data is typically characterized by low frequency with short periods. Well-understood methods like Bayesian regressions are better suited to sparse data. Methods that incorporate regularization, like LASSO regressions, are often used to address these problems to reduce the number of factors relative to the number of data points. Vector autoregression (VARs) can forecast several time series simultaneously, but are limited to just a few.

Dynamic factor models can be updated with partial data to form nowcasts. This can be an important consideration. For instance, economic releases trickle out of governments frequently. Methods like LASSO may ignore some of the data, while a ridge regression would be able to incorporate data as it is released.

Synthetic data can allow more scientific simulation because the data can be “fit for purpose” (explicitly designed for a scenario) and also balanced (containing enough data in all classes). While synthetic data may not always be realistic, that data can be designed for any scenario. A modeler can say, “If you don’t like my data, tell me what conditions you think we should model and I’ll produce them for you.”

Another option is to model multiple scenarios and then ignore the less important ones. Selection bias can become a problem when the training set is selected or when the synthetic data scenarios are defined. Monte Carlo methods can be applied to deterministic data (like coupons) and stochastic data (like price evolutions) simply by mapping an inverse cumulative probability distribution. Well-understood distributions are often chosen to reduce the computational expense. For instance, *copulas* are commonly used to model joint distributions. Quantile-quantile (QQ) plots, likelihood ratio tests, the Kolmogorov-Smirnov distance, and the Anderson-Darling statistic can be used to test the suitability of the chosen distribution. Resampling methods for synthetic data creation include cross-validation (CV), jacknife, bootstrap, and subsample.

Cholesky decomposition can generate simulations while preserving correlations between holdings, but generative adversarial networks (GANs), competitive learning methods like self-organizing maps, restricted Boltzmann machines (RBM), and convolutional Wasserstein (WGAN), can better capture the features of real data while limiting overfitting. Another option is the use of parametric Monte Carlo methods like regime-switching.

High dimensionality can be a problem both for models and for data. While there are techniques for managing and reducing large data sets and high dimensionality, they often lead to problems. For instance, they can diminish the interpretability, explainability, and fairness, including mixture models, random coefficient models, and quantile regression models. Reducing dimensions prior to machine learning, such as mapping to a lower dimension with PCA, is convenient and easy. Undercomplete autoencoders are also effective and behave like a nonlinear PCA. Penalizing predictors with methods like LASSO can incorporate dimension reduction in the ML algorithm. Trees are also good for high-dimensional learning.

Distribution assumptions and conditionality frequently lead to problems by assuming normality, stationarity, no autocorrelation, or other unrealistic simplifications. Traditionally, to address these problems, high bias models were adapted when necessary. More recently, better models have become available that do not require these assumptions. Robust learning may also help with some problems. Multicollinearity usually requires dimension reduction, but adjustments for multicollinearity are rarely effective.

Selection bias for samples can be a significant problem. Self-reported hedge fund and private equity returns are a good example of selection bias that leads to survivorship bias.[9] Similar problems can crop up with sparsely labeled data or incomplete dimensions. For instance, many mortgage applications that were robo-signed before the Great Financial Crisis of 2007–2008 (GFC) were often incomplete. An analysis that excluded incomplete forms would be biased toward lower default rates.

Privacy can be a critical factor when the data involves information in areas like patients in medical studies or clients in a private wealth analysis. Sources for this type of information can include *material nonpublic information* (MNPI) or *personally identifiable information* (PII). Privacy and confidentiality can be addressed in many ways, with solutions that are often used together, including:

- **Obfuscation,** which involves hiding sensitive data in plain sight through transformations.
- **Differential privacy,** which is a form of obfuscation that adds random values to data in a way that does not bias or degrade the signal, such as with a Laplace mechanism. Diversification and k-anonymity are differential privacy techniques.

- **Homomorphic encryption,** which allows operations on data without requiring decryption.
- **Federated learning,** which uses a decentralized and distributed data for collaborative training. As such, the access and reliability of a partner’s technology can be limitations. Federated learning can also involve *secure multiparty computation.*

Manipulability is a concern for many models. Clients can manipulate credit models to get better loan terms. Buyers and sellers can manipulate the order book to trick automated market makers and high frequency traders (*spoofing*). Bad actors can also create false trends and spikes to trick flesh-and-blood investors. If a model is designed so that it anticipates manipulation, it can not only detect and react to it, it can also encourage good behavior through social engineering. Models with well-known parameters, like FICO scores, fall in this category.

Confidentiality and disclosure are important to safeguard intellectual property (IP). Ring-fencing and restricting data destroy the “peace dividend”[10] and prevent researchers from freely seeking peer review and assistance.

Strategy, Objective, and Conditions

In this section, we will discuss strategy, objective, and conditions, specifically by:

- **Removing time dependence** from the independent variables so we can employ machine learning techniques that were not designed for time series
- **Examining fixed horizons versus path dependency,** particularly when it is necessary to expand our forecasts with dynamic and multi-period models
- **Assessing allocation, selection, direction, timing, and quantity** decisions

When considering machine learning *design patterns*, it is frequently useful to reframe a problem to apply a different type of technique. One way to do that is to classify numerical data, such as predicting buys and sells instead of returns. Classification allows a regression problem to be reframed into a classification problem that better identifies the desired output (a buy or sell order) and reduces the burdens of prediction.

Classification permits different algorithms to be tried and can simplify the problem, making up for the loss of information in replacing measurements with classes. This also has the benefit of addressing estimation error and prediction error by limiting the range to a few classes rather than an arbitrarily precise number.

Because of noise and limited predictive power, false precision and false confidence should be avoided. Some investors use 1/N weighting schemes to avoid predicting outcomes that are too specific, like investment weightings.

Classification of numerical data often requires arbitrary thresholds that inject bias. For instance, what constitutes a buy? Should data specify daily returns or monthly rolling returns? Should returns greater than 1 percent be considered, or should the threshold for returns be above 5 percent? Frequently, different thresholds are required for different populations (e.g., different stocks) and can sometimes be dealt with by summary statistics, like mean absolute deviation (MAD).

Multimodal distributions (or multimodal distributions of R^2 values) may indicate multiple populations with their own mean-reverting tendencies. This is the basis of the Market Profile technique. It is also why many traders do not put great value in significance measures. If they apply a statistical technique to a bimodal population, it may have a low information coefficient (IC) because the IC was evaluated on the whole data set (but it may have a high IC for the most relevant portion).

The reverse is also possible. Classes can be characterized as modes in a distribution with a cross-entropy loss function. The variance can be interpreted as uncertainty. Here are the pros and cons of some of the techniques used to address these issues:

> **Neutral classes** like “hold” can provide a buffer between binary classes. Despite the benefits of this concept, the addition of a neutral class may complicate the learning, adding more complexity without

providing tangible benefit. When reframing a neutral classification to a regression problem, distributions like the Tweedie distribution can be useful. More often, it is helpful to reframe a neutral class problem (loss, noise, and profit) to a binary problem (loss versus profit) to avoid an imbalanced data set that contains many trivial market movements and relatively few consequential price changes. Obviously, the downside is that small returns are not predictive and a two-class model may generate excessive churn and turnover.

Multi-label patterns can permit classifiers to predict magnitude as well as direction, such as "large loss," "small loss," "small gain," and "large gain." They can also be used for data that may contain similar labels (such as those that overlap, like "profit" and "windfall," or those that are redundant, like "win" and "rally"). When classifications are reframed as regressions, models can use simple functions like sigmoid, which allows values between zero and one, instead of softmax, which is binary. A flat pattern will map labels to a single response, while a cascade will use a hierarchy.

Ensembles are powerful and ubiquitous in ML. They can be used to improve prediction and also to reduce bias and variance (but often not estimation noise).

Cascades split a ML problem into stages that may branch like a tree. This can make it difficult to model interactions between stages and may make the code complex and modular. Often, a cascade is much simpler to design than a comprehensive model. An example could be a single model that predicts transactions by a market maker whose output is used to train two subsequent models: one that predicts transactions from market makers to retail investors and another that predicts transactions to "informed" investors.

Precision versus sensitivity. A powerful use of cascading is to train a directional model that feeds a sizing model to take advantage of the trade-off between precision and sensitivity. Predicting buys and sells should emphasize finding as many opportunities as possible (sensitivity, which is also referred to as *recall*). However, once the trades are selected, it becomes more important to identify the trades that are more likely to be profitable (precision or *positive predictive value*) and magnify them with larger bets.

Imbalanced data can be addressed by cascading a classifier that identifies a population (minority versus majority) and then cascades the population to a separate algorithm. For example, a classifier may identify high volatility populations and pass high volatility, low volatility, and anomalous outputs to either a high volatility regression learner or a low volatility regression learner to predict returns—and then discard anomalous events.

■ ■ ■

This chapter illuminated some of the technical solutions to the investing decisions that we discussed in Part II. Because we are focused on the investment process rather than the technical application, the chapter was organized in the same general order as Part II and can be applied to the topics discussed before it, like feature engineering. Chapter 17 will serve the same purpose for Part III.

---

1. For instance, OLS requires linearity, normality, homoscedasticity, and no multicollinearity, autocorrelation, or endogeneity.

2. D. H. Wolpert and W. G. Macready, "No Free Lunch Theorems for Optimization," *IEEE Transactions on Evolutionary Computation* 1, no. 67 (April 1997), 67–82.

3. In 1946, Charles Ellis wrote *The Loser's Game*, which describes winner's games (where the winner determines the outcome by achievement) and loser's games (where the loser determines the outcome by making a mistake).

4. A Chinese wall, or information barrier, is a policy preventing communication between employees to avoid sharing confidential information, such as material non-public information, or, as in this case, sharing strategy design assumptions with the testing team.

5. To avoid comparing a tuned algorithm with an untuned one, or an optimally parameterized one with a suboptimally parametrized one, a researcher can tune all the algorithms. If not all of an algorithm's hyperparameters are tuned intentionally, there is a chance that the default settings will be better for one algorithm than another. In this case we would not be comparing algorithms, but default settings.

6. J. Pearl, "Causal diagrams for empirical research," *Biometrika*, 82:669–710, December 1995.

7. It is worth noting that transaction data reflects the execution of ideas but not necessarily the intention. Trading errors, like "fat fingers" and homonymous tickers, are common. So, the subsequent market reaction may be more telling than the trade order itself.

8. Intentional delays are common also. For instance, when reporting to the Swap Data Repository (SDR) or when making regulatory filings, some people delay submitting some or all of their data.

This intentional delay is intended to limit the usefulness of their investment position reporting so their competitors will have difficulty deciphering the trade and its intent.

9. A related, but contrary, example involves the infamous bet between Warren Buffett and Ted Seides. The two agreed to wager a million dollars on the performance of hedge funds versus passive investments. Investors with access to the best hedge funds can outperform passive indexes, even after subtracting the high fees hedge funds charge, but betting on the average performance of hedge funds is entirely different.

10. The peace dividend was a term used by President George H. W. Bush and UK Prime Minister Margaret Thatcher in reference to the collapse of the Soviet Union in 1988–1991. The implication is that reducing defensive efforts frees up resources for more productive endeavors.

PART III

BUILDING OUR PROCESS

Prediction and Simulation

In Part I, “Planning Our Work,” we planned our business, product, and organization and began building our quantitative investing process in by identifying and forecasting the factors that drive our investment process. In Part II, “Data, Features, and Response,” we began laying out our quantitative process. In Part III, “Building Our Process,” we will detail how to make those plans a reality. We will learn how to:

- Understand how various risk and return models are designed and used
- Survey technical forecasting methods for financial time series
- Incorporate realism, like costs and impact, that often determine a viable strategy from a naive one
- Build portfolios using various schemes and optimizations, as well as methods for the selection of individual investments
- Backtest our model results to assess their viability with more sophistication than we could in the forecasting stage
- Buy and sell investments in an efficient way, considering complexities like transition periods and taxes

11

Alpha and Risk Models

Greater Than the Sum of Their Parts

The first few chapters of Part II focused on data and inputs. As we progress through Part III, we will shift the focus to model design. The investment process is not neatly segmented into atomic steps, so the chapters overlap in task and purpose. For instance, a model used to forecast interest rates may also be used to generate factors as an input for a corporate bond model, or as the ultimate return forecast for a Treasury bond purchase.

Classes of models. We will build three basic classes of models. They interact with each other in complex ways.

- **Data models** organize the data for use in scoring models.
- **Scoring models** process factors to determine ranks and statistics that can be used by functional models.
- **Functional models** perform the investment tasks like allocating assets to classes and selecting investments.

Scoring models. In this chapter we will discuss the first three of the four scoring models: alpha models, risk models, and integrated models.

Functional models. Later in Part III, we will talk about the functional models that apply scoring models to affect an outcome. Functional models address asset allocation, security selection, timing, weighting, execution, and implementation. Oftentimes, integrating models may be prohibitively complex. To make combining models tractable, models are often evaluated in stages instead of integrating them. Multistage calculations and implementations can disregard the interaction between stages.

Alpha Models

The term *alpha model* is really a misnomer. Alpha models are more properly described as performance models or return models since these characteristics are not necessarily alpha[1] (such as the relative attractiveness of investments that are primarily concerned with non-systematic risk) or beta (such as index or *smart beta* investments). Alpha models are most important to researchers and portfolio managers.

Many investment managers think of the methodology of their alpha model as the “secret sauce.” An investment manager frequently identifies with his methodology, regardless of whether his work is based in theory or experiment.

Most quants are educated in science, technology, engineering, or math (STEM), and frequently physics. Many scientific discoveries occur within the same time frame, as scientists are immersed in the same culture, system of scientific discovery, and body of knowledge. Claerbout’s principle states that the entire environment, especially the process, is what makes the alpha model work, not one brilliant atomic concept.[2]

Quants preach systematization, but it is the quants themselves that are the product. The entirety of the manager’s process is what is responsible for the gap between smart beta and hedge funds (and other alpha vehicles). The reason why so many smart beta funds perform poorly out of sample is because they are disengaged from the process that made them work in sample. For this reason, *autonomy* and *resilience* are important characteristics for alpha models, so they might be more independent from the process and tinkering of the manager.

Even the design of a fully autonomous system requires many discretionary decisions that must be understood to foster confidence and persistence when the system is underperforming, or in order to know when to “pull the plug.” A plethora of minute details must be defined, any of which can significantly alter the operating conditions and results of a strategy.

These details include a universe of available assets, states, investment expressions, hyperparameters and hyperparameter reevaluation frequency, factor evaluation frequency, model retraining, rebalance triggers, confidence limits, and trade horizon. Until someone sits down and starts to code up a strategy, it is difficult to imagine how many assumptions need to be made, and it is easy to unwittingly impose an assumption by omission.

Deviating from a good system because of fear and uncertainty is a common failing of systematic managers. Many quants have more training in math than in discretionary trading, and a seasoned trader is likely to agree that the market induces participants to make bad decisions (e.g., behavioral heuristics). That is the nature of markets and the essence of behavioral finance. Management is often a negative influence, pressuring the quant to do the wrong thing. “Plan your work and work your plan, take responsibility, and be decisive” may be good advice, but it is difficult to follow unless the assumptions of a system are well known and defensible. “Great is the enemy of good,” until the fund underperforms. When clients complain and emotions tempt management to go “off script,” the value of diligence in planning well will become apparent.

For example, the analyst may reach one of the two following conclusions:

> We chose a trailing three-day stop-loss of 12 percent. We determined this after hundreds of historical and simulated studies convinced us that research is more reliable than to use a 15 percent stop-loss or to wait and hope for a recovery. Hope is not a reliable process.

Or:

> This is outside of our operating range. We don’t have enough data to be confident in our signal. We should deleverage to a neutral position until circumstances change. I understand it means crystalizing our losses, but to do otherwise would be a blind gamble. We should only invest when the odds are in our favor.

Many funds advertise the mechanical nature of a fund as evidence of resilience. Independence may sound attractive, but the illusion can hurt a manager beyond their tendency to overfit. The secrecy it engenders often prevents the manager from sharing necessary information during due diligence. Investors can be fooled, but allocators who have interviewed thousands of managers understand that rigid, secret formulas eventually fail, but process, hard work, and determination can overcome.

Following is a list of key considerations in developing our model:

Taxonomy. Generally, models rely on one or more frameworks:

- Historical data, especially statistics about returns
- Forward-looking historical data, such as implied interest rates
- Quantitative approximations of qualitative theories (financial, economic, behavioral)
- Qualitative theories of reality

Despite the appearance of uniqueness, there is a surprisingly limited diversity in alpha models. To put a label on something is a way to disabuse it of the sense of specialness and to expose the true value of the manager—his competency, diligence, assumptions, implementation, discipline, operation, and management. One way to look at alpha strategies is to categorize them in four major groups:

- **Product structure or statistics,** such as mechanistic, dislocation, market/bookmaking, scalping and skew, insurance diversification, or arbitrage
- **Regime detection,** such as event-driven or global-macro
- **Fundamental,** such as value, yield, or quality (management, fraud, governance, accounting quality, reporting language, revenue resilience and variance, etc.)
- **Technical,** such as momentum, trend, insurance buying, growth, contrarian, insurance selling, sentiment, or mean-reversion

Funds often combine these categories or, from time to time, will favor some over others.

Alternatively, models can be categorized by the investment style of the manager. Categorizing by style can help when combining quant managers with less quantitative strategies. There are many style categories. Four popular ones are:

- **Equities**, such as fundamental growth/value, market neutral/directional, sector, short
- **Macro,** such as thematic, commodities/currencies, systematic
- **Event driven**, such as special situations, activist, mergers and acquisitions (M&A), distressed/restructuring/credit, private placements/initial public offering (IPO)
- **Relative value,** such as sovereign/corporate/asset backed securities (ABS), fixed income (FI), convertible arbitrage (convert arb), volatility, infrastructure/real estate

Quantitative versus qualitative. In general, the benefit of quantitative strategies is the diversification of bets (and its corresponding reduction in the reliance on good fortune); of course, diversification results in the dilution of resources available for each analysis. The analyst may reduce concentrated risks by using technology with limited sophistication to enable the analysis of vast numbers of opportunities. Centralizing good judgment in a strict process—or even a diverse web of disparate models—often forces quants to rely more on past data and algorithmic forecasts than thoughtful analyses of specific situations. The systemization of alpha and experience, intuition, and judgment are essential to successful quant models. Nonetheless, quant strategies that emphasize edge and grind can lack the flashy appeal of a good narrative and the worn story that some people will always be smarter than machines. Some may be smarter, yes. When properly used, machines do not try to outsmart humans; they think differently and play to advantages people do not have.

Like some qualitative strategies, some quant strategies are naive. There is no shortage of quant strategies professed by technologists and amateurs. Single-factor strategies and a reliance on unimaginative factors and investment choices are commonplace. These strategies are publicized due to their easy-to-explain design, as well as their ability to be a straw man for the anti-quant and passive-investment narrative.

At the extreme, quant strategies are as complex as thoughtful qualitative strategies, including planning for contingencies that often occur at the worst time for violated assumptions, execution failures, operational errors, unexpected liquidity events, financing problems, and counterparty defaults (unintended, like Lehman Brothers collapse of 2008, and intended, like short squeezes). Quant strategies have the distinct advantage of diversifying the risk of these problems, while suffering the disadvantage of a lack of focus on a particular event unless it has been identified and guarded against beforehand.[3]

One clear example of an overreliance on theoretical outcomes is the naive application of leverage to scale the performance and volatility of a portfolio, such as the optimal portfolio with lending and borrowing. Leverage can reduce the benefits of diversification as well as amplify errors in portfolio design. A similar pitfall occurs when an overfitted risk control overlay (*drawdown control*) responds too slowly out of sample and misses the opportunity to de-risk prior to a rout—but does so at the trough and misses the subsequent rally, compounding the damage.

Flexibility. Autonomy and resilience require flexibility. Flexibility is part of what makes quant models similar to discretionary models. The broader and more inclusive a model is, the more flexible it must be, as even the most influential trends can be unpredictable. The more niche a strategy, the more mechanistic and reliable a rules-based system can be.

Start with a Principal

Because so many little decisions are required, there is a danger of cumulative subliminal (or intentional *data snooping*) bias. The choice of a guiding theory is influenced by history, but it can be circumspect and deliberate. *Anchoring*, *framing*, and *confirmation* biases can be powerful influences. Model choices are also numerous, including:

- Independent and dependent variables
- The precise definition of these, including their expression
- How they are acquired, measured, and cleaned, as well as missing values and outliers
- Dealing with sampling and estimation error
- Weighting, such as favoring more recent values with exponentially weighted moving averages (EWMA)
- Robustness, such as resampling methods
- The measurement period and frequency

- Seasonality and other adjustments

Even weighting based on signal strength can magnify the bias in the signal. To avoid using signal strength for weighting, some quants equally weight (1/N) or weight based on something other than the "alpha" signal, e.g., risk parity.

Parsimony is often recommended, since including fewer variables makes most theories easier to understand and solve. Adding variables may increase the R^2 for technical reasons (e.g., overfitting), but may not improve the model or its information ratio (IR; Figure 11-1), much like increasing return history or frequency in an attempt to improve standard errors.

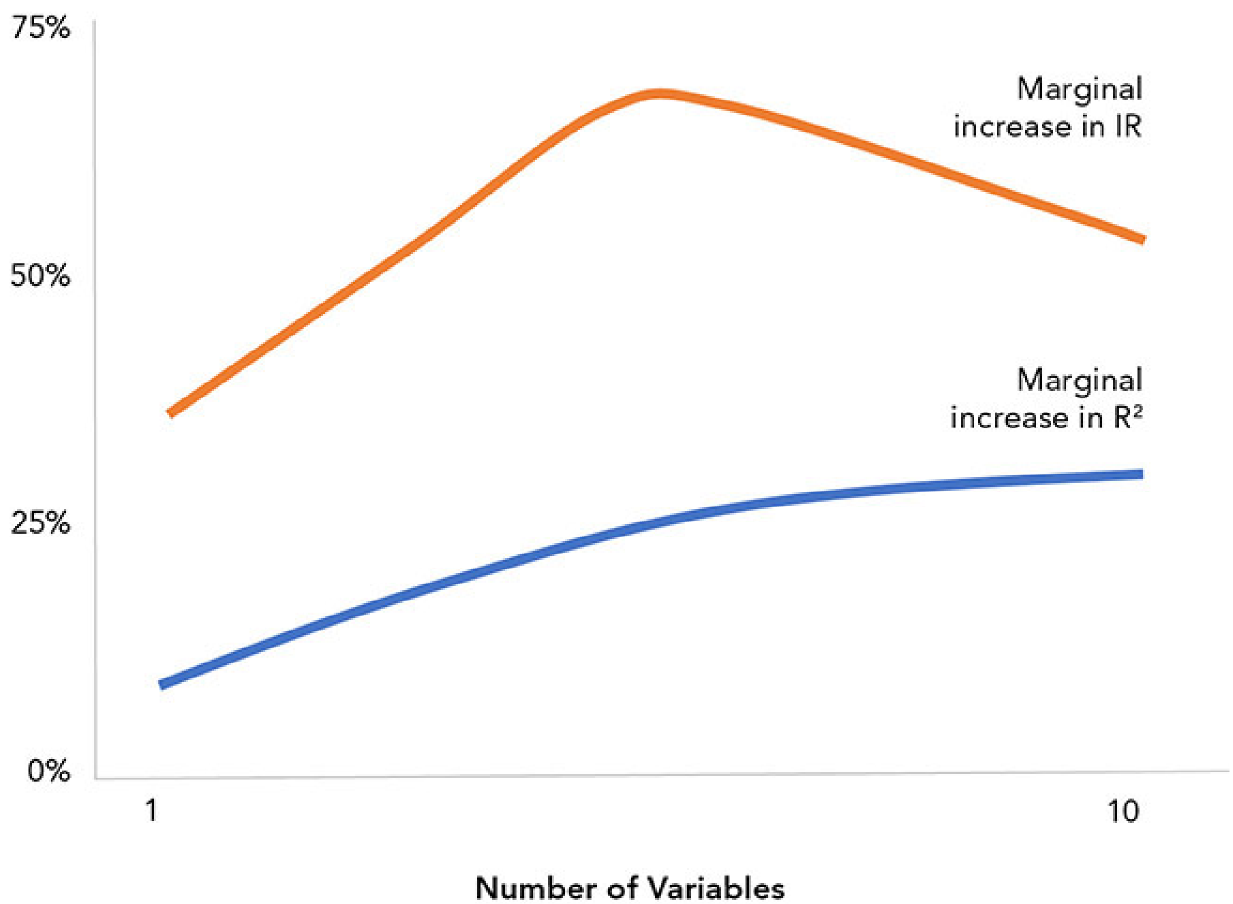


FIGURE 11-1 The number of variables vs. IR and R^2. Adding values increases R^2, but the diminished benefit can be seen in the marginal increase IR.[4]

When risk management *is* the alpha. Alpha models are designed to predict cash flows and appreciation. A common *anchor* for these values is zero. If cash flows or appreciation are negative, they are usually termed as losses, and the models that predict losses are thought of as risk models. However, this is arbitrary; they are still part of the “alpha” model used to determine relative attractiveness. The assessment of negative flows and appreciation are common for credit instruments, mortgages, and investments that hinge on binary outcomes like drug approvals, wildcatting, and private investment exits (such as initial public offerings or IPOs).

Risk Models

Risk models refer to the variation of returns, probability of impairment of investment operations, or the potential for financial ruin. Risk models are used by investors but may also be designed specifically for regulatory and compliance reasons. For instance, a Comprehensive Capital Analysis and Review (CCAR) or Basel III model may not produce output that is best suited for the investment process, but comply with internal and external regulations and budgets. Risk is a confusing word; it is not separable from other aspects of the investment process, and often is not definable in absolute terms. Consider four stages in the process:

1. **Alpha model (selection):** finding investments with good risk-adjusted returns, including investments where returns are determined by minimizing bad outcomes. This model is managed by the investment team, and any responsible selection process naturally incorporates risk measures.
2. **Risk model (allocation and construction):** combining investments in a way that produces a portfolio that is better than the sum of its parts. Combining assets involves using different kinds of risk measures with different focuses. Although there is overlap between alpha and risk models, risk models are more concerned with the quality of returns (reliability, consistency, variance) of the entire portfolio. Balancing wrong-way risk and right-way risk, correlations and covariances, and other interactions can be more important than the attributes of an individual

investment selection. The most obvious application of this model is in the timing and weighting of the investments within the portfolio.

3. **Risk model (leverage, rebalancing, and overlay):** controlling and managing the risk of the portfolio as part of the systematic process. The same models used in Step 2 can be applied to the systematic management of the portfolio. As the basis for combining the selections changes, it may be necessary to adjust the portfolio to change the overall risk or to adjust the weightings of the holdings to reflect the new reality of constantly changing inputs. Depending on the urgency, cost, and policy, these adjustments may be invasive. For example, an equity portfolio may be quickly and cheaply levered by buying or shorting a blunt index to adjust the beta. The opportunity cost of this method is implementation shortfall.
4. **Attribution and regulatory model (leverage, rebalancing, and overlay):** analyzing and reporting the ongoing behavior of the portfolio or strategy. Attribution is the final step in the feedback loop and may be used as a learning tool. The model may recommend the use of a method predefined in Step 3, a more invasive rethinking of the investment selection, or even a rethinking of the overall investment policy. This model may also be used to assess "hard stop" breaches in regulatory or organizational limits. Models built for business continuity or systematic resilience can be too "big-picture" for the investment staff. In many cases investment staff incentives are intent on performance. Myopic or self-centered concerns may make these broader risk matters appear arbitrary and counterproductive and may produce agency problems. We will discuss this step later in this section of the book.

Defining market risk. The organization, product, and model purpose will determine the objective of the risk model. Some organizations exist only to control or minimize risk. Rather than limit risk, investors should use risk as a currency to be spent (*risk budget*). Understanding the sources, character, magnitudes, and variations of their risk can help investors deploy their risk budget efficiently for larger and higher-quality returns, and for the fulfillment of their incentives through investment selection and sizing. Common concerns include:

Impairment of investment activities is perhaps the most unfavorable outcome; even a pathological gambler should want to avoid insolvency. *Permanent loss of capital* is a close second; irrecoverable losses are universally unpleasant. Kelly-sizing (Chapter 12) maximizes return while managing the "risk of ruin," but is modified or dismissed by many investors since this method ignores other risks.

Timing, covariance, and direction. Characterizations of pain (such as right-way and wrong-way risk, upside and downside risk, and drawdown magnitude and length) are naturally intuitive characterizations, and are part of any substantial investment conversation. Internal metrics (such as winners versus losers, and covariance) are also unavoidable because they are easily understood and visceral. Longitudinal (timing) and cross-sectional (covariance) models are both used.

Variance and dispersion. Moments and other measures of periodic risk are important for regulatory and business reasons, and theoretical risk tolerance and asset allocation assumptions often require them. They are baked into many processes, models, policies, and regulations, but are less important and practical than performance for most investors. Fear (measures of pain) and greed (measures of absolute or relative performance) are the most pertinent gauges of investor concern.

Diversifiable, idiosyncratic, and hedgeable. It is wise to limit an investment's risk dimensions to those that are predictable and controllable. By identifying the characteristics of an investment that are mispriced (edge) and minimizing the uncontrollable (hedging), an investor can align his or her results more through skill than luck.

Uncertainty. The nebulous but critical nature of risk measurement often results in a dashboard, or even a single measure to summarize the state of a portfolio or organization. There must be a way to distill complex information so that senior decision-makers can monitor their charge, building uncertainty and learning into the final output (such as by using Bayesian updates). This could involve providing confidence bands or distributions rather than just point estimates or smoothed time series. Uncertainty can be overwhelming; define methods, actions, and policies to rely on when the "fog of war" makes the way forward uncertain.

Without planning, it is easy to succumb to nihilism, paralysis, or gambling during difficult times. Explicitly incorporating measures of uncertainty in decision-making (such as stochastic dominance and game theory) and organizing courses of potential actions (such as *risk triggers*[5]) can help form a framework. The plans we make may not be used when they are needed (such as if the portfolio managers did not foresee exactly the circumstance that were encountered or if politics derail a carefully designed plan) (Step 3 above), but a plan's existence can provide an essential framework for analyzing confusing and emotional situations, as well as for some level of confidence in resulting decisions (Step 4 above). "We do not rise to the level of expectations; we fall to the level of preparation and training." (Archilocus).

As with alpha models, high-bias theoretical models can uncover forces, while more flexible empirical models may produce unanticipated results that elude suspicion. For example, the fallen-angel bias of corporate bond indices is one instance where biased models may confirm structural risks. Judicious use of bias is an advantage of economics and statistics.

Language matters. An important outcome of the risk model is the framing of investment conversations. Like risk premia, risk metrics focus and add meaning to the currency of risk. Take care to use the jargon properly, to add meaning without complicating the message. Words have power and loose diction, e.g., inferring causation rather than tendency, can cause damage—leave creative language to salesmen and spin doctors. When engineering an investment thesis, a lucid understanding of risk can be as important as a thorough understanding return drivers.

Integrated Models and Factor Alignment

Integrated models combine alpha and risk models to make risk-adjusted recommendations.[6] Less frequently, models that integrate execution strategies may be used to plan ahead and anticipate decisions by the execution model for more cost-effective decisions. This can be detrimental if the execution environment differs from what is anticipated. For instance, an integrated model may choose one asset over another because the chosen asset historically had a smaller bid-offered spread than another asset, even if the other asset enjoys a higher expected return. If the bid-offered spread of the other asset narrows at execution time, the portfolio may end up with a suboptimal allocation without the benefit of lower cost. Tax-loss harvesting is an example of a strategy that may require execution-aware security selection.

Integrated models are helpful in estimating values in a compatible way. It is common for the alpha and risk models to be separate and theoretically incompatible, often produced by entirely different research groups or software vendors, because they are often used by two different groups. The alpha model's ultimate goal is reliable performance, while the risk model may be designed for a different purpose and used in unintended ways. Integrated models can provide an exogenous input into the investment process (allocation, selection, timing, and weighting). In building a model, our success can be tripped up by many shortcomings.

Oversimplification. Many financial methods involve decomposing hopelessly complicated relationships into more relatable components, like risk premia. Often, these more isolated components are intractable. Another problem with this methodology is the need to account for the interactions between components and other influences, like constraints and costs. The asymmetry of market dynamics, such as the relative difficulty in shorting versus purchasing single-name stocks or capacity constraints, precludes the use of simple models. To predict and measure the errors in our approach, we apply the backtesting analysis (Chapter 14) and attribution analysis (Chapter 18).

Ambiguity. The imprecise definition of risk, and the disparate needs of various organs of an investment firm, result in a reliance on risk models that may be incompatible with their corresponding alpha models. For example, semi-variance or value-at-risk may be perfectly appropriate for liquidity managers, but inappropriate for an asset allocation model when permanent loss of capital is the true concern.

Optimization. The portfolio optimizer presents the most obvious conflict between alpha and risk models, because its purpose is to balance the performance and risk. Incompatibilities such as different sample periods or misspecified measures, or overlaps between alpha and risk factors, would produce a suboptimal portfolio. Optimizers (sometimes called "error maximizers") tend to amplify errors, making them an attractive target for critics.

Inaccurate optimization (transfer coefficient) can easily be measured by the disconnect between investor recommendations (information coefficient of the risk-adjusted alpha model) and portfolio performance. This may be attributed to portfolio-level risk control (such as leverage, volatility, or sector bias), when the real culprit is a more insidious misalignment of component models. These misalignments can affect every aspect of the investment process, not just optimization.

Circumvention. It is natural for an aggressive organization in a challenging environment to favor those employees who have a more direct and measurable effect on the immediate needs of the firm (it is best to be "close to the money"). High-performing investment managers often pour great effort and resources into an alpha model and neglect the risk model. The disparity is often compounded by an attitude among investment staff that risk management limits and hinders their opportunity set (and compensation).

It may seem prudent to purchase a sophisticated and defensible third-party risk model ("best-in-class") rather than build one that complements the internal alpha model. This can result in using a model built for a disparate purpose, such as using a capital adequacy or business continuity model to provide inputs to an asset allocation function.

Companies that rely on academic research can be lured into using inappropriate risk measures, prioritizing elegant and tractable models (e.g., closed form or stochastic representations). Theoretical insights that can be broadly applied and used to forecast interesting outcomes. Without modification, these insights may not adequately describe current portfolios and situations or the subtleties required to transact in the real world.

Lack of transference. Even the most relevant risk models are often inadequate. Investment processes are usually designed as a series of modules or stages that cannot easily capture the interaction between their functions. For instance, it would be beneficial for an execution model to fully understand the criteria of the selection model. If it did, it could choose to purchase a slightly less attractive asset that is cheaper to execute (such as a more liquid asset that suffers less market impact). Various machinations are used to compensate for the lack of interaction between the investment stages due to the complexity of designing and the difficulty in operating integrated models. Like a junior trader, the execution model can be given guidelines by the selection model (such as a ranking of preferences), but that is not the same as having the nuanced understanding of a portfolio manager who can account for subtle interactions or unanticipated events. Second-order errors can sometimes develop into big problems.

Misalignments are not limited to the process flow. Interactions occur in all dimensions of investing, including the evolution of the portfolio. Multiperiod implementation is a difficult problem to solve, and is often approximated or ignored in practice.

■ ■ ■

“Divide and conquer” is a useful strategy but produces complications. The common approach of separating returns (alpha) from risk minimizes the importance of the interaction between the two, and ignores the reality that return and investment risk can often be two interpretations of the same phenomenon. The nebulous nature of language and the imprecision of thought creates further confusion. An awareness of subtle and significant shortcomings can create a strategic advantage when applied consistently over time.

A proper understanding and respect for investment risk (“risk culture”) and its role in reliable performance—as well as a meritocratic management structure and an appropriately aligned incentive scheme—will be reflected in the governance of the firm and its management of investments. It will be observed in thoughtful and precise descriptions of events, opportunities, and proposals. It will be felt by employees, seen in their ownership of tasks and outcomes, and will flow through to client satisfaction and retention during trying times. Performance fluctuates and is ephemeral, but process and culture are resilient.

---

1. Alpha is frequently used to describe the value of an investment process in a general sense. Alpha is actually the portion of an investment's returns that cannot be explained by a relationship with a benchmark or other reference. Practically, this means that alpha should not include that portion of the return stream that could be replicated by a cheap systematic strategy, like a passive index or a simple active strategy (also referred to as smart beta). An investor should only want to pay high fees for the portion of returns that are not achievable without the manager's special skills and advantages.

2. "The idea is: An article . . . in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete . . . environment." D. Donoho and J. Buckheit, "WaveLab and Reproducible Research," Department of Statistics Technical Report 474, Stanford University, 1995.

3. Two additional and serious wrong-way risks of quant strategies are the possibility of an unimaginably senseless error that no human would conceive of and the scale that automation can apply, making a small error substantial through repetition.

4. K. C. Ma, *How Many Factors Do You Need?* KCM Asset Management, Inc., 2010.

5. Risk triggers are simply a series of conditions and responses, such as "email the portfolio manager if the portfolio's volatility increases 10 percent" or "call an investment committee meeting if losses in any holding are greater than 20 percent in one day or 35 percent in one month." Risk triggers can give stakeholders, such as investors or board members, confidence that concerns will be dealt with and not exacerbated by delay and confusion.

6. Transaction cost models will be discussed in Chapter 15, and attribution models will be discussed in Chapter 18.

12

Asset Allocation

Choosing Investments Holistically

Portfolio construction is a holistic combination of investment strategies, encompassing strategic and tactical asset allocation and security selection, while balancing risks and constraints. It involves selection, weighting, and timing, which results in a portfolio of assets based on views, preferences, conditions, significant events, more mundane history, and constraints. Many research teams focus on the alpha model, but the alpha model merely maximizes the reward objective (which may contain risk penalties), often using time series and cross-sectional techniques. Portfolio construction, which balances risk, return, and interactions can be even more impactful.

Active versus passive management. Passive managers tend to prefer fewer, broader investment vehicles like index funds, and less frequent rebalancing. Although passive managers may evaluate and rebalance their portfolios less frequently, unagressive rebalancing often reflects a lack of confidence rather than passivity.

Attempts to actively manage broad asset classes or factors in competition with many experienced and well-resourced portfolio managers is challenging. Similarly, timing is notoriously difficult and perilous. Niche strategies that exploit inefficiencies and dislocations caused by incompatible investment goals are more likely to be successful. Many large-cap stock valuation discrepancies are due to estimation error.

Fixed income (FI) choices are numerous. They include diverse structures, such as those with embedded exotic options. FI structures and nonlinearities are idiosyncratic and FI markets are often inefficient, allowing dislocations to persist. The markets are also *segmented*, creating apparent arbitrage opportunities that may not “clean up.” FI instruments often trade in large quantities, or *round lots*. Many of these investments can be actively managed, sometimes even mechanically, to produce reliable and repeatable profits. For this reason, about 85 percent of fixed-income funds are actively managed compared to 70 percent of equity funds (estimates vary, of course). Specialization does not always yield better results.[1]

Sometimes different allocations are managed using varying degrees of activity. Techniques like *core-satellite* and *portable alpha* can separate inexpensively managed passive beta investments from specialized and expensive actively managed allocations. In these ways, active management can be limited to circumstances where it is most efficiently applied. Some investors are constrained because of limited assets under management, liquidity needs, expense limits, shorting constraints, or other policy limitations.

Whether managing funds actively or passively, numerous choices must be made. A modest menu of these choices appears in Figure 12-1. It is easy to assume that passive strategies do not require active choices or that systematic strategies are not arbitrary and do not require intervention. For instance, even passive strategies require weighting and rebalancing rules, and these rules can often be overruled by the investment committee.

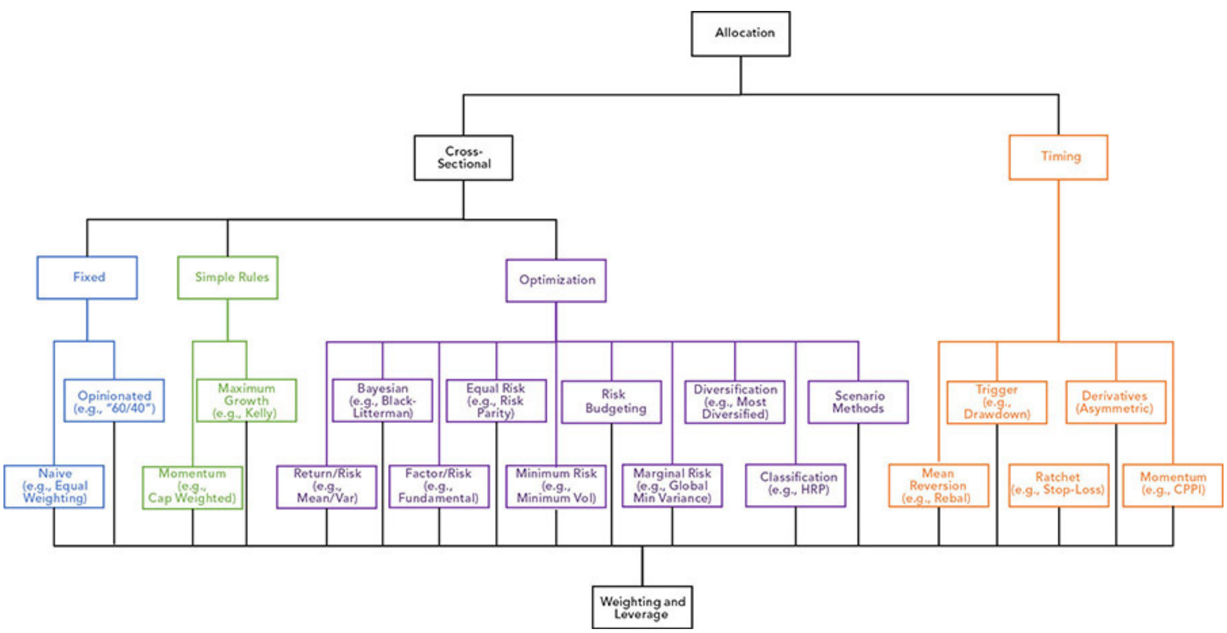


FIGURE 12-1 A taxonomy of some asset allocation stages

Cross-sectional weighting can seek to maximize returns or other factors. Often, weighting is used to diversify investments by combining them to improve risk-adjusted returns. Frequently, some variation of Markowitz optimization is used, but there are many other techniques that can be applied:

- **Fixed proportion** techniques can come from uniformed (or nihilistic) assumptions, a lack of sophistication[2] or lack of confidence, or from an unwillingness to educate investors.
- **Simple rules** (heuristics) ignore interaction effects. These factors may include transformations or combinations like the Sharpe ratio. They may involve a simple scaling based on the size or rank of the factors.
- **Optimizations,** including scenario-based methods, involve a balancing of more than one competing factor, such as return and risk accounting for interactions. Sometimes the solution is easy. Optimization usually involves assumptions and compromises. More realistic solutions may require more time and computing power than is practical.

Timing can also be used to smooth a return stream to reduce risk. These timing methods frequently involve some sort of leverage control applied to individual holdings or, more commonly, the asset allocation, such as the stock/bond/cash ratio. This is sometimes referred to as *drawdown control*.

These techniques include risk on/off, stop-losses and take-profits, insurance programs, and asymmetric investments, like options. Time series methods are often applied as an *overlay*, without integration in the cross-sectional asset allocation method—or after it, as a “bolt on.” One of the primary problems with these techniques is that the triggering methods can be perilous and are often tricked into reacting too late to protect the portfolio and persisting too long to participate in a recovery.

Cross-Sectional Techniques

Diversification. Most cross-sectional techniques are founded in the assumption that diversification mitigates risk. Diversification is widely praised as the “only free lunch in investing,”[3] but, like most concepts in investing, it is often misapplied and misquoted. Whether through overdiversification or simply inept asset allocation, it is important to recognize that noise alone does not diversify properly; it can dilute the portfolio return towards the risk-free rate.

Uncorrelated assets with positive returns diversify in a constructive way. The more certain one is about the value of an investment, the less interested she should be in diversifying it with a less promising investment.[4] Moreover, adding riskier investments to a portfolio will generally increase the risk of the portfolio and must be justified by that investment’s return potential and diversification benefits.[5] While risk is unavoidable, in the same way that costs and fees are ever-present, diversification benefits can be unreliable. The classic diversification of stocks and bonds varies greatly with time.

Noise is biased. Adding complexity and randomness is not neutral, because uninformed decisions are biased toward the wrong investments. Competitors, frictions, and heuristics conspire against investors. As Chris Brightman wrote, “The noise we observe is not white noise.”[6]

Time does not diversify in the sense that it does not increase the value of the portfolio. Adding the number of time periods does not improve averages, though it may sometimes reduce variance (if variance is not correlated with the mean). Multiple investment periods allow randomness to converge with the law of large numbers, which is similar to executing repeated random trials. That is different than combining uncorrelated investments. If the portfolio starts with a “good run,” time will likely draw it closer to the mean. This effectively narrows the confidence interval around the mean compounded return, but it does not dissipate the risk.

This is even more true when human behavior and investor utility are considered, rather than merely terminal wealth. Performance incentives may allow managers to crystallize compensation in good years, inducing them to “roll the dice” more often. Similarly, myopia and behavioral heuristics compel both investors and managers to judge performance based on annual investment periods, rather than on the appropriate investment horizon. By compartmentalizing returns in periods, good years are often attributed to investment acumen while bad years are “diversified” by “staying the course.” Bad trades become long-term investments.

The time diversification misconception may also result from the way risk is measured, rather than how it is perceived. As profits accumulate, the risk of shortfall decreases, but the potential for large losses increases.[7] Since many risk measures are based on some form of standard deviation, which increases with the square root of time, the risk calculation compares favorably with returns that compound geometrically.

Equal weighting rules require the least skill. 1/N-weighted portfolios, which own each investment in equal dollar amounts and expect a return corresponding to its correlation. Even when skill is used to select assets, equal weighting requires only direction and not magnitude. Equal weights diversify predictions and prevent concentrations based on inaccurate confidences. The embarrassing difficulty of investing is evident by the competitiveness of this portfolio. Some investors prefer equal weighting because it emphasizes smaller capitalization investments and is a simple mean reversion strategy.

Fixed proportions permit allocations to be unequal, but they are otherwise much the same as equal weighting. These allocation schemes include the popular concept of stock/bond proportions like 60/40 and 70/30, or 100 minus the client's age. Their use is usually justified by the simplicity of the narrative and the belief that planning (participation in the market through asset purchases) is more impactful and easier to get right than investing (choosing which assets to purchase and when). As we discuss later in the chapter, these proportions are not truly measures of risk. They vary with the time and composition of the components (e.g., which stocks are in the equity portfolio, the duration of the bond portfolio, etc.).

Simple rules can be described by complex equations but ignore interactions. They include:

Signal (or forecast) strength or confidence combined in some relative ranking is a common naive construction method. This may be as simple as weighting by factors such as fundamental valuations, Sharpe ratios, or the inverse of a risk measure. Often, interactions between signals are ignored at this stage, as well as more nuanced effects like trading costs.

Investors who use signal rank must rely on backtesting to evaluate the complexities, and investors who optimize also often outsource some of the complexity, testing, and tuning to a later stage. One obstacle of this method is that when signal strength or confidence is optimized, it is possible for positive signals to produce a negative allocation, creating a conundrum for allocating incentives to different alpha teams.

There are some important *trading rules* in the context of building a response function to train learners based on events. Market technicians may use entry and exit rules like those governed by signals. Traditionally, these rules are not combined[24] with a concern for interactions. When combining rules with signals, or when considering interactions, rules can be converted into signals or embedded into the utility function.

Momentum includes market capitalization weighting, or weighting in proportion to price. Market cap weighting is used by many indices and is tempting for managers, but unless a third-party fund is used, it may be difficult or expensive to fully replicate a fund of this type. Sampling techniques, such as *stratified sampling*, can effectively accomplish partial replication. Other momentum or mean reversion methods often involve individual time series but can be complex.

Maximum growth methods are less concerned with risk and correlation (except for the risk of ruin), and include the *Kelly criterion* (also called the *growth optimal portfolio*) and its variations. Many of these techniques can be implemented as an optimization as well, but are more often implemented for individual investments without concern for interaction other than availability of capital. Cognitive maps and other *uncertainty reduction techniques* are used to sort seemingly unrelated relationships to form an implementation goal. Techniques as simple as solving systems of simultaneous equations can transform vague concepts into actionable investment decisions. Systematic applications of these methods are commonly incorporated in quantitative methods, including those that handle optimizer conditions. Even simple pairwise rankings of preferences can produce powerfully explanatory representations of relationships. Concepts like the *problem space matrix* can expose and organize domain knowledge and fuzzy concepts into powerful rules. They can also expose biases and manage uncertainty. If they can be systematized, they can become part of a process that combines human knowledge and leaps of cognition with structured and relentless quantitative analysis.

Ordering. Some less quantitatively inclined managers avoid the thorny topic of optimization altogether. For instance, they may rank sort portfolio combinations based on factor exposures and test quantiles. They might then invest in the top 10 percent portfolio and short the bottom 10 percent of the portfolio.

Meta-learning. Meta-learning is a way to combine signals by applying machine learning to the alpha forecasts themselves rather than the predictors they use as inputs. An example of this is cross-sectional meta-learning, which can help weight or turn contemporaneous signals on or off. Time series meta-learning can combine rules and forecasts of different frequencies, horizons, start times, and durations. Meta learning presents some challenges: for instance, revisions to forecasts can create a time series problem. While some term-structure signals can be evaluated, not all periods are of equal value for every signal. Differences between beginning of day and end of day, and accounting periods, can be significant. In addition, market impact can sometimes be felt long after the trade, although it may diminish (or explode) during slow nighttime hours. Some of these problems can be avoided by using multiperiod optimizations.

Optimization techniques seek to find the “best” or *optimal* combination of signals involving interactions. They are distinct from forecasting techniques. In this context, “optimal” is a technical reference that indicates that the results should be treated as a guide, rather than a hard-and-fast policy. The optimal portfolio is only optimal if the data given to the optimizer is an accurate representation of the future and if the conditions and assumptions of the analysis are accurate and complete. This is unlikely to be the case.

Realistic optimizations can be prohibitively difficult or otherwise imperfect.[9] Practicality requires numerous simplifying assumptions or approximations spawning many classes of optimization (such as dynamic programming) and many objective functions:

- **Return/risk,** including *mean/variance optimization*, has been the reluctant go-to model for asset allocation for decades.
- **Bayesian**, including the *Black-Litterman* model, uses views and updates to align with forward-looking outlooks and changing conditions.
- **Factor/risk**, includes *fundamentally weighted* allocations.
- **Equal risk budget** techniques, including *risk parity*, assume returns are equal and include equal volatility and maximum diversification techniques.
- **Minimum risk**, includes *minimum volatility* methods.

- **Risk budgeting** invests in proportion to risk contribution rather than capitalization.
- **Marginal risk**, including *global minimum variance*, ignores returns.
- **Diversification**, including *most diversified*, assumes all Sharpe ratios are equal.
- **Classification**, includes *hierarchical risk parity* that allocate based on distance measures.[10]
- **Scenario-based** methods, such as stress tests, can use historical data or synthetic data to examine hypothetical performance during past or forecasted situations.
- **Stochastic model predictive control** or *optimal control* are state-based methods.

Most of the optimization methods listed above are relatively simple, well-known among investors, and can be executed mechanically. Scenario-based solutions are more salient and less mechanical. This may be a drawback for quant research but a boon for interpretability and explainability. Scenario-based methods can incorporate simple instantaneous shocks (e.g., GDP increases 3 percent) or complex (e.g., multiperiod path dependent shocks, such as GDP increases 2.5 percent to 3.5 percent over the next year). Scenarios can include expected outcomes and stress tests or be based on past events that are easy to explain (such as "what if we had another Great Financial Crisis [GFC]?").

Scenarios may be used to manage uncertainty (using ranges and multiple conditions) and instability (by resampling or bootstrapping, for example). Data can be estimated from views or outlooks, simulated, or drawn from historical samples. Historical samples can be modified in many ways including shifting or compressing time (such as using minute bars as if they were daily values). Historical data can be drawn from events or extracted to simulate a particular condition, such as a period in time when inflation increased quickly or the yield curve inverted.

State-based methods are common in engineering but less well known among investors. *Stochastic model predictive control* can solve problems in an *open-loop* or *closed-loop* (using feedback or *recourse*) fashion. *Certainty-equivalent* variations use deterministic models. Kalman filters are a simple and popular predictive control method. Optimizing signal revisions is a useful application of these techniques. They can often be solved using *dynamic programming*.

Strategic allocation, tactical allocation, and security selection. In Chapter 2, we discussed the purpose and process of strategic and tactical allocation, and of security selection. However, problems may arise if the distinctions are blurred or if managers emphasize or even exclude some stages.

Strategic portfolios are allocated based on policy set by upper management or boards. They define broad asset classifications in ranges or bands and are intended to be held for long periods unless significant events transpire that violate their assumptions. By reducing the set of choices and extending the investment horizon, management can focus on secular or mechanistic characteristics, broad themes and trends, and long-term structural considerations, such as liabilities. This reduced problem specification makes it relatively easy to apply sophisticated methods like optimization.

Institutions like pensions, endowments, foundations, and sovereign wealth funds tend to place a strong emphasis on this stage of the investment process. Because of the size and lengthy time horizon of these funds, granular active management can be challenging and expensive. For all but the most well-staffed institutions, time spent on significant allocation decisions can have a far greater impact on the long-term results than dissipating limited resources on a more exhaustive management approach.

Tactical investing involves timing and selection. Tactical portfolios are what most people envision when they discuss portfolio management.

Tactical allocations are more granular and respond to events more quickly (monthly, daily, or even intraday, in the case of compelling economic releases). Asset categories, such as sectors like healthcare instead of classes like domestic equities, can still be broad enough to allow most quantitative methods to apply timing. Even investors with long horizons frequently tinker with their portfolios, perhaps because of a compulsion to act or a need to justify their fee.

Security selection is used to choose precisely which investment to employ to best express the investment thesis. Selection often depends on excruciatingly detailed analyses of capital structures and implementation details.

Many constrained funds, market neutral investors, and arbitrageurs produce impressive returns by selecting investments without regard to asset allocation at all. Those who try to apply sophisticated optimizations may find their analysis intractable due to the size of their investment set and the nonlinearity of the investments themselves. For this reason, managers often prefer suboptimal portfolios optimized at the asset class or sector level rather than the individual investment level.

Optimization

There is a distinction between forecasting techniques and optimization. Optimization and related methods can fine-tune forecasting methods (hyperparameterization) and combine forecasts. Conversely, forecasting methods can be used for optimizing the allocation of assets. The distinction is that forecasting methods strive to increase the value of a signal, while optimizers combine signals with the goal of preserving as much value as possible. Optimizers can add value by combining components so that the sum is greater than the parts, but the signals are not improved by the process.

In optimization, accuracy and the mitigation of parameter estimation error are prohibitively expensive, so assumptions and simplifications are necessary, leaving the most comprehensive and detailed analysis to the backtesting and attribution phases. People are ingenious and methods are rapidly evolving; we describe the most relevant challenges and solutions specific to quantitative asset management.

Implementation is one of the most important details that separates a successful professional investment process from a haphazard and amateurish endeavor. As Part III progresses, we will discuss modeling implementation, the details of these models, and how the models are executed.

Portfolio optimization problems have become more challenging since they were originally popularized in the 1950s. Correlations are weaker and more conditional, making volatility and drawdowns more “contagious,” markets more segmented, and risks more complex and accessible through secularization. This section will briefly discuss some of the challenges, partial solutions, and enhancements that have been devised to apply technology, specifically optimization, to portfolio construction.

Compromises. Complex mathematical solutions to portfolio construction are fraught with compromises and deficiencies of the worst kind for technical methods—those that are catastrophic, immediately observable,[11] and intuitively obvious, like instability and extreme sensitivity to inputs. This challenge is compounded by the complexity of the task and the realization that much of the benefit in using technology is that the solutions are unintuitive but better than intuition. Investments that dominate in isolation may be detrimental in combination with others, and vice versa. These effects are often due to a complex interaction between costs, risks, constraints, and other details.

Since many of the reasons for security selection and weighting are fleeting and difficult to explain in all of their nuances, some managers focus on alpha and must make the often-erroneous assumptions of independence and intransigence of outperformance. For these mangers (especially those with unsophisticated clients or high-variance, high-risk investment strategies), the benefits of an easily explainable process often outweighs the precision and extensibility that technology can offer. Technological solutions, for all their faults, address shortcomings directly and with steady progress.

At the same time, there are no better solutions available for complex portfolios aside from “common sense,” which is easily confused by complexity and heuristics. Increasingly, unintelligent machines are making quick work of besting thoughtful investors penny-by-penny, leaving people with only risky and confounding opportunities for which neither success nor attribution are clear. Moreover, machines are evolving; there are still many opportunities to take advantage of quick but naive machines, but these opportunities are increasingly fleeting and dangerous.

Portfolio factory[12] (Figure 12-2). Optimizers facilitate a clean and formulaic investment process to tailor portfolios to the manager's beliefs and outlook. Here are key steps to the process:

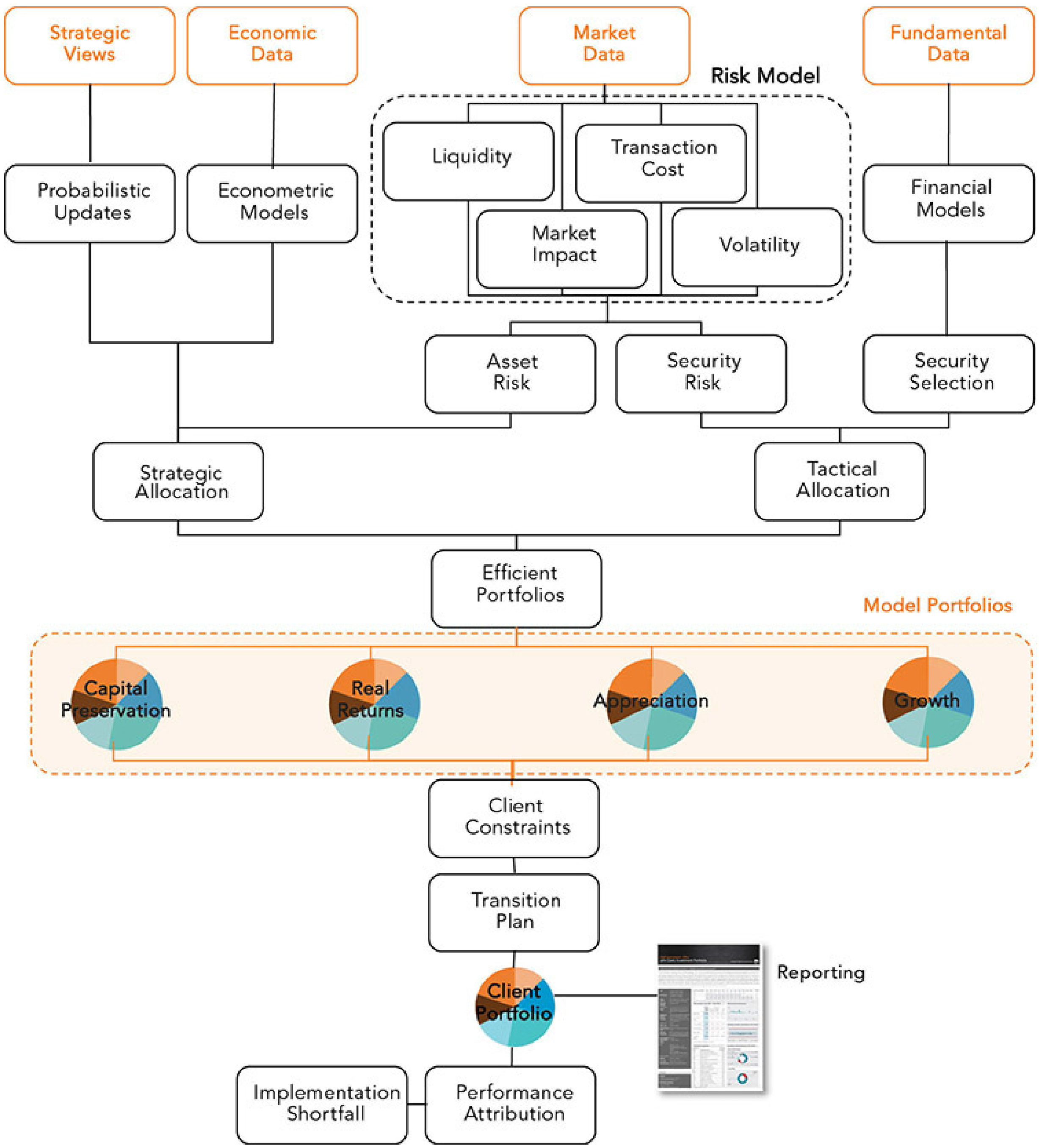


FIGURE 12-2 A simple portfolio factory

Establish a baseline. Every rational investment portfolio—intentionally or not—expresses an opinion, belief, or "view"[29] of how the market will perform. Allocations expend resources, including a risk budget, to deviate from the views that are implied by an inexpensive, diversified, unconstrained portfolio. To determine the manager's baseline views, managers can iteratively back-out (reverse engineer) implied capital markets assumptions using a model like Black-Litterman.

Create models. Applying investment policy and discretionary views to the unconstrained portfolio results in a master optimal model portfolio. An efficient frontier can be used to create model portfolios for various risk tolerances. The manager can change the curvature of the frontier by choosing investments that are more or less correlated. The curvature determines the trade-off between risk and return, making the choice of risk appetite fairly rational. For a frontier that is flat, appetite is arbitrary, but for steep frontiers, risk can often be well rewarded (and caution can be expensive).

Tailor the portfolio. Optimizers can transform model portfolios into appropriate products by applying constraints. The constraints may be based on different benchmarks and can be thematic to provide a menu of products such as ESG, inflation protection, and growth—all based on consistent institutional views. For managed wealth, client goals and constraints can be applied to produce portfolio weights that are appropriate for each client. Sophisticated optimizers can manage taxation, including locating trades among various qualified and non-qualified accounts (*householding*). Noble Prize–winning economist Robert C. Merton proved that when efficient frontiers are combined, they result in an efficient amalgamation. This allows entire portfolio families to be combined to form new families.

Advantages and trade-offs. The optimization methodology has many advantages for an institution or growing advisory. This process is clear to follow, easy to explain, and theoretically appealing. It is fair, equitable, and defensible. No client can assert that they were treated differently or with less professionalism because of discretionary decisions. Model portfolios defend against allegations that a particular client was disadvantaged by his bespoke allocation. Model portfolios are the culmination of the most skillful opinions (house view) using the full resources of the company (taking advantage of pooled resources and economies of scale). Model portfolios also allow client-facing employees to focus on building their book of clients. A consistent institutionalized process builds a brand that elevates employee decision-making and minimizes key man risk. The three most significant trade-offs for technical solutions include overfitting versus imprecision, bias versus variance, and accuracy versus computation time.[14]

Assumptions can be used to reduce the computational burden of quantitative allocation techniques. A benefit of analytic techniques is the ability to generalize rules. Most practitioners recognize the necessity of using nonparametric techniques that require iteration. Many academics pursue closed-form solutions that can be solved analytically but require unrealistic assumptions. In the pursuit of elegance, complexity like taxes or fees are often simplified or ignored; common assumptions include normality, rationality, and risk-aversion.

Oversimplification, especially with single-horizon methods, is almost always necessary. There are many reasons why multiperiod methods are important:

- Implementation and rebalancing nearly always begin with an existing portfolio that requires transition, often with great concern for taxation.
- Transaction costs can have repercussions over multiple time horizons.
- Hedging is employed over different time frames for different reasons.
- Risks, objectives, goals, views, signals, forecasts, and constraints often span periods and have different frequencies, starting points, horizons, and durations.

- Multiperiod optimizations are difficult to apply to complex conditions and constraints and require vast amounts of memory and computational capacity. Their requirements increase geometrically as portfolio composition and conditions become more complex.
- Conversely, it can be difficult to form precise and accurate views in a single time period. Estimating term structures of views may create an illusion of precision by "filling in the blanks" with uninformed data.

Constraints, both hard and soft, are often used in an attempt to impose real-world conditions on technical solutions, or to tame unruly techniques to make them more reliable, countering their tendency to exploit and amplify minor errors. Because constraints work in direct opposition to the design of these techniques, their application often makes the techniques intractable. Constraints can be used to create "guardrails" to help errant optimizers converge on a stable global solution, corralling the algorithm and forcing it to provide a solution that is determined by the constraints and not the science.[15]

Objectives need to be simplified to permit statistical or machine learning. The complexity of reality (like market impact) combined with the influence of human psychology (including political pressure to stop-out or ride profits and losses) make modeling difficult. Frequently, researchers and managers overcompensate for misspecified objectives with backtesting and overfitting.

Risk is a nebulous concept based in psychology and expressed mathematically. There is no one way to express these fears in numerical terms. Attempts to ignore psychology, like the Kelly system, are difficult to apply under real-life circumstances where human beings hold the ultimate responsibility. We will discuss the measurement and management of risk in Part IV.

Technology requirements trend toward increased memory and faster processing. Some risk limits and transaction cost models, cardinal constraints, and other nonlinear constraints can defeat simpler conic optimizers. In these cases, QP relaxation and second-order cone programming (SOCP) may be used to solve these complexities without resorting to full scale optimization.

It is tempting to apply scattershot technological solutions to solve difficult analytical problems. However, this can result in incomplete, poorly understood, overfitted, or over-constrained solutions. A growing menu of techniques is available in the quest for a sensible outcome that converges, rather than a robust and proper outcome. Some ingenious solutions include Bayesian optimization, fuzzy, particle swarm, neural networks, agent-based simulations, and other forms of supervised and unsupervised learning.

Concentration and herding are problems in quantitative investing, despite its highly competitive and secretive nature. However, similar to scientific discoveries, most quantitative investing methods are developed in a shared environment with unavoidable influences and dependencies. As ingenious as they appear, quantitative methods share exposures. Shared exposures are also the source of quantitative strategy rallies and crashes ("flash crashes"). Independent return streams are the ideal outcome, but the limitations of orthogonal returns reward the analyst with the ability to categorize and simplify asset allocation.

Unintuitive, unstable, and sensitive solutions are common in portfolio optimization. Inexplicable outputs ("corner solutions" and frivolous trades) clash with decision makers' intentions. This may be due to either interaction effects or estimation error, as well as the propagation and compounding of errors. Financial data frequently lacks stationarity. Financial analysts rarely consider s*ignificant digits*, and their analysis can suffer from the illusion of precision and the tendency to replace allocations due to slight changes in preferences.[16] Understanding the sources and relative importance of these challenges can help focus limited resources and troubling assumptions. Estimation errors are not all equal in importance. Optimization is usually more sensitive to expected return than to risk, and more sensitive to risk than covariance. Sensitivity varies with the utility and objective functions.

Incorporating uncertainty in the estimation process addresses the illusion of precision. Many techniques have been designed to address uncertainty, including those that use Bayesian probability, fuzzy math, and bootstrapping. Dynamic programming, stochastic programming, and robust optimization (worst-case utility) have also been used to address uncertainty. Distribution-based risk models, including value-at-risk (VaR), and coherent metrics, like conditional VaR (CVaR), can incorporate non-normal distributions and tail risk. These problems are usually easy to solve.

Combining multiple views, outlooks, signals, and forecasts is essential for sophisticated investment management. The difficulty of reconciling views that may be contradictory are compounded by mathematical hindrances like nonlinearity, non-additivity, non-commutability, multicollinearity, and non-stationarity. Views may vary by:

- **Time**—horizon, frequency, strategic vs. tactical, long-term versus short term, turnover
- **Measurement**—including amortization, tax realization and matching (last in first out or LIFO, first in first out or FIFO, etc.), and attribution (including incentive allocations)
- **Periodicity**—ticks, bars by time, currency, volume
- **Granularity** asset class versus sector versus instrument, holistic versus scaled versus tiled
- **Quantity and relationship**—grouped versus absolute versus relative to other views, linear versus nonlinear, equality versus inequality
- **Absolute or relative** to a benchmark (difference versus tracking error)
- **Conditional**
- **Return based or liability driven**

There are solutions of varying complexity and efficacy, including averages, meta-learning, and Bayesian methods, such as the Black-Litterman model and Attilio Meucci’s entropy pooling.

Verification of results is a benefit of scientific and quantitative investing. Examinations of the validity of a methodology is built into the investment process. These include out-of-sample results and scenario and sensitivity testing.

Implementation is difficult to incorporate in an optimizer without compromise. Examples of trade-offs include:

- **Conflicting objectives and constraints,** such as those for different family members, business units, or accounts, including different cost and risk aversion
- **Conflicting goals and liabilities**
- **Delayed and inter-temporal execution** including market impact mitigation, temporary substitution of investments, wash sales, and capital raising trades
- **Subscriptions and redemptions**
- **Separately managed accounts (SMAs)**
- **Tax lots and locating investments** in qualified and non-qualified accounts, including locating among different family or company accounts like hold-for-investment (HFI) versus for-trade accounts

Constraints and Other Considerations

Constraints are often invaluable for an investor to address quantitative asset allocation. *Investment constraints* are used to make the solutions applicable to clients and regulations. *Technical constraints* are used to make the optimization tractable and to make its solutions sensible.

Investment constraints can also be used to enforce exogenous requirements that produce appropriate solutions that would otherwise be suboptimal, such as *client constraints* including social restrictions; *regulatory constraints* such as short selling restrictions; and *discretionary constraints* such as portfolio management (PM) decisions. Investment constraints attempt to identify human preferences, so they come in many varieties. *Holdings constraints* are a common place to start restricting the portfolio. *Risk and return constraints* include risk-adjusted return, consistency, variance, drawdown, periodic, catastrophic, point and composite risk, path-dependent risk, and tail risk. Risk appetites and tolerances are often estimated and oversimplified in an attempt to codify investor behavior.

Investors are complex, and inconsistent to the point of contradiction. Even sophisticated institutional investors will complain about their peers outperforming and about absolute losses when their budgets are under pressure. Intelligent and understanding investors will be “compassionate” but still dismiss their managers for being unable to maintain multiple contradictory mandates.[17]

Budget and group constraints can set boundaries on individual assets and on systems of proportional relationships for groups. There are many common groupings, like Global Industry Classification Standard (GICS), geographies, and styles. Nontraditional groupings include principal components and importance. Other unsupervised and hierarchical methods are also common. Assets and factors can be included with positive filters or excluded with negative filters (prohibited investments). Negative filters are a simple way to limit unwanted investments and are easier than favoring sought-after assets, like ESG. However, they can have unintended consequences like favoring unwanted assets that are not explicitly excluded.

Relative constraints like risk-adjusted returns (Sharpe ratio, Sortino ratio, etc.), *active weight*, and *tracking error* are often necessary to express goals. Few investors have the luxury of being judged on absolute return alone, especially when traditional and inexpensive benchmarks can outperform well-paid fund managers by a wide margin over extended periods (after the managers extract their fees).

Abstraction of observable data—including *latent* variables, like economic forces, groupings and sectors; and abstractions, like factors—allow more precise expressions of intentions, although they create implementation complications. By focusing on optimizing signals rather than assets, we can drastically reduce the opportunity set of available choices to those that are most worthwhile, increasing the effectiveness of the optimizers.

Long-only constraints are among the most restrictive, performance-destroying constraints, because they prevent the optimizer from underweighting assets below zero. One workaround involves addressing underweights when providing signal specification during the feature engineering stage. *Shorting* has a number of drawbacks. It is more expensive to enter and exit positions; it involves negative carry and opportunity costs; short interest must be paid, and dividends and buybacks relinquished; and short squeezes are commonplace. In addition, short positions can be called or substituted by the counterparty, and their potential for loss is unlimited. The competition among short-sellers is more sophisticated and often well-capitalized with access to low borrowing rates. The financial system contains mechanisms to cure distressed companies, including bailouts, mergers, acquisitions, and negotiated settlements, creating a hazard to short sellers. Cognitive biases and social forces are stronger with shorts, and price action is often punctuated by large, sudden jumps. Rallies tend to last longer than routs, limiting opportunities for shorts. While long positions grow as a percentage of the portfolio, and the longs' investments' prices appreciate to an investor's benefit, short allocations grow as they fail to perform.[34] This creates natural pressure to "double down" and requires a dynamic hedge that incurs expenses on the way down and the on the way up (*negative convexity*).

Leverage costs, including those for *constrained long-short* portfolios (e.g., 120/20), do not net. Similar to shorting, leverage is messy and asymmetric. Great care must be taken when employing leverage. Academic and theoretical assumptions often fail at the worst times and are often multiplied by leverage, sometimes geometrically or beyond a tipping point (*metastable*).

Risk-adjusted returns are not always a goal in isolation. They may be part of a plan to manage liabilities or life goals, like major purchases. Strategies that incorporate liabilities and goals into the plan can include segmented portfolio assets that are earmarked for these objectives. A common example is investing in assets with fixed maturities, or cash flows that balance specific obligations or time frames. Similarly, many institutions, like pensions, optimize *surplus* (the difference between their assets and liabilities).

Liabilities and goals are difficult to hedge because they are driven by forces that are not often reproducible using available investment choices. For instance, while inflation protection is available in some forms, the specific form of inflation that drives a portfolio's liabilities may not be hedgeable, like wage inflation or beneficiary class migration (such as active to retired or disabled).

Additions, distributions, and withdrawals are related to liabilities and goals because they add or subtract value from our portfolio, getting us closer to or further away from our target. Ideally, a multistage optimization will include systematic cash flows and determine an optimal allocation and spending schedule.

For pensions and endowments, actuarial analyses may define the outflow term structure. For private wealth, life events and regulatory requirements for spending down qualified accounts may be the driver. For tax-loss harvesting strategies, inflows are usually required to refresh the tax lots (delaying *burnout*). Taxes are often baked into the expenditure schedule.

Taxes and other frictions can be complex, with path-dependent constraints such as holding periods and other restrictions. For instance, capital gains and dividends in the United States are taxed at different rates depending on whether the asset is held for a long or short term. Conversely, capital loss carry-forwards are taxed at the same rate regardless of the holding period, creating an arbitrage referred to as *tax-loss harvesting* and an aversion to *wash sales*. Taxes and tax-loss harvesting are discussed in detail in Chapter 16.

Further confusion comes when swapping out tax-efficient investments like total return swaps, municipal bonds, and exchange-traded funds (ETF) for common stock, corporate bonds, and mutual funds that do not offer the same tax treatment. Often, the equivalent return and allocations are difficult to reckon. Choosing which investments to locate in qualified accounts and non-qualified accounts is also a key decision, since the capacity of qualified accounts available to the investor is usually limited or held by related family members or organizations.

- **Sentiment and heuristics** famously conspire against utility and risk aversion. Daniel Kahneman wrote of many examples in his book *Thinking, Fast and Slow*.[19] Theoretically, institutions should be rational, but political forces and short-term pressures are often greater in an institutional account than a private portfolio.
- **Externally applied constraints** like *stop losses* and *take profits* can be path-dependent and difficult to implement in an optimizer.[20] Frequently, their inclusion requires excessive assumptions and simplifications. They can be applied in the response function during feature engineering, or suboptimally after the optimization.
- **Satellites and portable alpha** are often used to separate asset allocation decisions and their associated constraints. This is similar

to separating strategic and tactical decisions or using siloed asset classes. All of these methods have the advantage of simplifying decisions at the expense of optimality. Though the solutions may be mathematically suboptimal, estimation noise may make optimality irrelevant. Satellite and portable alpha techniques have the advantage of separating high-cost and high-fee strategies from inexpensive ones. This allows allocators to compensate skill and avoid paying for tracking. For instance, a long-only manager can be asked to manage a long/short SMA that contains only the deviations from the benchmark. The allocator can combine this SMA with an inexpensive index fund to produce the same return as the long-only allocation at a fraction of the fee.

Technical constraints are designed to prevent the optimizer from choosing a solution that is impractical in a way that would be obvious to a sensible investor. There are many ways to rein in a misbehaving algorithm.

Cardinality constraints, such as specifying the maximum and minimum number of investments, may require mixed integer solutions that cannot be solved with standard convex optimizers.

Turnover constraints on the average of number of buys and sells are a way to limit costs and taxes without the complicated calculations that are required to measure them precisely. Turnover constraints lack the ability to vary costs based on specifics, like alpha decay, asset class, trade size (market impact), fraction of *average daily volume* (ADV), and holding period. Some managers prefer prioritizing turnover-constrained optimization by signal strength, but it is somewhat counter to the idea of diversification. Sometimes, signal strength is not a predictor of value. Fully optimizing for taxes and other forces will not necessarily reduce turnover and vice versa. Diversification constraints like cardinality and long-only/long-short requirements can be used to force sensitive algorithms to diversify holdings (or to limit churn and positions that don't "move the needle"). If signal strength were predictive enough, diversification would be detrimental because it would dilute the benefit of the signal. Diversification is used to balance the portfolio and spread the investments among many signals. Explicit diversification rewards and penalties are sometimes used, but are more complicated to implement.

Complex metrics are sometimes necessary to describe constraints that may seem obvious to a person but are surprisingly difficult to define precisely. Risk measures (and to a lesser degree return metrics) are notoriously difficult to express and include absolute values or relative measures like tracking error, point values (like means and maximums), probabilistic measures, extreme measures, higher moments, and many more.

Technical complexities are inevitable when applying realistic constraints. These complexities are often expressed in complicated and challenging ways:

- **Interactions** between constraints are difficult to measure. One solution is to measure shadow costs in the Lagrange multiplier of the optimal solution of the objective function.
- **Misalignments** of the alpha and risk model can make constraints imprecise or even perverse.
- **Discretization,** like other trading and position constraints, reflect the reality and friction of real-world implementation, especially for small portfolios that cannot trade round lots. They include transaction size (minimum, maximum, round-lot, fractional) costs and fees.
- **Semicontinuous, conditional, and soft** constraints can be expressed as inequalities instead of as simple levels.
- **Factor mimicking portfolios.** The translation of factors to investments is often ambiguous and polluted by investor constraints. Managing the abstraction that is required for factor-based investing and the inherent *basis risk* is a difficult problem in quantitative finance.

Dealing with Uncertainty and Other Enhancements

No optimizer is perfect for most real-world problems; nearly all practical solutions require compromises. Choosing a specialized optimizer or modifying one may be beyond the ability of many investors. Not only is it expensive to purchase a complex and powerful optimizer, but it requires mastering complicated software with many configurations that are easy to misuse without even knowing it.[21] Sophisticated modern optimization software is built with nuanced programming decisions that may not even be disclosed. Analysts should be wary of the following:

Uncertainty, including *estimation error*, *forecast uncertainty*, and *non-stationarity*. Errors in goal estimates can be far more impactful than errors in other parameters like risk and covariances. Understanding the impact of uncertainty on allocation may guide priorities but does not absolve analysts of addressing the errors in other variables. One example is the tendency for covariances to vary slowly over time, then converge quickly to unity during crises (*stress correlations*). Autocorrelation and non-orthogonality are less intuitive but also important.

Opportunities for improvement. Methods like robust optimization make allocations less dependent on estimation error. Hierarchical risk parity and other methods are being invented all the time. Many techniques have been devised to mitigate inaccurate assumptions. For instance:

- Maximum drawdown can sidestep the autocorrelation sensitivity of some other risk measures.
- The Cornish-Fisher expansion can help with the normality assumption in parametric methods.
- Random matrix theory can help make the covariance matrix more relevant.
- Behavioral finance can help identify the desire to sell liquid assets to raise capital for the next trade rather than sell failed trades or trades that exceed their target.
- Quasi-diagonalization reorganizes the covariance matrix for better results.

Robust methods. Optimizations are often referred to as "error maximizers"[22] because they can be extremely sensitive to small changes in their input parameters (including both estimation error and forecast uncertainty). Hypersensitivity can make an optimizer's recommendations unintuitively unstable and cause their allocation recommendations to be concentrated, greatly emphasizing some holdings with an insignificant numerical advantage. One solution is to evaluate a range of values for the optimization parameters and observe the variation in outcomes. A drawback to this parameter search technique is that the recommendation set will be non-parsimonious. Most, if not all, investments in the original set would be assigned some weight, and the outcome would typically be less than optimal.

Uncertainty sets can be used to identify the “worst case” rather than the average (expected) solution. Take care to avoid outcomes that are so pessimistic that they are unusable. Robust methods can be adapted to multistage problems.

Bootstrapping can be used to resample parameter values from distributions repeatedly to generate a multitude of optimization scenarios that factor in variations in risk, return, and covariances. These deviations from the estimated values produce a range of outcomes that can be used to evaluate the sensitivity of the model. By understanding model behavior, a manager may become more tolerant of outcomes and less likely to revise forecasts and incur the associated costs of new portfolio adjustments.

Resampling usually produces better results out-of-sample even though the resampled “efficient” frontier will not be optimal. The resampling noise can overwhelm the estimate leading to unrealistic outcomes. This is sometimes addressed by incorporating uncertainty in the utility function. Resampling methods can be difficult to implement with certain common constraints, like turnover.

Resampling is dissimilar to shrinkage and Bayesian methods like Black-Litterman. Those methods tend to move estimates toward an average or a proscribed view. While robust methods can be relatively effective in certain cases, despite the extra calculation times, robust methods often underperform shrinkage.

Dynamic programming (stochastic control). Robust methods address an optimizer’s single-period results. But investors often want to optimize multiple goals over time for many reasons, including life goals for wealth clients, projected constituent mixes or plan changes for pensions, and multiperiod implementation for large changes in investment portfolios.

Dynamic methods use *backward induction*, recursively using Bellman’s principle to traverse a multinomial scenario tree from the terminal wealth state to the initial investment. This method reduces all past information to a vector of states. Approximating continuous processes with discrete nodes may introduce *discretization errors*, so a balance must be struck between accuracy and computation time. Because these models can require a many scenarios, practical implementation is often almost impossible. Dynamic programming can solve complex multiperiod problems including transaction cost and tracking error complications.

Stochastic programming (including recourse models) explicitly treats *parameters* as uncertain (non-deterministic). While effective in certain applications, stochastic programming requires a great deal of computing power and can be used for:

- **Parameter estimation,** like risk and return
- **Probabilistic constraints** like VaR (which often involve SOCP to avoid mixed-integer solutions)
- **Multiperiod analysis** like dynamic programming

Synthetic parameter values may not be realistic but allow for experimentation that would be unavailable with historical data alone. For multiperiod analysis, think of each variable as a probability distribution used in a series of successive path-dependent scenarios. One may imagine scenarios organized in a tree structure—the root node would be the present and the terminal leaves would be the realizations at the investment horizon.

Interior nodes can be simulated in a couple of ways: *in parallel*, which can reduce run-time by employing simultaneous simulations, or *sequentially*, which can reduce run-time by requiring fewer simulations by selecting only the most realistic branches and avoiding extremes. If extremes are of interest for determining tail risk, they can be emphasized. Hybrid approaches are also possible. Like value-at-risk models, parameters values can be generated by *assuming distribution functions*, fitting distribution parameters to historical data, or bootstrapping distributions to data, or *by using models* like vector autoregression, PCA, or clustering.

Parameters can be *anticipative*, incorporating past uncertainty, or *adaptive*, with subsequent determinations or *realizations*. Without anticipative parameters, the model would be deterministic, and without adaptive parameters the model would be static.

Recourse models combine anticipative and adaptive parameters, allowing the model to make decisions, such as those needed for rebalancing. Non-anticipativity constraints enforce the scenario structure at each stage so that identical paths produce identical decisions. *Fuzzy sets* extend the concept of stochastic parameter estimation by representing variables as distributions.

Asset Classes

Investing is is largely a social science. Asset classes, with their salience and relevance to current events, dominate allocations partly because investment committee and client interactions frequently degenerate into discussions of asset classes and asset class risk premia.

Nearly every investment is realized using financial instruments (or hedged with them). The US *wealth portfolio* is defined in terms of assets. Financial products are often built with investment instruments that are unrelated to their advertised use and behave in unexpected and complicated ways. Financial engineering often involves creating simple-looking products that are complicated “under the hood,” like CDOs, stablecoins, and volatility ETFs.

Embrace detail and complexity. Quantitative investors favor broad strokes and large data sets to allow them to apply their rules to a multitude of opportunities. Understanding the peculiarities and influences of asset classes on asset allocation can avoid consequential missteps and assumptions.

Equities and fixed income are considered traditional assets that behave in a well-understood way. However, they are neither traditional from a historical perspective nor well understood. Noise, variability, adaption, and complexity make even these basic asset classes befuddling.

Equities. Equity exposure is common and used as a naive metric for portfolio risk. There are many reasons for this approximation, chiefly because it is easy to explain to clients (and this is mostly because it is an often told narrative, not because it is true or particularly easy to understand). The use of equity exposure as a proxy for risk is especially important to investment managers when explaining more complex topics involving risk, like target date funds. Another compelling reason is that naive portfolios, like 60/40 or 1/N, can be frustratingly difficult to outperform after costs and fees.

"Equity" can mean many things and can include a spectrum of risk in absolute terms, from a highly volatile biotech stock to a docile utility. Many other dimensions make the fraction of the portfolio invested in equity a poor surrogate for risk, including *home bias* and the dominance of US markets. Home bias is common abroad as well, though it is often unjustified by relative performance expectations or risk premia; a client's liabilities and goals often need to be financed in the local currency, which may justify home bias. Critically, the relationship between equities and other asset classes is unstable and unreliable as a strategic assumption.

Interest rates. The unrefined interpretation of asset allocation looks to equities for return and fixed income for diversification. Though yields have been trending down steadily for much of recorded history, fixed income return variance is too large, varied by geography and subclass, and too great in duration for most investors to benefit from this trend.

On a practical note, correlations between interest rates and other asset classes vary greatly and often change signs. Correlation hedges performed well for about two decades until the GFC of 2007–2008, when they were primarily used to counterbalance "risk" assets, performed again in 2011, but were reduced to hedges again in 2013, and continued to offer tepid yields for the next decade.

As with equities, the simple, homogeneous model of a fixed-income asset is far from reality. In particular, bond indices are often extremely diverse, debt-weighted by nationality, and issuance-weighted by company —favoring countries and companies with high debt-to-equity. Many managers skew their bond portfolios to take on credit risk in search of yield, especially in the era of "financial repression" and *zero interest rate policy* (ZIRP). Negative implied risk premiums and even negative rate bonds have been issued when governments intervened in markets.

Various equity products and other fixed- income instruments include instruments like convertibles and structured credit products. These instruments can be complex and different from their perception amongst investors. A thorough analysis of a prospectus or offering memorandum may be required to understand these products, and that is a “high hurdle” for most investors (recall Credit Suisse’s CoCos).

Emerging markets and some other illiquid markets are common assets whose behavior deviates from more traditional investments. The efficacy of investing in these asset classes waxes and wanes, as does the general performance of active management. They provide niche opportunities and inefficiencies that can be exploited by a savvy investor, but also have higher costs and fees.

In the mid-nineties, emerging markets (EM) were touted as diversifiers, but we now know that they can become correlated with more liquid markets. EMs lack diversification benefits when they are needed most (their correlations with developed markets increase in crises). Moreover, stressful periods result in significant wrong-way risk by creating illiquidity and volatility.

While they may not always diversify well, EMs are diverse and idiosyncratic. Common simplifying assumptions are unreliable; for instance, it is common to group commodity-driven EMs together, but different geographies may be driven by divergent forces that may not appreciate and depreciate in tandem.

Another common distinction among EM economies is hard currency versus soft currency. Passive investing, or generalizing by broad groups, may be attractive because investing directly in some geographies may be difficult or impossible. Even so, tracking error can be five to ten times higher for passive EM indices (relative to developed markets indices) and may not justify the risk.

Alternatives, currencies, and commodities can provide portfolios with diversification through returns that cannot be built with more staid assets. Some institutions may invest in alternatives (“alts”), but not in the underlying instruments and techniques that alts employ, including leverage, derivatives, and shorting.

Alts often exhibit seasonality, extreme distributions, and low correlations with “traditional” assets. They are also often complex and volatile, requiring specialization. Naive investors may generalize, especially when constructing historical data series, which may lead to poor assumptions. Combining disparate investments because they do not fit neatly in a traditional classification overemphasizes their diversification benefits and makes their future risk and performance difficult to forecast. Some alts, like real estate, hedge funds, and private equity, are appraised or report infrequently, requiring a de-smoothing or liquidity adjustment. Other biases result from self-reporting or poor access to data.

Commodities are heterogeneous and are well diversified when combined. Unlike their well-deserved reputation for seasonality and volatility, they can have a fairly normal return distribution when deployed as a group. Commodities are difficult to apply to more specific investment needs, like liability matching, inflation hedging, and geography proxies.

For instance, copper may be significant for Australian, Chinese, and Indian equities, but not for American stocks, because these economies are in different positions on the supply chain. Australia is directly affected by the price of copper because it sells raw earth to China and India who process and use the metal. The United States purchases consumer and durable goods that use copper, which involve many other costs including labor, transportation, and demand.

Similar representations of commodities may behave differently and may not be appropriate for a particular use (like CPI versus wage inflation for liability hedging). Investable instruments tend to have complex structures and change frequently. Deliverable baskets, carry, and convenience make simple-sounding products dangerous for investors. Even investable indices often encompass the credit risk of the issuing entity.

Commodities are frequently generalized as drivers of other market forces, like EM equities and inflation. The reverse is also true: proxies like mining companies are frequently substituted for commodities to avoid regulatory hurdles, or to improve liquidity and access. Frequently, the assumptions that underlie these decisions are imperfect; though gold mining company stocks and corporate bonds are sometimes used as a proxy for the metal, they are an imperfect match.

Real estate. Chapter 4 discussed the many nuances of alts like real estate, hedge funds, and private equity. These assets are priced infrequently and inconsistently. They often rely on appraisals, comparables, or book value, especially valuations in some popular indices. Some real estate investment trusts (REITs) can be priced based on transactions involving similar properties (*comparables*). Real estate often suffers from overgeneralization by substituting REITs for direct investments and by combining different groups (residential, commercial, etc.) and geographies into the same category. In the opposite way, the assumption of geographic diversification failed catastrophically with correlation products in the GFC of 2007–2008. No perfect solution currently exists for real estate as a class.

Hedge funds are similarly complex and heterogenous. It's common to group by style, but even managers within categories can perform differently. Though funds-of-funds add an extra layer of fees and only encompass a subset of available managers, they do explicitly address some of the problems that afflict hedge fund performance data, including survivorship bias, performance transparency, and audited valuations. For these reasons, if an allocator is forced to treat hedge fund styles as a homogenous group, it may be better to track a fund-of-funds (FoF) index instead of a hedge fund index or even individual funds. While these are all valid and substantial concerns, the most relevant consideration is access to quality managers. Few investors can identify and invest in top-tier funds, and these funds behave differently than the index. Don't invest like Yale unless you can invest like Yale.

Private equity allocation proxies suffer in similar ways to hedge funds but have additional hurdles like being measured by *internal rate of return* (IRR) and *multiple on invested capital* (MOIC), which may be an inappropriate measure of both performance and economic value. There has been some progress in creating benchmarks for optimization that address frequency, reporting lag, and autocorrelation.

Private equity (PE) portfolios are more likely to rely on a small fraction of investments compensating the loss of many others. Objectives are more often defined in terms of terminal wealth than correlation with other investments. Diversification can be attempted in many dimensions including *cohort* (for example, the *J-curve*).

There are different types of PE funds, such as buyout funds, with different J-curve profiles, return distributions, and correlations. Portfolios may suffer style drift as investments mature. It may be particularly important to diversify the liquidation schedule as committed capital becomes distributed capital.

Factors create special challenges due to the abstraction and misalignment involved in addressing the forces that span investable instruments. Mapping drivers to factors and factors to investible assets without a strong chain of causality can result in overfitting and other biases. Improperly applied factor techniques may result in the same output as a simple minimum volatility optimization.

A common method of approximating an integrated factor methodology involves filtering (and possibly ranking) a universe of investments on factors and then optimizing the subset of selected instruments. Factors can also be used explicitly as constraints in the optimization. This can even be accomplished without addressing expected returns and their inherent estimation error.

Time Series Techniques

A well-constructed portfolio founded on factors and diversification will offer protection over time by creating a cushion through price appreciation and income. Staying in the market is generally more effective, easier, and cheaper than timing the market ("time in the market versus timing the market"). Opportunity cost—the other side of the power of compound interest—is a powerful force. Avoiding the worst days is less beneficial than participating in the best days. Market timing has negative edge. Risk management is not fruitless: it is easier to forecast and avoid risk than it is to forecast and earn excess return.

Systematic and integrated portfolio management through time should be incorporated in a sophisticated portfolio. For this reason, it is best to take a layered approach, with increasing difficulty of execution and diminishing efficacy.

Thoughtful portfolio construction is usually accomplished with the cross-sectional techniques discussed earlier. A portfolio's expected return distribution can be reshaped through factor-based, risk-budgeted, diversified asset allocation. Management incentives (such as clawbacks, high-water marks, and other fee structures) are also effective in schematically aligning best efforts with desired outcomes, mobilizing the full potential of the investment team.

There are funds that specialize in a specific asset class or technique that may limit the manager’s ability to use cross-sectional methods. Some funds, including many commodity trading advisors (CTAs), avoid this layer altogether by design. Maximum growth strategies, like the Kelly criterion, use sizing techniques that take advantage of an investor’s probabilistic “edge.” They are popular on Wall Street (e.g., in hedge funds and proprietary desks) but often left out of academic courses and avoided by wealth managers.

Rather than managing risk by diversification, wealth maximization seeks a geometrical increase in wealth by managing deployed capital. However, the high concentrations and large losses that frequently result from this strategy can be difficult for investors to stomach. While most diversified strategies suffer large losses occasionally, the large P&L swings of these methods can lead to ruin if the losses occur before enough wealth accumulates. Practitioners often allocate a suboptimal fraction of the recommended amount (such as the “fractional Kelly” method) to avoid harming a portfolio before it has a chance to gain traction. As with modern portfolio theory, most maximum growth strategies rely on several simplistic assumptions. These strategies can be effective for leveraged, illiquid, and concentrated investments, but are not as practical for large multi-asset portfolios, especially those under public scrutiny.

Routine risk management is used to actively monitor and mitigate day-to-day expected risks before they get out of hand. Low-cost, short-term strategies with an operational edge can smooth the return stream without a significant drag on earnings.

Many of these strategies can be prohibitively expensive in practice and may be impossible to execute when needed most. For instance, shorts can be exorbitant (or impossible) to borrow when they are the most useful.

Maintaining protection over extended periods with derivatives or dynamic hedging are notoriously expensive strategies if used continuously. Many funds, including those managed by arbitrageurs and insurance companies, have been able to profit by taking the other side of these trades (*self-insuring*) and accepting the risk of loss (“picking up dimes in front of bulldozers”) or managing a *matched book* (offsetting positions through scale).

Structured notes and other instruments are designed to make money over time for the investment bank at the expense of the client. In general, buying insurance is only worth the expense if the event being hedged is existential or if the reduction in basis risk is important. Protection sellers, like insurance companies or market makers, reap large profits while diversifying away existential risk. Protection buyers, like homeowners or portfolio managers, may not have that ability or risk tolerance.

To buy protection frugally, some level of systematic timing must be employed. Timing requires enhanced risk management, such as using *risk triggers* that protect the portfolio when proscribed rules are breached. A common trigger is an increase in rolling VaR beyond a certain threshold. *Ratcheting*, such as increasing the strike price on a trailing stop, is also a simple and common trigger used to protect appreciation.

Systematically scaling leverage based on quantitative drivers can be complex. It gets even more complex as managers tweak (and potentially overfit) their data. Active management of liquidity and other risks is a common temptation. Timing techniques are employed to manage the entire portfolio leveraging process, often as an overlay—and often to the detriment of the core systematic strategy. Many good funds have been ruined by hedging overlays that performed poorly out-of-sample. Success in one aspect of investing can lead to overconfidence and tragedy.

Crisis management is the final layer of the process. It is employed to prevent irrecoverable losses, garner windfall profits from tail-risk hedges, and identify opportunistic investments. It is critical that these methods are explainable and defensible. When the “fog of war” rolls in and pain accumulates, so will doubt and criticism. Ad hoc solutions and deviations from planned defenses are a poor substitute for discipline and resolution.

Conversely, understanding the parameters and limitations of our strategy is equally important. If assumptions or investment theses are violated, the strategy should be shut down. Good quantitative investing is based on repeatable exploitation of edge. Relying on luck and hope is a poor substitute for a good plan and is strong evidence of inept management.

Timing strategies rely on good problem specification, like most plans. It’s vital to accurately identify and target the risk measurement, or a combination of measurements. Often risk measures required by regulations are not sufficient to account for investors' reactions to loss. Volatility, skewness, max drawdown, expected shortfall, tracking error, drawdown depth, and drawdown length are among the most important factors. Even small losses over extended periods can lead to large capital outflows.

Broadly speaking, active risk management includes *weighting* schemes like tactical asset allocation, style rotation, rebalancing, and leverage schemes that are either triggered by indicators or seek to target specific metrics by scaling. Some levered strategies can short.

Leverage schemes and, to a lesser extent, weighting schemes can be triggered by *regime indicators* like economic numbers and correlations, or *risk indicators* like drawdown or volatility.

These indicators can be used for different hedging purposes: *integral* hedging methods like delta hedging and managing a matched book (market making); *long-term* cyclical or secular hedging, like economic sentiment; *anticipatory* hedging including regime detection and forecasting, correlation changes, and contagion; *reactive hedging* like mean reversion, trend following, and drawdown reduction or control; and *recovery*, including increasing aggressiveness and reinvesting in core strategies. *Dollar cost averaging* is one of the simplest of these methods.

The application of leverage schemes is varied. *Momentum and mean reversion* techniques can cause great financial damage when the timing is off because they rely on identifying reversals, described as "catching a falling knife." Frequently these methods are reactive. Not only may they result in participating in the maximum loss, as well as hedging or reversing after the maximum loss, but equally damaging, they may also experience re-risking long after the market has begun to rebound.

Dynamic hedging includes volatility targeting and the active management of other Greeks. *Asymmetric* and *product-based solutions* like volatility caps and swaps, puts, spreads, and collars tend to be expensive. *Portfolio insurance schemes* reduce the tails of the P&L distribution by moving the mean left. They include constant proportion portfolio insurance (CPPI), ratcheting strategies like time invariant portfolio protection (TIPP), and value-at-risk insurance methods. For example, CPPI is a simple dynamic asset allocation method that automatically de-risks as the portfolio loses money and increases risk as profits accumulate. A *cushion* of safe assets protects the portfolio from falling below a critical *floor* value. This procedure creates a synthetic put. As attractive as CPPI and other insurance schemes are, they have their weaknesses. For example, CPPI can get stuck at the floor (*cash lock*) or fall through the floor in fast markets (*gap risk*).

■ ■ ■

Thoughtful asset allocation can dominate portfolio performance. Quantitative methods have become nearly unavoidable for institutional portfolios. These methods have many faults that need to be managed with complex and sometimes contradictory techniques. Diversification and optimization are commonplace, but common sense and timing have their place in portfolio management. Discretion is often employed at the most critical times, even if they are not advertised as key elements of a strategy. Good planning is the best defense against poor decision-making, and systematic design and testing form an excellent foundation for navigating the inevitable difficulties that we will discuss in Part IV.

---

1. Japanese investors are less enthusiastic about active management, which is why only 40 percent of equity funds are actively managed in Japan, compared to 70 percent in the United States.
2. An equally weighted portfolio can be used to overweight small cap stocks relative to a capitalization weighted portfolio, but that is only one simple method of overweighting small caps.
3. Harry Markowitz, "Portfolio Selection," *Journal of Finance* 7 (1952), 77–91.
4. As Andrew Carnegie said in an 1891 address at the Pierce School of Business and Shorthand, "Put all your eggs in one basket, and watch that basket."
5. Diversification often involves reducing tolerable risk, like the standard deviation of returns, and can ignore painful risk, like maximum drawdown or CVaR. Critically, adding diversifying investments may not account for out-of-sample risks—the unknown unknowns—but instead it can increase their likelihood by exposing the investor to more unvetted investments rather than a few well-researched ones.
6. Chris Brightman, "Our Investment Beliefs," Research Affiliates, October 2014.
7. The probability of shortfall decreases because realized returns converge to the mean, but the chance of a large loss to occur during the holding period increases because the number of trials increases. That is to say, the chance of an unlikely event occurring is larger, but the probability of recovering from it is higher, provided the loss is not catastrophic.
8. Technical analysis signals are notoriously contradictory. The market technician's art is evaluating when to favor one signal over another and in interpreting the signals before they are fully confirmed. Some consider technicians students of human behavior; see Martin Pring's *Technical Analysis Explained*, 5th ed. (McGraw Hill, 2014).
9. All of these optimization methods have technical and structural tradeoffs but they also have practical compromises. Most perform well only in certain regimes (e.g., high volatility, high growth, or inflation) and poorly in others. Maximum Sharpe ratio, for instance, often performs well during growth periods but not during volatile or inflationary periods. Others, like risk budgeting are suitable to many regimes but do not perform particularly well in general.

10. Group bounds constraints can force traditional optimizers to behave similarly to classification methods. Group constraints can be used for direct indexing and thematic portfolios. You can find examples at www.QuantitativeAssetManagement.com.

11. Subtle and difficult-to-find errors are arguably worse than those that are immediately observable. Obvious errors are embarrassing to miss (or willfully ignored because they are difficult to correct). Errors that are discovered easily by critics can have worse consequences than those that are hard to find.

12. K. Bayer, "Vom traditionellen Management zur Portfolio Factory," in Christoph Kutscher and Günter Schwarz, *Aktives Portfolio Management: Der Anlegeentscheidungsprozess von der Prognose bis zum Portfolio* (Verlag Neue Zilricher Zeitung, 1998).

13. F. Black and R. Litterman, "Asset Allocation Combining Investor Views with Market Equilibrium," *Journal of Fixed Income* 1, no. 2 (September 1991): 7–18.

14. Computation cost is measured by statistics like convergence, FLOPS (floating-point operations), Big O (orders of growth), and NP-hardness (non-deterministic polynomial-time hardness).

15. Just as purchasing an appropriately risky investment is more efficient than combining a volatile one with a straddle hedge, specifying an optimization properly is more efficient than adding constraints to a poorly designed optimization.

16. Qualitative investors (such as investment advisors) often respond to unstable solutions with noncompliance (*implementation error*) because they do not trust the effectiveness of the quantitative trade recommendations. This distrust, combined with an aversion to taxes and fees, makes it difficult for the portfolio manager to entice them to adopt important portfolio adjustments when they are needed.

17. A particularly difficult mandate to manage is an inflation adjusted absolute return mandate, like CPI plus 200 basis points. Managers often resort to risky assets to achieve the spread over inflation, but the mandate implies low risk, steady returns.

18. As a the price of a long position decreases, its allocation decreases also, reducing the risk as the long loses money. As the price of an investment increases, the market value of a short position also increases, adding to its allocation as it loses money.

19. Daniel Kahneman, *Thinking, Fast and Slow* (New York: Farrar, Straus and Giroux, 2011).

20. The triple barrier method is one convenient implementation of stop losses and take profits.

21. A simple example of this is the look-through calculation option for Bloomberg's Port function.

22. Richard O. Michaud, "The Markowitz Optimization Enigma: Is 'Optimized' Optimal?," *Financial Analysts Journal* 45, no. 1 (Jan.–Feb. 1989): 31–42.

13

Security Selection

Details Can Dominate

A sprawling and evolving universe of investment vehicles offers investors the flexibility to look beyond the picked-over set of common, hackneyed investments. Motivated managers can find a niche that has less competition and more edge. It is not uncommon to overlook the variations, even in traditional investments.[1]

When applying quantitative techniques to asset management, asset classes are abstracted as inanimate forces. It is easy to neglect the tenacity of investment managers as expressed through their security selection and trading. Techniques for security selection strategies are even more diverse and fluid than the broad range of investment classes in which they are categorized. Rather than products created by people, investment opportunities are often the result of deliberate actions performed by professionals who are laser-focused and constantly innovating for an advantage. These people are not all perfecting their technique by repetition but are often angling for an edge through improvisation. The smallest sub-strategy can represent someone's entire career of hard work in developing and refining his technique and intuition.

Surveying these strategies allows us to add variety to our toolbox, as well as diversification benefits. Richard Grinold and Ronald Kahn's Fundamental Law of Active Management (FLAM) states that the information ratio is proportional to the square root of breadth—FLAM is the codification of the intuitive concept that more opportunities may result in greater profit if skill is available to exploit them and if the opportunities are different enough from one another. A menu of strategies allows us to segregate investments into products that charge fees appropriate for the skills employed:

- Low-cost *index funds* for class exposure (beta)
- Modest *smart beta* products for style, factor replication, and risk premia
- Expensive *active managers* for "residual" returns
- *Overlays* and hedges to manage the overall portfolio

Categorization gives us a convenient language to describe these activities in an imperfect and theoretical way. We will use one of several common security selection technique taxonomies. We will not attempt the impossible task of putting every strategy in a box. We are only trying to provide a sense of the extent of choices available.

Given these artificial classifications, some data providers have gathered performance statistics; while these time series are flawed, they can provide insight. Some are self-reported and contain survivorship bias, some have backfill bias, and others report incompatible metrics. It is important to remember that these are *average* returns for each category and the benefit of active management is in the right tail. Average active investments are typically not worth the expense and effort.

Not just for quants. Our focus on quantitative asset management does not preclude us from investing in qualitative strategies. We can use quantitative methods to identify and manage investments in qualitative strategies; we will discuss some of these in Chapter 18. Asset allocators, including large institutions, funds of funds, and multi-strategy funds, trading desks, and proprietary trading firms, can optimize investment strategies and allow purely qualitative managers to invest within those assigned limits.

Qualitative techniques are commonly (and often poorly) approximated with *quantamental* algorithms and smart beta methods, whose innovations compress active management fees by commoditizing some techniques that require active human management. The distinction between qualitative and quantitative will continue to blur as tools and techniques advance.[2]

Many market technicians and other quantitative managers follow the rules-based commodity trading advisor (CTA) model developed in the 1970s and reinvigorated in the 1990s. They use systematic automated approaches, often with significant leverage, in the pursuit of momentum, reversal, or contrarian strategies.

Investors in managed futures, short-term traders, and other managers frequently limit themselves to highly liquid and easily accessed instruments, like futures contracts, ETFs, and indices, to trade frequently and realize a beneficially asymmetric return distribution. These strategies often involve overlays and de-risking rules that may force the fund into cash for periods of time. At the other extreme are funds that specialize in direct investments, private deals, and bespoke and esoteric secularization.

Primary distinctions. Following convention, we will discuss these strategies as:

- **Macroeconomically driven and thematic** strategies
- **Cross-sectional** strategies (like arbitrage) and *operational* strategies (like securitization and market making)
- **Multi-strategy funds and funds-of-funds,** which manage groups of managers, often quantitatively and more efficiently than individual allocations but with caveats and extra costs
- **Equity based**, including hedged, neutral, directional, and constrained, along with equity-driven event-driven strategies, which are less likely to be quantitative
- **Event-driven**, including the effects of corporate events on capital structure

Strategy categories frequently overlap. For instance, business valuations can be used in a straight common stock valuation (equity based) and in the evaluation of the entire capital stack (event driven), as well as to determine the best expression of a theme (macro), and to compare with other investments (cross-sectional).

This list excludes many niche strategies, especially those that invest in alternatives like electricity, weather and catastrophe insurance, and collectibles. Many specialized strategies that were not listed above hinge upon tax and legal structures like flow-through entities. Fund structures like real estate investment trusts (REITs) and master limited partnerships (MLPs) are well known and widely used.

Macro and Thematic Strategies

Strategies driven by macroeconomic (*macro* or *global macro*) events have a broadly appealing narrative for allocators but are difficult to time and execute. All asset classes can be employed to express macro views, especially products best suited to sophisticated investors, such as derivatives, currencies, and commodities. Some managers use proxy instruments, like mining companies, rather than more direct expressions.

Both bottom-up and top-down methodologies are used. *Quantitative* top-down processes are well suited to managing these portfolios although *discretionary* non-systematic methodologies are also common. Funds can position for:

- **Future events**, like recessions and BREXIT, the United Kingdom's (UK) exit from the European Union (EU)
- **Themes**, like ESG
- **Trends** and countertrends (like those used by "systematic diversified" funds)

Macro multi-strategy funds can combine systematic and discretionary strategies. Top-down strategies often invest in liquid investments, like sovereign bonds, or less liquid instruments, like credit default swaps. Bottom-up strategies can aggregate large numbers of small models or ideas. Directional and relative value trades (like convergence, stat arb, and dislocation trades) are both common, but directional trades reflect opinions about future economic events rather than the current dislocations in tradeable instruments that relative value managers use.

Systematic factor-based strategies are the focus of this book and include some tactical asset allocation rotation strategies. Systematic strategies often include elements of strategy categories, like macro. As the name implies, they strive to minimize arbitrary meddling.

Thematic funds can be highly constrained like sector funds, or they may be broader and more conceptual. Some investment processes can funnel house views into several thematic products using different constraints and filters (Chapter 6). An example of themes includes:

- **Environmental, social, and governance (ESG) and socially responsible** funds, including alternative energy (like biofuels, fuel cells, electric, fusion, geothermal, hydroelectric, hydrogen, solar, and wind), energy storage, biomass and waste, carbon capture and storage, carbon emissions and trading, clean energy, climate change, energy efficiency (production, transmission, and use including transportation)
- **Emerging and frontier markets,** like *China A-shares*
- **Formula or structured funds** engineered for specific risk and return targets or cash flows
- **Index tracking** funds
- **Insured, guaranteed, and capital protection** for risk adverse investors
- **Legal and financial** contracts and negotiations like litigation outcomes and life settlements
- **Natural resources** including metals and energy
- **Religious- or political**-view-compliant funds like *Islamic* (*Sharia*) funds

Cross-Sectional Strategies

Cross-sectional (*relative value*, or RV) strategies are diverse and sometimes esoteric. RV strategies overlap with other categories like mergers and acquisitions. Many are well suited to quantitative analysis involving vast amounts of data and fleeting arbitrage opportunities. Generally, they benefit from a current and temporary discrepancy between two or more investments.

The relationship may be simple, as in a pure arbitrage between the same investment transacted in different locations, or pair trading between two highly cointegrated stocks. Or they may be complex, as in the Treasury bond basis or VXX (iPath S&P 500 VIX Short-Term Futures ETN) versus VIX (CBOE Volatility Index). Some relative value trades may rely on statistical relationships, while others may be mechanistic. Riskless arbitrage is rare and often requires a large investment in technology and privilege (as in high frequency market makers with co-located infrastructure). Often, low-risk strategies involve the potential for catastrophe due to timing or credit issues, as with short squeezes.

Systematic factor-based strategies belong in this category, including those that discriminate and categorize cross-sectional features. These strategies can be straightforward, like those of CTAs, or more subtle.

Volatility arbitrage is a common feature of both time series and relative value trades in instruments that contain optionality. Frequently, the optionality of an instrument is measured by its implied volatility—a number that must be estimated using a formula and several assumptions. A simple example of a volatility trade is to use the difference between implied volatility and historical (observed) volatility. Volatility trades can be complex and can involve a great deal of math.

Correlation trading is common in many of these strategies, especially highly structured and tranched credit products like collateralized debt obligations (CDO). Structured products are designed for several features in addition to fee generation:

- **Increasing access** to various investors with different risk appetites through technology, like tranching
- **Yield improvement**
- **Credit protection** through technology like overcollateralization

Relative value trades that use these products can involve correlation trading. In this context, the correlation that is being assessed is not between the values of the assets but rather the joint credit loss probability (the chance that assets’ credit will degrade at the same time). Due to data limitations, these analyses often focus on defaults rather than more general credit transitions. Joint defaults can be low-probability events that require sophisticated math, such as various copulas, to estimate.

Correlation traders trade implied correlations in a manner similar to how options traders trade implied volatility, and they manage the higher order Greeks similarly. Correlation trading is notable for its role in the Great Financial Crisis (GFC) of 2007–2008.

Fixed-income arbitrage (curve trades, correlations, spreads, volatility) is common due to the great variety and technical nature of fixed-income investments. *Asset-backed strategies* (ABS) are flexible and often clever. Borrowers can include companies that cannot find low-cost financing through traditional methods (like bank loans). Their value is often determined by a good purchase price rather than high quality, and is specific to highly technical details like vintage, pool, and servicer. Mortgages and real estate–backed strategies are in a separate category and are not included in ABS. ABS are supported by cash flows and collateral including:

- **Tangible** assets that are either fixed (machinery, land, timber) or non-fixed (cars, boats, aircraft)
- **Future cash flows** (residuals), which may include stock (unrealized harvests) or trade between a commodity-producing emerging market (EM) country and a developed markets (DM) consumer
- **Intellectual property** (entertainment, patents)
- **Financial assets** (loans, credit cards, receivables, life insurance), which may include bridge or other short-term financing for real estate or companies
- **Derivatives** of those assets (pools, portfolios, collateralized loan obligations or CLOs, CDOs, or collateralized bond obligations [CBOs])

Capital structure, credit arbitrage, and high yield focus on the convergence of corporate rates using relationships and instruments like curve trades, credit default swaps (CDSs), basis, senior and subordinated debt, term structure, and structured positions. The strategy can be RV or long-short.

Convertible arbitrage exploits the relationship between convertibles (and their embedded option) and non-convertible instruments (equity, debt, rates, volatility), and can involve synthetic convertibles. Convertible bond warrants are frequently illiquid. The value of the call and redemption feature are more often the focus of this strategy than the underlying equity. Convertibles famously displayed their illiquidity risk in 2008 when the shorting of financial stocks was suspended.

Corporate bond arbitrage trades the spread between corporate bond yields, corporate bond yield curves, and benchmark yields like those of sovereign bonds.

Emerging market debt is often considered an RV strategy. The illiquid and esoteric nature of these instruments leads many investors to assume that expert local advice is critical in this strategy. Some research alleges that macroeconomic analysis, rather than nuanced instrument-level analysis, dominates EM excess returns.

Mortgage-backed securities (MBS) strategies are similar to ABS strategies but use mortgages as collateral. The MBS market is large and technical. Many variations and subtleties are reflected in various issuances and the structures and derivatives built from them. Residential MBS (RMBS) are generally backed by home loans that are supported by government agencies or issued by private institutions. Commercial MBS (CMBS) are often considered part of the real estate RV category.

Sovereign bond arbitrage exploits supply, demand, uncertainty, and other highly technical aspects of sovereign bonds. These trades may involve isolating forward rates, reconstituting an issuance that had been stripped and sold in parts, cash-futures basis relationships, and other derivatives strategies.

Volatility strategies can be any combination of directional, market neutral, or arbitrage opportunities and can exploit the difference between realized, implied, and expected volatility. Expressions usually involve imperfect and complex derivatives. Hedging can be complex, non-linear, and dynamic.

Alternative strategies are also numerous, and include:

Collectibles and cryptocurrencies are popular for reasons that may be economic but often are not. The skills required to be successful are different from most traditional investing activities. “Liquid alts” may attempt to replicate collectibles and other investments but frequently lack fidelity. Similarly, vehicles intended to trade in traditional markets attempt to replicate crypto instruments. The converse is true as well: stablecoins may seem like traditional assets (e.g., money market funds) expressed as crypto instruments, though their failure exemplifies the difficulty inherent in replication methods.

Commodities are usually considered global macro trades, aside from the trade financing ABS transactions discussed previously. They are often touted as inflation hedges. Many investors try to cut corners by investing in more convenient vehicles, like ETFs and mining companies, but these are poor substitutes for pure commodities trades.

Currencies can be used for speculation or hedging. Speculation is frequently based on economic, technical, or time series methods. Currency arbitrage can be similar to sovereign bond arbitrage, since the more popular currencies are exceedingly liquid and precisely priced ("priced to perfection"). The *carry trade* is a common strategy in countries like Japan, where investors can borrow low-yielding yen and invest in high-yielding dollars. The variance of currencies is frequently greater than the rates component of foreign-denominated debt. Many choose to hedge the currency risk of foreign debt, but not more volatile equity or real estate exposure, supposing that the benefit of foreign economic appreciation is expressed in the currency as well as the equity component.

Energy strategies—distinct from commodities in general—typically seek to take advantage of discrepancies in investment vehicles, infrastructure investments, or the supply chain ("upstream," "midstream," and "downstream"). They often involve MLPs and can be isolated from equity market direction (either equity, rates, or commodities). They may involve sub-industries, such as drilling and mining, transportation, utilities, and power generation.

Real estate RV strategies attempt to profit from the yield differences of various real estate investment vehicles. These may include direct investments in residential or commercial real estate, as well as REITS, but not mortgages or mortgage-backed instruments (which have their own category), or other assets and asset-backed securities.

Operational strategies like securitization and market making can also employ cross-sectional aspects. Usually, this manifests itself through the converse of identifying dislocations (by diversifying them away).

Multi-Strategy Funds and Funds-of-Funds

Many allocators combine external and internal managers with other investments to build their portfolios. Exposure to these managers through third parties like multi-strategy funds and funds-of-funds (FoFs) can be beneficial. They may be:

- **Diversified** or **concentrated** (thematic)
- **Conservative** (market neutral or arbitrage)
- **Defensive** (short or countercyclical)
- **Strategic** (opportunistic, equity hedge, emerging or frontier markets)

Funds of Funds allow allocators to outsource some or all of their external manager sourcing, due diligence, fee negotiation, oversight, reporting, hiring, and firing. They offer indirect investment in external managers through such vehicles as limited partnership shares.

Multi-Strategy Funds are similar to FoFs but have greater control and access to their managers, who are generally employees of the fund, as in proprietary trading firms. Rather than relying on limited data, multi-strategy funds may have real-time reporting of all positions monitored in a “war room” or “control room” and with rapid recourse. These firms tend to be quantitatively managed to take full advantage of information transparency. One disadvantage they face is the need to employ their managers in-house. Employing fund managers, rather than participating in external funds, involves managing a limited supply of qualified talent and a more cumbersome hiring and firing process.

There are many advantages to using FoFs and multi-strategy funds. It may seem inconsistent for an investor to desire external managers to handle their money without the benefits of professional FoF managers to choose and manage the external funds.

> **Access to managers** that are soft- and hard-closed. Investing in external managers is most useful when an allocator can pick the best mangers, but the best managers do not often accept new investors (*soft-closed*) or any increase in allocation (*hard-closed*). A fund with an existing investment in a manager may be able to increase investments in soft-closed funds. Similarly, allocators can buy shares in FoFs and multi-strategy funds to get access to hard-closed managers. FoFs can invest modestly in a new manager and add to the soft-closed investment if the manager succeeds. This can be viewed as a *real option*[3] on the future performance of a fund, and can provide a competitive advantage in "picking winners."
>
> **Outsourcing overhead** benefits from economies of scale. It lowers entry and exit barriers, especially for smaller allocators and allocators new to a strategy. Outsourcing can offer superior resources and talent at a fraction of the cost. *Economies of scale* extend beyond shared resources. Examples include better negotiating leverage for fees, *sidecars*, and *separately managed accounts* (SMAs).
>
> **Increased liquidity and decreased minimum investment size** (often less than $50,000) is attractive for the same reasons as access to managers. While the managers that these funds invest in may have long lock-ups and restrictions, FoFs may provide much more frequent liquidity by exchanging shares between fund investors or by making up a deficit in assets through internal resources or loans.
>
> **Diversification** benefits from a shared portfolio of investments. For many allocators, the minimum investment in an external private manager may be too large to allow them to build a well-diversified

portfolio. Cover for those afraid of being “wrong and alone” can be a powerful motivator, especially among less ambitious allocators, such as some bureaucratic pension managers who benefit little from increased performance but suffer greatly if they are singled out for poor decision-making (*headline risk*).

Disadvantages to using FoFs and multi-strategy funds include:

An extra layer of fees and those fees are spent on third parties rather than internal sustainable capabilities. This “rent versus own” decision can be unattractive.

Limited influence over external investment activities has pluses and minuses. Control comes with obligation and responsibility. Internal management has nowhere to hide when results are poor or criminal activity by external managers is discovered (“headline risk”).

Increased liquidity can allow the “weak hands” of other investors in the FoF to flee the fund. Redemptions may force the manager to sell good liquid assets or funds (rather than poor illiquid ones) to raise capital.[4] This can leave remaining investors with a degraded portfolio. Funds may also need to borrow money. Influential investors can negotiate segregated portfolios with managers to avoid this.

Overdiversification, “groupthink,” and *herding* can poison nearly any investment process. FoFs who cater to many clients can succumb to the desire to keep everyone happy.

Another layer of lock-ups and gates may be added by a FoF. These restrictions tend to be more lenient than in the managers that make up the FoF or multi-strategy portfolio. The FoF lock-ups and gates may also obviate or mitigate the first layer. However, they can be more stringent; a FoF or multi-strat fund may lock up or gate even when the constituent funds do not. When more lenient terms are offered by a FoF, they may be contingent, e.g., “best efforts.”

Equity-Based Strategies

Equity strategies need not restrict themselves to common stock and may include investments throughout and outside the capital structure of companies. They are generally categorized as *long-only funds*, which are the most common offering (there are many equity funds that are not long-only). Equity funds buy stocks that are in indices or selected with valuation techniques that may be as simple as positive or negative screens, sampling, or business valuation. *Short bias* funds are often different from long funds in both technique and operation. Shorting is not the mirror image of purchasing and tends to burden the investor in several ways, including finding the instrument to borrow, owing the dividend to the lender, having the instrument called back at the worst times, regulatory restrictions (like banning and limitations on transactions, including the *uptick rule*[5]) and adverse taxation. For instance, in Australia, profits on shorts are taxed, but there is no credit for losses.[6]

Long-short funds, often labeled *equity hedge*, are net long and short, based on the manager's opinions and preferences, but are not mandated to have a specific market exposure. Their beta exposure is often 50 percent (for instance, a long-short fund may have 75 percent exposure to long positions and 25 percent exposure to shorts to net to 50 percent) but can be leveraged two times or more, or be net short. Long-short managers are typically "stock pickers" whose positions are concentrated in fewer than 100 stocks. This is in contrast to *market neutral* funds that generally invest in large portfolios with more than 100 stocks.

In Chapter 12 we pointed out that the long-only constraint is one of the most damaging because it limits a manager's ability to express a negative opinion. An absolute return manager with a long-only constraint is unable to express any negative opinion. A relative return manager can only express a negative opinion to the extent that the investment is represented in his benchmark.

Equity non-hedge makes explicit use of market direction and may be focused on a precise theme, sector, or small group of stocks. *Market neutral* funds are typically more diversified than long-short funds and have less than 10 percent correlation with market movement (beta). This "neutrality" isolates the fund's performance and investment themes from the broader economic forces. Managers may choose to be market neutral so they can focus on security selection for arbitrage or short-term trading.

Market neutral is more common in fixed-income and quantitative investing, where relationships between investments are more precise. Arbitrageurs may exploit an inefficiency, dislocation, or anomaly, like a basis trade, or a well-defined opinion, such as a mispriced forward rate that implies an unlikely Fed rate hike. These precise opinions are usually the result of a “bottom-up” analysis. Trade expressions that involve more intuition—like broad macroeconomic themes or mergers and acquisitions—have their own classification.

Pairs trading is a form of *statistical arbitrage* (“stat arb”) that buys and sells company exposure to profit from relative valuations. It is important to note that it is common to profit from both a long position and offsetting short in the same trade. While some funds are focused on a few thoughtful trades, most stat arb portfolios (“books”) have hundreds of investments with high turnover and corresponding costs. Stat arb books often contain more longs than shorts because of the relative difficulty in shorting. *Index arbitrage* is a related strategy that offsets a portfolio, such as the S&P 500 index, with its constituents, such as the underlying stocks. Often this involves sampling and partial replication due to the effort and cost required to fully replicate the reference portfolio.

Directional and non-directional categorization is related to the ability to short but refers to the intentional bias to participate in beta. A neutral fund may periodically become long when there are more buying opportunities than shorting opportunities, if the managers are bullish on the market, or if they want to reduce tracking error. As the name implies, *quantitative directional* funds use models to discriminate and include techniques like statistical arbitrage, which may be better categorized by some as relative value (cross-sectional).

Factor-based, thematic, and sector funds employ factors like growth or value, themes like climate change, or sectors like healthcare. *Emerging and frontier markets* are a specialty as well. These concepts are all used in equity funds—especially sector investing—but are not limited to equity instruments.

Business valuation. Quants have practiced the traditional Graham and Dodd method of security valuation for decades. Factor analysis (*quantamental*; see Chapter 7), combined with other alternative data, like regulatory filings and social media sentiment, chip away at the human edge in company analysis. It is unlikely that the best analysts will be replaced by machines anytime soon (many more will find cover under noise and obfuscation). Wall Street exists, in part, to maximize compensation for employees. Employees are protected by their relationships with their managers and their clients, which affords them the luxury of underperformance.

Data can be misleading in many ways. Good due diligence and a simple “strengths, weaknesses, opportunities, and threats” (SWOT)[7] analysis can sometimes best the most sophisticated and well-financed algorithmic strategies. Whether performed by a computer or a person, strategy analysis is essential to many long-term investments. Some items that are typically considered include:

- **Economic, regulatory, and industry environment,** including the stage of the business cycle, cost of materials, and supply chains
- **Business model,** including the value proposition, products or services, resources, competencies and advantages, reputation, production, and distribution
- **Competition** including barriers to entry and exit, market size and access, potential for growth (green, brown, blue), differentiation, scale, reputation, lobbying, rivalry, demand, substitution, bargaining power, future technology, and consumer preferences
- **Operational** including access to and cost of capital, debt burden and credit rating, liquidity, cushion, free cash flow, and efficiency

Event-Driven Strategies

Event-driven strategies typically require complex assessments of corporate actions and events. These changes are usually triggered or catalyzed by external circumstances.

Determining the primary driver of value may require an assessment of the effect of uncertainty created by an announcement (rather than an actual event). Understanding the relevance of news and its affect on price may involve operational, legal, and regulatory expertise, which may require outside consultants. An event, catalyst, optimal expression, horizon, and exit must be identified. As a result, event-driven investments are often evaluated based on fundamentals rather than other quantitative methods. These changes may emanate directly from decisions by the managers, which may impact the value of equity, debt, or derivatives, including:

- **Capital structure** changes like debt and equity offerings, buybacks, tenders, and exchanges
- **Mergers, acquisitions, spinoffs, and restructurings**
- **Changes in credit quality** and their effect on the existing and future cost of capital, such as downgrades and bankruptcy

Event-driven strategies are categorized by the type of specialized skills needed and markets that they exploit. Event-driven groupings are broadly divided into several areas, including:

Merger arbitrage (M&A) or risk arbitrage strategies are low-leverage strategies that typically focus specifically on the M&A announcement and take positions in the most responsive or impacted parts of the capital structures of the firms involved. Actions can be friendly or hostile, and involve parts of or entire companies, as well as cash or transfers of investment vehicles. Commonly, the acquiring company must entice the target to sell by making a high bid. As a result, the equity value of the buyer will generally decrease and the value of the acquired equity will generally increase. For a share transaction, a manager may short the buyer and go long the target. For cash transactions, usually only the long position is held, and the trade is unhedged. Deals that appear overpriced can be exploited with the opposite positioning. Analyses may require expertise in regulations in different geographies for international transactions. M&A is a low-beta, cyclical strategy with high volatility around the event time.

Activists attempt to influence corporate events like spinoffs, dividends, and buybacks through actions like management change, operational streamlining, synergies, and efficiencies. They may acquire influence through board seats or alliances with other shareholders (while overcoming “blocking factors”).[8] Targets may have financial problems such as assets that are not properly reflected in the company valuation. Targets may have operational issues, including poor strategy, deviation from core competencies, agency problems like nepotistic management, or inept execution. Solutions include management changes, reorganizations, cost-cutting, and spinoffs. Managers are often concentrated in about a dozen companies, some of which are inactive. Positions are typically unhedged, and interactions may be unfriendly, confrontational, or publicly exposed.

Distressed strategies and credit arbitrage involves corporate credit that is heavily discounted due to financial, management, operating, or legal troubles. These strategies may involve issuers who are considering or navigating reorganization, bankruptcy, or restructuring. Distressed and credit trades often require expertise in legal and regulatory matters, including bankruptcy, reorganization, and liquidation. Most distressed investing is related to American companies and a result of US bankruptcy and reorganization laws. Managers often hold debt positions but may invest in any instruments inside or outside the company’s capital structure, including options. *Reorganizations* can deleverage the capital structure by negotiating and exchanging debt and other instruments for new securities and may involve activists, including obtaining seats on the creditor committee. *Arbitrage* strategies are typically expressed as fully hedged combinations of debt, such as senior and subordinated bonds. Aggressive tactics include offering rescue financing with a better claim on assets than existing (legacy) financing, leaving the original lenders with a poor recovery.

Private issue, Regulation D, PIPE, and SPAC funds often seek to capitalize on an illiquidity premium, information asymmetry, or constrained access to capital or markets. Upward valuations may result from a catalyst like an exit (public offering, merger, or acquisition) or company growth (like credit improvement, increased viability, or exiting bankruptcy) that results in the appreciation or conversion of existing investment instruments. This often requires a strong understanding of business valuations.

Special situations is a term for event-driven strategies that are not included in the other categories discussed. They are usually equity centric and specific to a single company.

■ ■ ■

Security selection may encompass many human decisions and behaviors, as well as many automated actions. The distinction between qualitative and quantitative has become less clear over time and will continue to blur as tools and techniques advance. Some funds focus exclusively on selection, while others avoid selection entirely, focusing strictly on asset allocation. Frequently, security selection is too difficult to incorporate in a holistic portfolio optimization and is treated as a separate step, either prior to or after optimization. While this is not always optimal, it can make the asset allocation procedure tractable. Separating security selection also allows the selection to change dynamically between portfolio allocation rebalancings. In many large portfolios, external or internal managers are allowed to alter their selection at will within the constraints of the portfolio's last optimization.

---

1. Even the SEC neglected to consider out-of-mind investments in 2020 when it revised rule 15c2-11 with pink sheet stocks in mind and neglected to address the over-the-counter bond market.

2. It is already clear that large language models are successful at doing research that was the exclusive purview of humans only a short time ago.

3. Stewart C. Myers, "Determinants of Corporate Borrowing," *Journal of Financial Economics* 5 (1977): 147–175.

4. Recall our discussion about the difference in fund structure between ETFs and mutual funds in Chapter 1.

5. Elizabeth M. Murphy, Amendments to Regulation SHO, 17 CFR PART 242 Release No. 34-61595; File No. S7-08-09 RIN 3235-AK35, United States Securities and Exchange Commission (SEC), 2010.

6. Some of the difficulties and expenses involved in shorting can be mitigated with derivatives, such as a total return swap (TRS). But, the swap counterparty adds counterparty risk. Also, since the counterparty may be hedging with the underlying stocks, TRSs may have many of the same difficulties as regular shorts. TRSs were discussed in Chapter 4.

7. The source of the term SWOT is uncertain, but likely first used by Albert Humphrey in the 1960s.

8. These efforts may exploit technical details, like the mechanics of voting (who votes, whether nonvoting shares count for or against, whether voting is cumulative or not, what constitutes enough votes, etc.). The companies themselves, may invent novel solutions for capital raising and voting control including innovative equity offerings like APEs (AMC Entertainment Holding, Incorporated’s Preferred Equity Units), PRPLs (Purple Innovation, Incorporated’s Proportional Representation Preferred Linked Stocks), or Bed Bath & Beyond Inc.’s warrants on Series A Convertible Preferred Stock. To read more about these events, we recommend the excellent and witty newsletter “Money Stuff,” by Matt Levine.

14

Backtesting

Predicting Risk and Performance

Over the past 13 chapters, we have defined our product, planned our process, girded it with governance, gathered our data, identified salient features, modeled our risk and return, and combined our models into investable recommendations. This journey required many exploratory decisions and approximations, so it would be prudent to test them under conditions that are more realistic or predictive than we were able to while building the model. Backtesting in quantitative financial analysis serves a similar purpose to clinical drug trials. We want to examine our virtual creation in a laboratory setting before unleashing it on the world.

Backtesting is essential for building trust and confidence in our strategy for ourselves, our investors, and our stakeholders. It is often derided for the uncertainty inherent in any ambitious undertaking. Nonetheless, backtesting is worthwhile for:

- **Improving logic.** We can test our models with more rigor, complexity, and detail than was possible while building them.
- **Building a pipeline of innovation.** A robust, reliable research process depends on an "assembly line" of strategy and a corresponding quality control mechanism.
- **Stress testing and sensitivity analysis.** Testing our models with manufactured and historical data sets exposes faults and weaknesses in our models and allows us to apply the *scientific method*[1] to our experiment.
- **Increasing trading efficiency.** Streamlining and squeezing every bit of profit out of a strategy and process can be the difference between a successful business and one that just looks good on paper.

- **Identifying limitations.** Using these tests identifies the boundaries of our operating envelope and creates expectations of operating results (Chapter 18).
- **Optimizing maintenance.** Planning efficient operation of our portfolio through practices like execution, rebalancing, and transition management (Chapter 16).
- **Planning for contingencies.** Create contingency plans for managing the portfolio under stress, when the market is in turmoil, clients are upset, and the plan seems questionable (Chapter 19).
- **Facilitating execution and order routing.** Build and test a computer program for use in live trading.[2]
- **Monitoring, reporting, and attribution.** The back-end of the backtesting analysis reports performance and risk statistics and compares these marks to historical values and trends; this same logic can create a dashboard to monitor live execution in context.

Features of Backtesters

Build or buy? All simulations are imperfect and rely on assumptions. The best reason to build a simulation is to gain a thorough understanding of its assumptions and repercussions.

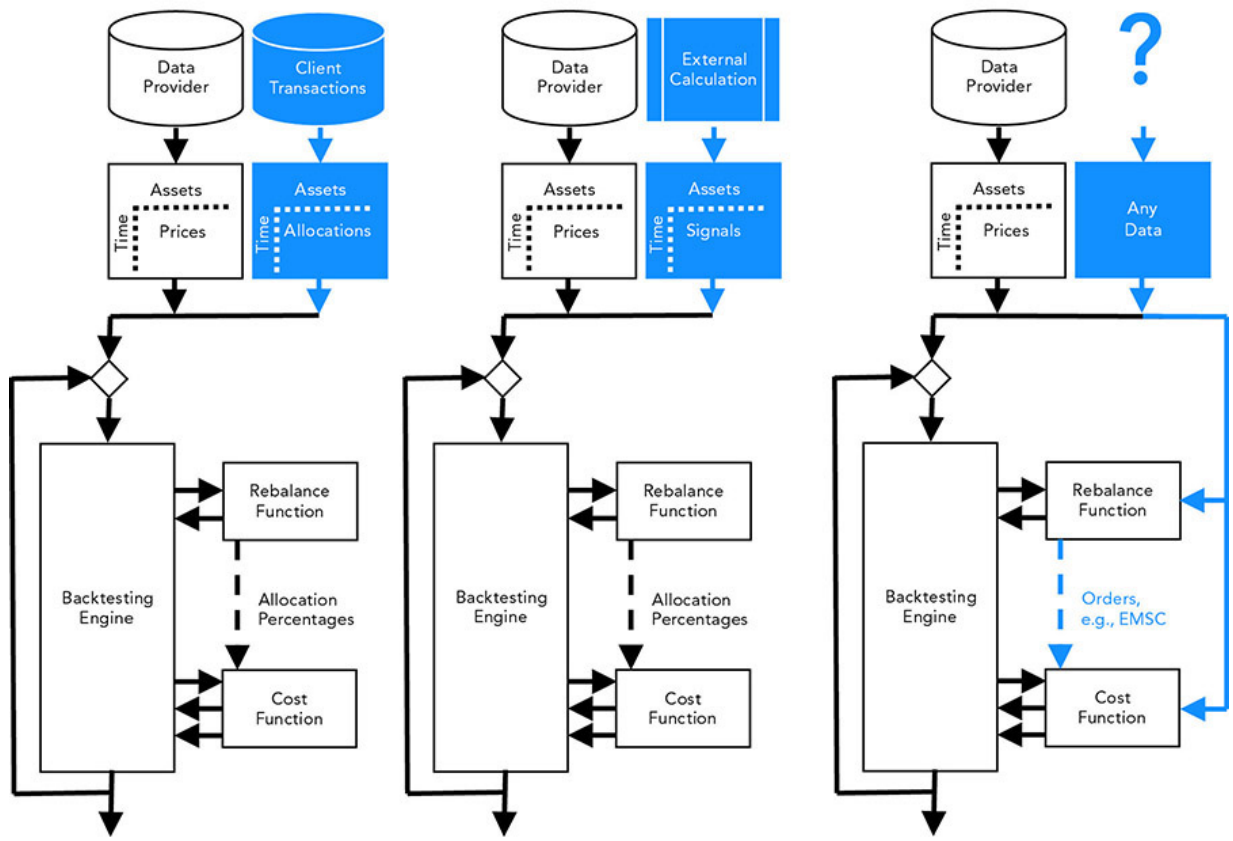
The ability to tailor a simulation to answer specific questions is essential. Although commercial developers spend a great deal of effort making their software flexible, uniform, and full of features, most of this effort is unnecessary for a competent quant team. Purchasing expensive third-party software can also require expensive operational procedures and frequently results in unsupported software forks as proprietary modifications are made.[3] A much smaller set of valuable features (that are necessary to the analyst at the time) and a cumbersome interface is far more efficient and affordable than a user interface built to be “all things to everyone.” Tailoring can provide a competitive advantage, an additional source of alpha, or a crucial feature.

Model error and operational error. Commercial software has a host of drawbacks; it is not entirely transparent, is often restrictive, and can be so complicated that operational error becomes a significant risk. If the modeler is not an expert, model error is of paramount concern.

Compare the results of proprietary models to commercial products to uncover both the overlooked details of the former and the assumptions (and errors) of the latter.[4] Building models to mimic what is available and then expanding on that foundation can start development on firm footing.

Integration, efficiency, and economic intuition. A good, reusable backtesting simulation must be integrated with data systems and research tools, upstream and downstream trading, execution procedures, and attribution tools. If the code is intended for trading after testing, it must be efficient. These requirements benefit from a technical sophistication that researchers and portfolio managers often overlook. Though technology is important, the modeler must understand market forces and execution dynamics. This domain knowledge may be intuitive to nontechnical traders but is rare among programmers, who too often seek a "one size fits all" technical solution. Take the time to be intimate with the data and the causal chain. It is intuitive and helpful to think about agents, rather than procedures and money transfers instead of returns. If agents are not suitable, the next best option is to think in terms of cash flows rather than more abstract concepts like returns.

Batch versus event-driven. A significant distinction when categorizing backtesters is whether they are designed to run in a preprocessed *batch mode* (Figure 14-1A) or online in *event-driven mode* (Figure 14-1B and C). Batch mode backtesters are often seen as unsophisticated tools, whereas event-driven simulations are more flexible and realistic. There are excellent reasons for using both modes; finding a simulator that can use them simultaneously would have formidable advantages.[5]



(A) Stylized backtesting flowchart used to replicate precalculated allocations (batch mode)

(B) Stylized backtesting flowchart used to generate allocations from signals (event-drive or hybrid mode)

(C) Stylized backtesting flowchart with an OMS like EMSC (event-drive or hybrid mode)

FIGURE 14-1 Operating modes of a backtester

Batch mode, also known as *vectorized* backtesting, has several distinct advantages. It is *simpler and more transparent* than event-driven backtesting. Designing and building a standalone simulation with limited capabilities is a relatively straightforward task. Unsophisticated backtests are usually written in batch mode for simplicity, and many are even written with simple tools like Microsoft Excel. Batch mode backtests are easily *scalable* and easily adapted to run in the cloud with nearly unlimited computational power. A well-designed event-driven simulation can also use these resources. Batch mode backtests are also *efficient*. Batch mode simulations can be preprocessed and need not execute at analysis time. They may run at any time, often in stages. This can be a weakness as well as a strength. As we will see, the ability of event-driven simulations to mimic “live” execution can be useful (e.g., with the EMSC object). Vectorized code is also more easily compiled, parallelized, and generally faster than event-driven code, which typically must be run sequentially.

Finally, backtests are *reusable and produce additive results*. Preprocessed results can be reused and recombined without being rerun. They may even include *transfer learning*.

Event-driven mode mimics human (agent) behavior that reacts to events like price changes, economic releases, transactions triggered by prior events, or other signals. Event-driven simulations can be *more complex and realistic*. Event-driven techniques provide the ability to incorporate many contingent and multiperiod details that are difficult to model in a vectorized format. *Interactions* between assets and across time periods, and path dependent calculations, are more straightforward to model with event-driven techniques than with a vectorized format. Common actions, like multiperiod execution, contingent orders, and hedges, are easy to implement with event-driven models.

For automated strategies, event-driven backtesting provides the valuable benefit of *reusable, consistent, and reliable code*. Event-driven code mimics sequential and contingent order execution and incorporates logic that can be "lifted out" and used for execution. Frequently, the same code can facilitate research, testing, live execution and order routing, and performance attribution. Reuse minimizes translation error when replicating the investment logic for different implementations. Highly optimized and high-frequency code may be too challenging to build for reuse. Because event-driven code is *modular*, different routines can be remotely located. They may be co-located at an exchange, or they may reside in the cloud for scalable resources and more robust protection from local continuity problems like power outages and internet failures.

Hybrid execution (backtests that combine vectorized and event-driven modes) can reduce simulation and development time. It can reduce the learning curve for current development, and when using legacy code, allow strategists to focus their time and energy on innovating and researching profitable strategies. Hybrid execution offers:

- **Preprocessing** for time-consuming analyses.
- **Integration** of many reusable studies.
- **Convenience.** Hybrid backtesters can combine disjointed analyses involving technologies that do not integrate easily.

- **Workflow.** Hybrid backtesters can combine the efforts of people across multiple technologies, geographic areas, and time zones.
- **Best of both.** All the benefits of event-driven models (noted above).

Interpretability is a primary outcome of backtesting. Because most backtesters are designed for specific circumstances, researchers may skip features that could be essential for attributing results. Some common oversimplifications include:

- **Record keeping** at the account level rather than for each security.
- **Using periods** instead of actual dates may lead to imprecise timing rules.[6]
- **Simple leverage and cash reserves,** instead of sophisticated cash management, including cash sweep and overdraft for cash drag, margining, and leverage.[7]
- **A focus on statistical measures** of risk and return rather than decision-centric attribution analysis, like Brinson's method of performance attribution.

Rookie mistakes. Avoidable errors are manyfold, but the most egregious include:

- Overfitting and "p-hacking"
- Look-ahead bias and data snooping
- Survivorship bias
- Muddling return distributions and predictive power
- Costs, fees, and capacity limitations, especially when shorting

Major Components

A backtester can be viewed as a combination of agents corresponding to market participants (Figure 14-1). An agent-based structure helps us visualize what we are simulating. Imagining the agents' goals and challenges helps prevent errors in the same way that valuation benefits from cash-flow modeling.

Since simulation requires data, modeling the actions of *the fund's information technology* (IT) group is a common first step. For event-driven systems, this data should be provided in a feed that is an accurate analog of live trading, including point-in-time (archival) data unadjusted for splits and dividends—just the way the algorithm would receive it when it is in "live" operation. The feed may also contain order book information and signals, including preprocessed output from vectorized simulations. For backtesters designed to create reusable strategies that can be turned on "live," accurate simulations of data feeds from many authentic providers can be a substantial undertaking, including the inclusion of divergent and complex structures and fields. The *data handler* may include proprietary cleaning and aggregating tools. The *backtesting engine* iterates and keeps track of performance to mimic the services of *the custodian (transactions), bank (borrowing and lending), and back office (accounting)*. It is the backbone of the simulation and does not vary with strategy. The simplest versions may keep track of profit and loss, but more sophisticated implementations keep track of *tax lots* and account location, costs and fees (including path-dependent fees involving *hurdles* and *high-water marks*), separately managed accounts (SMAs), and other complexities.

This module also may keep track of leverage, cash drag, overdraft fees, and lending and borrowing. Simple versions may charge the risk-free rate, while others account for time-varying lending and borrowing rates for excess cash and leverage. Accurate accounting for these rates is vital for long simulations that combine periods when rates were high (as in the early 1980s) and low (as during the era of "financial repression" and zero interest rate policy [ZIRP] from 2008 through 2015). These mechanisms, which should mirror the circumstances of the fund, can be complex. For instance, borrowing may be automatic (up to a limit), but repaying the credit line may require an explicit decision by the fund and, thus, an action by the fund module, since the use of leverage may be intentional.

The fund (rebalance function in Figure 14.1) is usually represented by a single step, but it is really composed of several agents. If properly written and thoroughly tested, the fund module can be "lifted out" and reused for live trading. This module can be iterated and run at every step, or an event manager can trigger the portfolio manager module.

The portfolio manager and *investment team* transmit the strategy response. It is important to have a feedback loop that accesses the execution results. Investment decisions are commonly made based on the composition of the portfolio and account-level profit and loss, rather than just market levels. The trading team or the broker may choose not to execute a request or to partially execute it. In some cases, the price received may differ from the pricing information available at the order time (*implementation shortfall*). The portfolio management module may have predictive algorithms that forecast these values, including *market impact*, and may adjust predictions based on the feedback provided by the broker's execution (the *market model*). This information can also be used for cost-efficient *routing* in the execution module.

The trading team chooses the trading decisions (the *execution model*), including timing, periodization, buying, selling, canceling, leverage, and sizing. Order logic can be complex; fortunately, exchanges have codified orders. Different exchanges may have different orders.[8] These details may be necessary for precise simulations. The *risk management function* can be contained in the portfolio management module since it generates decisions based on a risk management strategy. Risk strategy may involve path-dependent look-back and forward-looking predictive measures, like the investment strategy. The risk logic of the firm should be modeled, including triggers and regulatory requirements. Risk metrics may affect the cost of capital through borrowing rates and availability of leverage and shorts.

The cost or benefit of risk management extends beyond attribution statistics. Risk also shows up in the risk-free rate used to calculate performance through metrics like risk-adjusted return on capital (RAROC) or in performance incentives by way of the hurdle rate. *The broker* (cost function in Figure 14.1) simulates the market response to the orders created by the portfolio management module. It should be reactive and should not incorporate any actions defined by the investment or risk strategy. It "executes" orders based on a stream of data and simulated market reactions, like market impact, and should be smart enough to provide a realistic simulation of events. For instance, it might provide only a partial fill to an order that exceeds a large portion of the daily volume in that investment. A broker function capable of multiperiod actions may continue to execute a partially filled order in the next iteration until it is complete or canceled. Since the execution of an order may differ from the intention, it is valuable to provide feedback to the portfolio management module for the next iteration.

Broker calculations can be cumbersome and time-consuming, because market data can be voluminous. Calculating execution based on statistics like percent of volume (POV) can be prohibitive if repeated frequently, but careful design can resolve many of these problems.[9]

For more accurate and realistic simulations, including the ability to “drop” the portfolio management module into production, the broker module must be designed to behave like an actual broker communication. This enables orders to be routed to the actual interface by “flipping a switch,” without modification. Each broker and exchange may have a different method, and the backtester should route orders to different simulated destinations. Similarly, each broker and exchange may implement their various output and process queries differently. Aggregation modules, like Bloomberg’s EMSX, are used to unify the various methods. Otherwise, communicating to different parties may require different syntaxes and protocols. Simulation of these interfaces allows code to be tested and trusted when used in live trading.

Interactions with brokers or algorithmic execution should be simulated in the broker function, including querying information like the execution status of a trade. Trading desks and brokers can be compensated by formulas (such as price improvement in excess of VWAP) which can also be modeled and incorporated in the backtest simulation. Just as the many services of a broker are not necessary (but often required) for algorithmic execution, interactions with a trade order management system (TOMS or OMS) or an execution management system (EMS) may need to be simulated for “drop-in” code reuse.

Performance measurement and attribution, and transaction cost analysis (TCA). Proper forensic analysis is an important step, including decomposing and attributing results to specific investment decisions rather than blunt statistics. While simple Brinson attribution can be effective and easily communicated, complex machine learning classifications and predictions can also be employed.

In-depth TCA can be performed at this stage and used in the training stage for a model that is retrained periodically. In this case, it is important to exercise the usual precautions against bias and sensitivity, including a sensitivity and overfitting analysis of the TCA model.

Less intensive TCA that adjusts “on the fly” should be housed in the risk management module and used for decision-making in the portfolio management module. The TCA that is performed “in the loop” may be less expensive to calculate so it can be efficient during live execution. Multi-pass backtesting that reuses preprocessed values may allow many simulations while incurring diminishing costs to simulate complex calculations in a live system. The binning and VWAP techniques are two examples of this.

The forensic module can also be “lifted out” and used for a dashboard to monitor live performance. Many metrics and frameworks can be combined to form a mosaic; regulatory, formulaic, and behavioral measures can be incorporated.

Clients and boards often view risk and performance in different and contradictory ways. They may benchmark performance against peers in good times, and against the other reference portfolios when the fund underperforms an index—there always seems to be someone who is performing better than the portfolio. Rolling averages (like three-year and five-year returns) and more timely performance (“what have you done for me lately?”) are referenced at different times. Pain measures, like drawdown length and shortfall, may be critical to investor retention and continued employment (see Chapter 18).

The post-trade analysis should address many audiences, including investors, risk managers, regulators,[10] and clients. Various (sometimes contradictory) measures may be needed. For instance, high-frequency time-weighted performance may be available to the investment team, but the clients might receive less frequent performance reports based on modified Dietz accounting. Simulating what the client will see helps forecast the flow of funds and may be required to present hypothetical results.

A Word About Orders

Exchanges, brokers, and other actors all use their own vernacular to place orders. These exist for speed and precision, and often predate automated execution. These are legally sound and battle-tested ways to communicate investment decisions. They are easy to research and well documented, but we will provide a quick overview of the more common and essential commands. As we review some of these orders, it should be apparent that the backtesting technology required to simulate them can be complex but deterministic.

Important uses for these simple orders may grow into new strategies or strategy enhancements. For instance, an order that allows a broker to purchase one of several securities can permit best execution among a set of indifferent assets without micromanagement by the portfolio managers or traders. It can transform a hypersensitive and indecisive optimization into one that is capable of selecting the best assets at execution time.

General orders. The most general types of orders are buys, sells, and cancels. They often require clarification, including size, time, or level (limits, prices, or percentages). Orders can be aggressive (liquidity takers) or passive (liquidity providers), like those used by market makers. Passive orders may result in better prices or missed opportunities and are usually less expensive than aggressive orders. Different exchanges may offer various rebates to provide liquidity; some exchanges provides rebates for taking liquidity.

Time-in-force modifiers, like fill or kill (FOK), good till date (GTD), and immediate or cancel (IOC), are among the most common modifiers. Simulating these orders requires volume data to determine if the market could have supported the order size.

Contingent orders involve a condition, an "if this then that." Orders can be contingent on levels and percentages and may include a "plus" to allow for poor execution or slippage. A one-cancels-the-other (OCO) order is a contingent order. Basket versions of OCOs are also standard;[11] they may be called bids wanted in competition (BWICs) and offers wanted in competition (OWICs). BWICs and OWICs can net bid/ask spreads and eliminate the timing uncertainty in transacting a large number of trades, such as when rebalancing a portfolio. Other contingent orders include trailing stops and trailing limits.

Guaranteed orders. Just as insurance companies charge for convenience and serenity, brokers offer guaranteed orders in return for a larger commission. Orders can be guaranteed at a time (such as the open or close) or a particular level, like the TWAP. The use of these orders can make backtester predictions more accurate at a nominal cost for the convenience.

Best execution. When parameters are specified, they often result in opportunity cost. Depending on whether the broker is acting as principal or agent, parameters may delay execution until the broker can be assured that he can provide the required trade without undue risk to his business. As with nearly any convenience, orders with conditions can have hidden costs (including delays and worse prices) which could be avoided by managing orders actively by the trader. A good broker with an edge or an axe (or a bit of luck) can provide faster execution and price improvement—even while benefiting himself. Other structural benefits exist, like *pay for order flow*.

Order book (LOB) and depth of market (DOM). Markets are complex, inefficient, and disjointed. Orders aggregate from various sources (some of which may not be accessible in live trading) and stack by priority behind the Best Bid and Offer (BBO). This is why many backtesters approximate execution prices with the VWAP of historical trading data, assuming their order would displace actual trades that occurred on that day, and the displaced orders would disappear.

Access to the history of bids and offers available to the manager can allow a backtester to accurately determine the execution price if the order size is less than the market depth. It can also estimate the price if the order is larger than the book or if the execution algorithm is more complex than *sweeping the book* (sweep-to-fill order, “keep buying, I’ll tell you when to stop!”). Some of the most sophisticated backtesters simulate entire order books, but they require large historical data sets that include orders and trades.

Commissions and fees also vary by agent and may be based on complex formulas, including rebates for providing liquidity and transaction volume. Some traders, including high-frequency traders, can base an entire business model on arbitraging fee structures.[12]

Special Signals

Simulations can be awkward and complex. Illiquid and private assets, competing constraints, and taxes may be simple to explain but befuddling to model. A straightforward combination of several efficient frontiers involving exclusive choices (assets A and B, *or* C and D, *or* E and F) is often enough to produce an odd and unintuitive solution (e.g., a kinked or broken efficient frontier).

Complexities tend to compound simulations and make them intractable, which is why simulations are often simplified and tailored to specific considerations rather than running “[everything but] the kitchen sink.” Complex features may include multiperiod execution and transaction cost prediction. Shortcuts like turnover constraints can produce imprecise and detrimental results, limiting the opportunity set.

Other business models may demand specialized tools, like microstructure models and order book reconstruction and simulations. Derivatives, structures, and off-the-run bonds all require meticulous treatment.

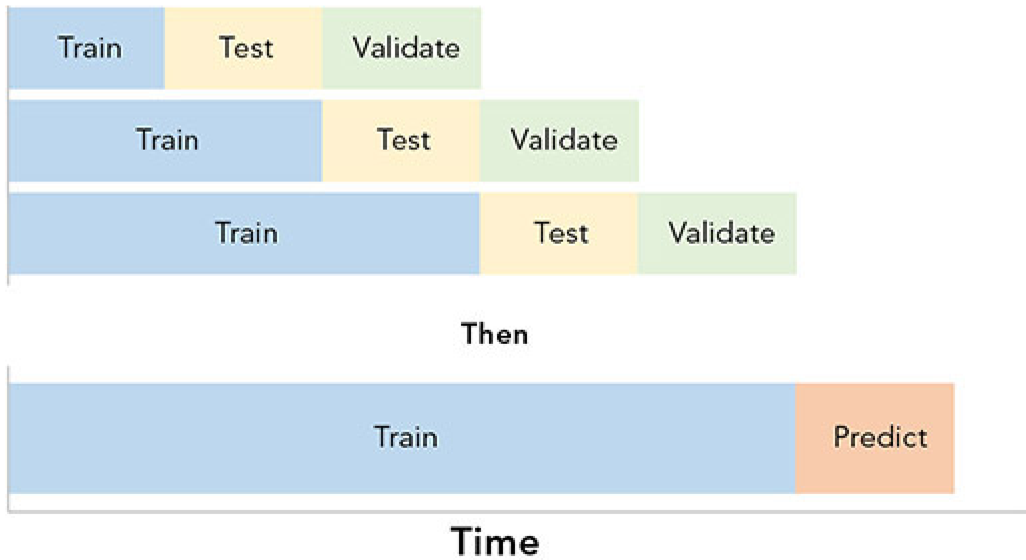
Tax-aware models that keep track of account location and tax lots can be a strategic benefit for both institutions (*hold-for-investment* versus *mark-to-market* accounting) and retail managers (householding). But tax effects can be challenging to model or benchmark, and include the limited benefit of tax loss carryforwards (versus the unlimited generation of taxes due to profits).

Shortcuts, like treating each lot like a different investment, can fall short. Complicating matters further, the high correlation between tax lots that refer to the same investment can degrade optimizations and other analyses. More specialized and complex modeling can be a strategic advantage.[13]

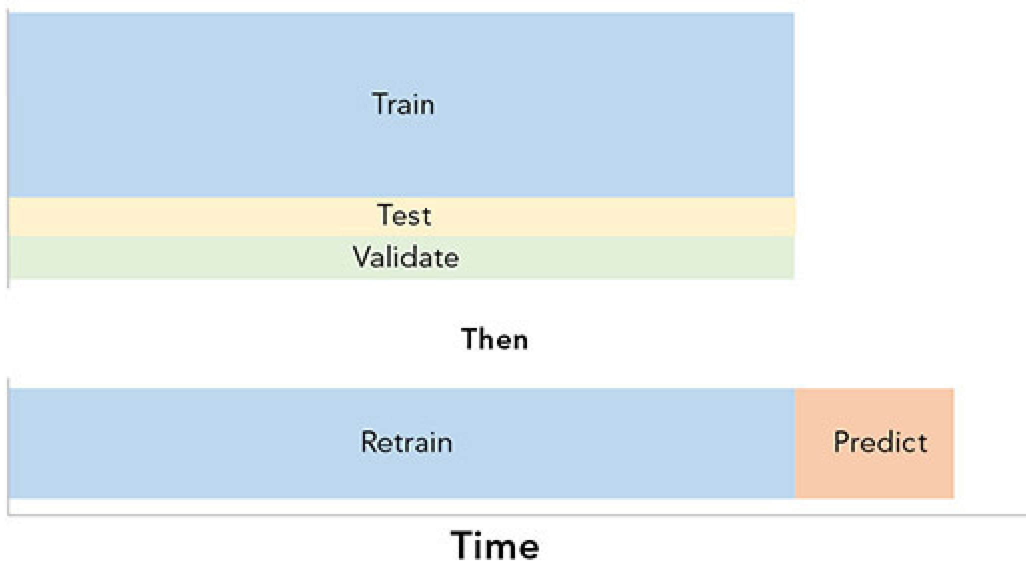
Generally, managing multiple accounts with a shared interest but different risk tolerances, return goals, liabilities, and tax treatments involves deliberate treatment. Investment portfolios commonly suffer from layered fees and nested investments, resulting in massive inefficiencies that could be unraveled and exploited through a thorough holdings-level analysis.

Data

Backtests may draw their data using multiple techniques and databases. Complex data sources are a necessary reality but may have shortcomings, such as look-ahead bias, information bleed, and data snooping. All these techniques require the separation of training, testing, and validation sets (Figure 14.2).



(A) Expanding window with "out of time" testing and validation



(B) "In time" training, testing, and validation

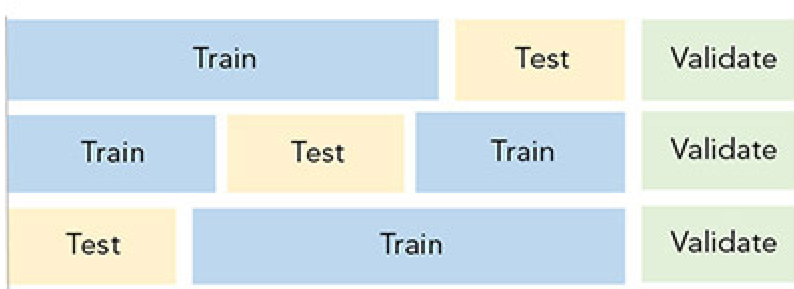


FIGURE 14-2 Validation modes of a backtester

When determining the retraining frequency, it is crucial to train and test a working model, including regular scheduled retraining, before it is deployed. Due to the psychology or the political realities of the workplace, there may be great temptation and pressure to retrain the model ad hoc once it is in production, especially during drawdowns. It is best to anticipate triggered retraining and rebalancing, and to include the these events in the testing methodology so they become part of the plan rather than an untested exception to it.

Historical data sampling is the most common process for backtesting. The *walk-forward* (cross-validation) and *expanding window* methods step through data as if replaying history. Walk-forward may do a better job of capturing regimes, while the expanding windows method may simulate evolutions and averages better. The history is often played from beginning to end and applies the backtest to as much data as possible, encompassing many regimes.

There are many variations of data sampling, including overlapping rolling periods, splitting the data into non-overlapping periods, and picking periods representing times of interest like scenario testing. Scenarios can also be used for meta-training to determine when the model is most effective or most sensitive. A common limitation with these methods is that the evaluations they produce may be salient, but they are also strictly anecdotal.

Historical simulations frequently fail due to unintentional bias in selecting parameters or time periods and the temptation to remove “anomalous” periods and outliers, which tends to inject look-ahead bias. Another common mistake is the tendency to use as much history as possible, which confounds results with the population statistics, like averages. Using scenarios, including stages of the business cycle, may produce more focused evaluations ("less is more"). Some models, such as state-space models like Gaussian mixture and hidden Markov, can deal with mixed regimes better than others.

Resampling, including cross-validation, draws data from history in *folds* that may or may not be sequential or randomly selected (bootstrapping). Repeated resampling improves historical sampling by increasing the amount of data available for backtesting and decreasing the potential for overfitting. However, it creates the additional problems of overlapping sets and the destruction of features like autocorrelation, clustering, and non-stationarity that may be present in the original data set.

Factor-based backtests take a more scientific approach by modeling investment performance based on driving forces like inflation and credit spreads. Since their inputs are theoretical, targeted, interpretable, explainable, relevant, practical, and more independent, the analyst can experiment with potential circumstances that "rhyme" with history or have never occurred. An example might be:

> What would the portfolio's reaction be to another credit crisis, given current conditions and bolstered protections instituted since the Great Financial Crisis of 2007–2008 (GFC)?

Factor-based models are tempting because they permit complex and precise scenarios to be specified directly and unambiguously. In addition to analyst-defined scenarios, this method is simple to use to answer challenges from critics and stakeholders by running backtests based on their assumptions. Factors can be combined to simulate events that have never occurred in the past but the theoretical nature of synthetic events requires caution, e.g., modeling the unobserved hypothetical interaction between factors can be uncertain.

It is difficult to accurately model the complex price responses of investments to their factors. Some more explainable versions of factor-based models rely on statistical models that use covariances for simple investments and cash flow models for nonlinear investments like derivatives and structured products. Complex models may include causality and machine learning.

Synthetic data generation attempts to solve the anecdotal simulation problem by producing artificial histories with "perfect foresight" rather than simulated factors. Using synthetic data can be akin to selecting a high-variance model rather than the high-bias factor models (although, synthetic data can be produced with a high-bias model as well). The high-variance nature of this method injects randomness and probability, easing the concern that a factor model may be too precisely and inaccurately defined. With complete control over synthetic data, this method can improve accuracy by increasing the number of draws while holding variance constant.

A simple type of synthetic price data might use Brownian motion, which preserves the correlation structure of assets using Cholesky decomposition. Cholesky decomposition addresses an important detail often overlooked when generating synthetic data: interactions.[14]

Another somewhat naive but popular method uses simple Monte Carlo draws from parametric or nonparametric distributions (like extreme value distributions). Analysts often resort to less realistic parametric distributions and copulas because of the difficulty in evaluating inverse cumulative distributions (CDFs). Goodness of fit is often misestimated when using imprecise and unreliable evaluations like quantile-quantile (q-q) plots.

More sophisticated methods may use generative adversarial networks (GANs) adapted to generate time-series data. Synthetic data can be produced with nonparametric methods.[15] In addition, if more control is needed, parametric methods can impose bias.

A nearly unlimited amount of data can be produced this way, solving the problem of underpopulated data sets or crucial minority data sets. Critical queries may be answered, like assessing the validity of a factor-based method by employing synthetic data sets that appear to mimic scenarios, albeit with a less direct specification. If necessary, appropriate data sets can be produced by generating many evolutions and selecting the ones that represent the scenario. Like factor-based methods, the value of generated data depends on its realism. Unrealistic synthetic data may confound some algorithms that are designed for time series learning.

Bias, Validation, and Hyperparameterization

Much of this book details biases in financial models. Adjusting statistics for the number of trials attempted and comparing results to the model's results based on arbitrary decisions (rather than model forecasts) can help manage biases. Common sense, domain knowledge, and practicality, like testing for no-arbitrage conditions, can be much more accurate and effective than theoretical tests skewed by inappropriate data (such as those that are multicollinear, autoregressive, or nonstationary).

A great deal has been written about the proper use of backtesting and comparing it to methods of scientific discovery. Increasing the number of trials properly also improves the chance of finding a spuriously profitable result. Many valuable suggestions require restraint. Research should be stingy with exploratory trials (resisting the natural temptation for trial-and-error research) so that backtesting is limited to testing and not the iterative improvement of a model. In many high-pressure, results-oriented investment environments, it may be too much to ask of a strategist to limit the breadth of his research for fear of stumbling on a lucky outcome.

It may be unreasonable to expect people to act against their interests, even under the threat of penalties, spurious results, and poor out-of-sample performance. They may have immense pressure to cut corners or even deceive to "buy time," especially if their incentives encourage bad behavior. There is no substitute for integrity, intellectual honesty, and good governance to help actors consider the longer-term effects of prudent research. The quant and the strategist are experts with unique insights, and they are responsible for explaining the reliability and veracity of their intricate, often arbitrary, and impenetrably confusing methods to management, salesmen, and clients.

Evaluating Strategies

Quantitative methods are often easy to categorize. They frequently rely on their precise implementation for an edge. Quants are naturally wary of giving away their "secret sauce." Probing questions that emphasize the shared nature of quant strategies and not the implementation details can help alleviate paranoia. The interviewer's knowledge and due-diligence checklist can justify the need to understand the fund's investment process in detail. It is essential that an investor or allocator thoroughly understand the strategy, including:

- How the strategies are born
- The sources of return and risk, scalability and capacity constraints, and external limitations
- Risk, sizing, holding periods, and leverage controls and discretion
- What the fund's "edge" is, including structural benefits or niche capabilities
- The effect of competition, including arbitrage
- The potential for data mining
- The quality of judgment in making the many intricate decisions involved in modeling, including data curation and cleaning

The robustness and reliability of the future success of a research process is often exposed by examining rigor and quality of the idea generation, data pipeline, feature generation, fitting of the model, testing, verification, and operational oversight, including the realism of execution models, operational robustness and scalability, and many details including monitoring hard and soft limits.

■ ■ ■

Backtesting is an essential step in validating the systematic investment process because it combines quantitative rigor with a degree of complexity unavailable upstream. Used properly, backtesting can help develop a more substantial, more reliable business—even though backtests have complexity limits and biases, too. Backtesting has earned a bad reputation because it relies on people with failings, blind spots, and temptations, so backtesting must be used with care and good intentions. Conclusions drawn from backtests can help promote proper discipline, but they must be adhered to and not disregarded without substantial evidence, even when faith in the system is challenged.

1. Ideally, we would use synthetic data to create vast data sets and precise scenarios, or interventions (*do calculus*). Even "plain vanilla" backtesting allows us to perform the essence of the scientific method: observation, research, hypothesis, experimentation, analysis, conclusion, and ultimately improvement and robustness.

2. You can find an EMSX simulator for MATLAB (we call it EMSC) at the book's website, www.QuantitativeAssetManagement.com. This simulator can be swapped out for the real execution object (e.g., EMSX) to seamlessly to turn your backtesting code into a live systematic trading model. "Bloomberg's Execution Management System, EMSX, is a sophisticated, broker-neutral trading tool that allows you to route seamlessly to nearly 2,000 broker-dealer destinations, including over 50 algorithmic suites." Source: Bloomberg, "Execution Management System," 2008.

3. Many "best in class" software packages are "most things to most people," but in making the software accessible to most people, developers often make it all but impossible to do some things. If you wish to incorporate ideas that do not fit their framework, such as those in this book, it may be easier to build an in-house system ("roll your own") that does what you want "organically" rather than improvise to try to make a package do what you need ("put a bag on the side"). Even well executed modifications to the software force you to work with an unsupported version of the software ("off model"), your bespoke version ("fork") will eventually become incompatible with future updates from the original manufacturer.

4. Although we have used many excellent financial tools, we have found errors in both their data and analytics. Any system, even one that is well designed, is built on assumptions that may not be appropriate to your "use case." Access to these assumptions and the ability to interrogate the code is essential to truly taking ownership of your process.

5. Several examples of the MATLAB backtesting framework, which uses both modes and was largely developed in my class at Columbia, is available on the book's website, www.QuantitativeAssetManagement.com.

6. Periodic specifications, like monthly rebalancing, rather than actual rebalance dates, may be wrong by a few days now and again due to holidays, weekends, employee vacations, and other technicalities. Buying or selling too early or too late may mean missing a colossal rally or stumbling into a rout, as well as producing inaccurate backtesting results.

7. The backtesting framework used in this book (and in my class) integrates many useful nuanced features including flexible and complex path dependent fund fee calculations (such as high water marks) and recordkeeping for borrowing (leverage) and lending (excess cash).

8. The order logic is handled by the execution simulator (EMSC in this case). The execution can be simulated in many ways. For instance, high-frequency data can be used to create a histogram of price and volume on each day (volume by price buckets). Synthetic orders for each bin can be placed in a matched book, managed by the simulator. If, on a given day, the market paid $10 for 100 shares, $12 for 50 shares, and $13 for 25 shares. Those three bids and offers can be entered in a synthetic order book. We may stipulate that we can transact 50 percent of the *average daily volume* (ADV). When the backtester tries to purchase 75 shares at the market on that day, it might receive 50 shares at $10 and 25 shares at $12 for a *volume-weighted average price* (VWAP) of $10⅔.

9. In the case of POV, pre-calculating the cumulative volume for each bar of each day allows the VWAP to be calculated quickly. A method involving binning was described in the prior footnote.

10. Many aspects of managing a strategy "take a back seat" to performance until they become the only important thing. Compliance and regulatory problems are in this category. Time management is a key skill for any project. Tending to these important but less urgent matters regularly can help make distressing times manageable. Avoiding them until they become urgent can be expensive and may derail the operation of the fund, or even cause it to fail.

11. Baskets can also be the opposite of OCOs; they can be all-or-none, or "take what you want."

12. Arbitrages that are illegal or difficult in traditional finance (TradFi) can be much easier in crypto (DeFi). Even DeFi arbitrages are subject to revocation and reversal under some circumstances. Like TradFi trades, DeFi trades are also subject to legal action and may be ruled illegal retroactively, even if the trade was not clearly known to be illegal at the time of execution.

13. For example, an expatriate account may benefit from timing investments (and their dividends) so taxable events occur when the beneficiary is on foreign soil.

14. If interactions are not considered, it is possible to generate synthetic data that appears good independently but not when used with other data. For example, if return distributions for two highly correlated stocks were generated independently, a Monte Carlo simulation might repeatedly draw an up day for one stock and a down day for another, and vice versa. By preserving correlations, the Cholesky method makes it more likely that the Monte Carlo draws of the two stocks will behave as expected when drawn together.

15. Look for examples of synthetic financial data generation using GANs at the book's website, www.QuantitativeAssetManagement.com.

15

Transaction Costs and Fees

Details That Matter

Transaction costs and fees are powerful determinants of profitability. They are complex and challenging to model. Ideally, they would be woven throughout the entire investment modeling process, from factor engineering through implementation, but the task's enormity and complexity prevents this.

This chapter and the next dive deeper into the intricacies of backtesting that we introduced in Chapter 14. We examine the agent paradigm more closely and discuss ways to model market behavior. We illustrate how to plan and gauge our reaction to these events to determine a suitable way forward. We also discuss contingency plans for when the market breaks the wrong way.

One of the most essential concepts for all but the most basic backtests is the *order book*. We must interface with or simulate a *transaction order management system* (TOMS or OMS) or *execution management system* (EMS).[1] Additionally, market structure is a vast and distinct topic; we will lay a foundation to enable our discussion of transaction costs.

Transaction costs and fees affect our simulation at many levels.

> **The custodian, bank, and back-office (accounting)** functions are at the core of a backtester. They must keep track of fees, which may be path-dependent, including *hurdles, high water marks, lock-ups, gates*, and other intricacies. They should also segregate and track various costs and performance to enable forensic analysis of the simulation.

The fund module must react to the simulated execution performed by the broker module, including timing, selection, periodization, buying, selling, canceling, leverage, and sizing. Large trades can alter the market (market impact) and may be reflected in adjustments to the raw data feed. A sophisticated investment algorithm will attempt to predict the transaction costs of potential trades and learn from previous iterations. *Prediction* in this module must only use *past data* to avoid look-ahead bias.[2] Cost simulations require a *market model*. A strategy that is not naive will have an *execution model* to formulate a response to costs for effective execution, considering complexities like rebalancing strategy, transition management, and withdrawal strategy.

The risk management function is normally housed in the portfolio management module but may be isolated and called simultaneously. Since shortfall and market impact affect risk, this module is also concerned with costs. The effects of these costs include liquidity, the ability to short, securities lending fees, and market impact. A thorough backtest should anticipate extreme risk management, including the flow of funds, “going to cash,” lock-ups, and gates. *Execution strategies* respond to costs but do not change the intention of a trade. They may change the method of execution and substitute similar instruments if costs induce one to become more attractive than another (e.g., tax-loss harvesting), but only if the investment thesis is unaffected. The *broker module* simulates incurred costs and fees with a different model than the portfolio manager and risk manager modules. For simplicity, the difference may be derived from historical values obtained by analyzing *future values* in the order book or merely trades, bids, and offers. *Performance measurement and attribution* (e.g., Brinson attribution) and *transaction cost analysis* (TCA) can decompose costs and fees after the simulation has executed a complete run. The risk management function simulates this process and the portfolio manager module’s reaction to it in the loop.

Orders and the Order Book

Simple backtesters are hampered by limited market data formats.

- **High, low, open, close, volume (HLOC) bars** (periodic summary statistics) are the easiest data to find but can be misleading as highs and lows may come from obscure and inaccessible exchanges and may report useless outlier values.
- **Best bid and offer (BBO)** data can augment trade data but lacks the depth of the order book and information from dark pools. It may be composed of quotes from venues inaccessible by the trading desk.
- **Time bars, ticks, and trades** are better, especially if they include bids and offers; however, they also lack depth and have a discrete frequency.
- **Raw ticks** may provide enough information to form better bars, such as *volume bars*, *money bars*, or bars based on complex domains like *imbalance bars*.
- **High-frequency data** may contain all the trades, bids, and offers and even order book information. This "big data," and to a lesser extent data with trades alone, may be so cumbersome that they require special tools, like MATLAB's datastore function or a tuned database like KDB. With a little practice, the use of these tools becomes second nature. It often takes time to extract a sample set from the larger data set. Saving intermediate tables can avoid needing to access the full data set unless the query changes (e.g., if a new predictor variable is needed or the researcher needs more recent data).

Order level data is essential for all but the most cursory backtest. Advantages include the *simulated execution* of complex orders using an OMS or EMS, e.g., using the EMSC emulator.[3] Proper order design and execution can add great value or clarify the consequences of important details. For instance, large orders must be timed properly to minimize market impact.

Order book reconstruction and *trade reconstitution* have been valuable tools since the beginning of markets. Significant trades are commonly obfuscated by methods like splitting the legs and distributing across time, brokers, and exchanges.[4] Watching the behavior of open-outcry participants and "reading the tape" was common before algorithmic trading.

Triangulation of tangentially related markets and products, such as implying underlier values from their derivatives, can reveal valuable insights. The complexity and variety of derivatives offer a treasure trove of implied data that is usually overlooked as “opportunities in work clothes,” as Henry J. Kaiser would say.[5] Derivative data can be exploited in order to gain insights into the underlier’s behavior and used be categorized as alternative data, but is commonplace now. Understanding the order book history is a valuable step in identifying opportunities and simulating backtests.

Algorithmic trading has revolutionized this adversarial contest with tactics like *spoofing*, *hidden orders*, *dark pools*, *liquidity arbitrage*, and *iceberging*. Powerful weapons in this arms race include *payment for order flow* (PFOF) and applying modern techniques and computers to dissect massive databases to reconstitute intentions and actions. Preprocessing statistics like *arrival price*, trade-weighted average price (TWAP), volume-weighted average price (VWAP), and *slippage* for orders can enhance market models significantly by speeding up repeated simulations.

Order book simulation facilitates more accurate in-sample and out-of-sample backtest results and model training. This is critical in high-frequency models and can add great value to lower frequency simulations as well.

Transaction Costs and Fees

The market is not indifferent; it conspires to siphon profits and bleed businesses dry. Many strategies are systematically designed to benefit from these fees and costs, generating an edge from flow, including those managed by banks, brokerages, and proprietary trading businesses. Businesses whose focus is the forecasting of income and appreciation provide the flow and pay for the privilege. For their part, professional investors benefit from salaries, bonuses, management fees, and incentive fees.

The stage of the investment process (contracting services, strategy decision, trading decision, and execution) is a natural way to classify fees and costs, in the same way that we used the agent-based paradigm to describe backtesting.

Fees and commissions are charged for services such as communicating an order and safeguarding an investment holding (e.g., custody services). They represent transfers of capital absorbed by the infrastructure (including brokers, custodians, government, and exchanges) before reaching the trading counterparty (buyer or seller).

Investment costs are incurred after the strategy decision is made and before the order is transmitted to the broker. These costs are a result of a suboptimal translation of an investment decision into an execution decision. A thorough backtester would simulate these frictions as part of the *broker* module. It is critical to remove friction simulations before the code is used in actual trading. Placing costs and fees in the broker module helps to ensure that they will be removed from the code, since the broker module is not used in live trading. These extra costs may crop up as a result of one of several factors, including:

- **Ignorance.** Trading decisions made by the investment team with a lack of situational awareness and poor communication with "boots on the ground."
- **Inefficiency.** Operational inefficiencies in transmitting the strategy decision to the trading team.
- **Constraints.** Constraints on the trading team can result in suboptimal orders. Some funds famously split strategy orders into parts. These funds transmit orders to the trading team in random fractions and random times during the day, rather than trusting the trading team to add value through timing.
- **Delays.** Ad hoc or disorganized processes can take more time and produce more errors than the systematized, planned, and preprocessed methods recommended in this chapter.
- **Errors.** Operational risk is ever-present and punishes in many ways (including lost time and opportunity, financial losses, regulatory trouble, outflows, and reputational damage). Leadership, governance, systemization, and automation can help to remedy the complexity, limit the damage, and lighten burden of operational risk.[6]

Trading costs, like market impact, result from trade execution itself after the order is placed with the broker. Trading costs are forecast by the *market model* using past data (*lookback*) and simulated by the *broker model* using future data (*lookahead*).

Models are needed to estimate costs and fees. These models are essential to backtesting but should also ideally be integrated with the alpha and risk models, and the allocation and selection decisions, including optimization. Implementing transaction cost simulations in the objective function of an optimization may be more feasible, albeit less realistic or precise[7] than expressing the transaction costs elsewhere. Elegant implementation in the objective can facilitate efficient calculation and develop a powerful economic intuition.

The *market model* is designed to estimate the fees and costs of trading. It can be used to calculate potential *net* profit and identify attractive substitutions. In backtesting, the market model is used by the *fund module* but can guide investment (strategy or alpha) and risk decisions throughout the investment process. It involves transaction cost analysis (TCA) and prediction "in the loop" using historical information. This model can be used in live trading. The *execution model* is used to optimize the execution costs of investment and risk decisions. It determines some substitutions and order details, including type, timing, sizing, and routing. This model can be used in live trading to make predictions and adjust orders.

The *broker model* simulates execution, including slippage. It is a trading simulator. It is not limited to looking backward from the observation date to examine historical prices and can use future information (examining data from dates after the observation date in historical data set) to provide better-simulated values. The *attribution model* analyzes risk and performance and attributes the causes to specific decisions and factors. This model can be used to analyze live trading.[8]

Visibility is another characteristic of fees and costs. Some fees and costs are listed while others may be "wrapped" or hidden and appear as frictions or slippage. Some may even be *latent*.

Observable, direct, and explicit fees and costs, such as taxes and commissions, are apparent and often scheduled. *Hidden, indirect, and implicit* fees and costs are less obvious because they are not explicitly charged, but easily identifiable when compared to an unencumbered transaction. Because they must be “backed out” or estimated, their magnitude (and sometimes their direction) may be debatable[9] but clearly present, impactful, and sometimes overwhelming.

Opportunity costs result from failing to *complete* an order optimally, including partial or missed orders and orders executed at worse prices. The magnitude and attribution of opportunity costs are often not easily quantifiable. Though they are authentic costs, quantifying their magnitude (and even possibly their direction) depends on assumptions about events that did not occur; for instance, alpha decay assumes that the order could have been executed with a theoretical degree of *slippage*.

Hard currency fees and costs are paid directly with currency transfers. These fees are unmistakable. *Soft currency and wrap* fees and costs are common vehicles for buffering and obfuscating the full extent of expenses. Examples include *bundling* services (such as providing data and analytics “for free”[10]) and *proprietary architecture (*a fund that holds another fund, compounding the two fees).

The predictability of the magnitude of the fees and costs can reduce a stochastic model to a deterministic one. *Scheduled (fixed)* fees and costs are often prenegotiated. They may be complex, time- and path-dependent, and client-specific (such as taxes) but are known with reasonable precision and certainty.[11] *Uncertain (variable)* expenses respond to parameters that must be estimated with less confidence.

The effects of costs and fees can be felt in many ways. The *efficient frontier* is a theoretical upper bound. The *accessible frontier* is somewhat lower (less profitable) and to the right (more risk). Specific decisions can alter the placement of accessible investments; for instance, guaranteed prices, bids wanted in competition (BWIC), and offers wanted in competition (OWIC) may sacrifice profit to lower risk.

Information ratio and *tracking error* are similarly affected. The information ratio is affected by the relationship between expected and realized execution and is dependent on forecasting skill (including timing), as well as execution competence (a “loser’s game”). *Limited arbitrage opportunities*, including *hard-to-borrow* shorts, allow inefficiencies to persist and make diversification benefits elusive. Similar to the long-only constraint, stocks that are difficult and expensive to short restrict arbitrageurs from profiting while reducing the ability to diversify portfolios.

Fees (Pay for Service)

The topic of fees, and payments in general, is confusing to many technically focused people; problem solvers and process builders who lack professional experience “in the trenches” should validate their models with savvy traders who may not be as technically competent but who have proven themselves “antifragile.”[12] Technologists should be cautious not to focus on the theoretical (or try to be too “elegant”) to the detriment of the mundane but critical elements of the analysis.[13]

In direct contrast, an entire ecosystem of sharps and systems has been designed to extract fees from the investment decision. Fees can be explicit or implicit, hard or soft, fixed or variable. Some of the many fees include commissions, custodial fees, transfer fees, clearing, settlement, taxes, levies, management fees, and incentive fees. Moving beyond the exploration of the risk-modifying effect of incentive fees, we will focus on the deterministic calculation of fees as it affects performance for backtesting.

Explicit, hard-currency, fixed commissions, fees, and taxes comprise the bulk of the expenses. Modeling them can be a complex task. Execution fees (e.g., by exchanges) involve many charges. These are usually expressed as a single fee, but may require the evaluation of numerous components. Brokerage commissions are typically charged by share or as a fraction of currency amount and may depend on a schedule of breakpoints that decrease commissions with volume and service type, such as voice brokerage or *direct market access* (DMA). Rebates may be absorbed by a broker or passed on to the investor. They can be based on:

- **Maker-taker** schemes rebate market makers (passive traders) as compensation for providing liquidity and bill market takers (aggressors) for using liquidity.
- **Taker-maker** schemes are the opposite of maker-taker schemes, and reward liquidity takers while penalizing market makers.

- **Commission** models may charge both market makers and market takers.

Implicit and soft-currency fees are vaguer, requiring more intimate domain knowledge, but are usually modeled with simple assumptions and estimates. Many misleading indicators, such as the *total expense ratio* (TER), exclude expenses like performance fees and costs like transaction costs.

It is important to recall that financial structures—and fee structures in particular—are often designed without any consideration for the mathematics required to value and model them. Like constraints, they frequently result from arbitrary conversations between lawyers, clients, and salespeople, and they typically take the form of legal contracts. The participants in these negotiations do not consider—nor care about—how seemingly simple ideas can be challenging to translate into formulas.

To understand the impact of these structures, it is often helpful to think as agents would, paying particular attention to the terms governing cash flow (precisely who pays whom, what, when, and how much). Minuscule details, such as *day-count bases*, can have overwhelming repercussions. Oversimplifications are inappropriate for many investment tasks.

Variable fees can be path-dependent. They commonly include hurdles and high-water marks, and extend over multiple periods. They may include explicit or “real options” (including clawbacks) and may be paid in cash or instruments of uncertain value (such as unvested stock).

Forethought and sound design can overcome most difficulties in modeling fees. Event-driven, agent-based backtests can account for nearly all contemporaneous charges. Unlike with costs, the process of modeling fees is complex but rarely difficult.

Investment Costs (Before the Order)

What are the primary causes of investment costs incurred internally, after the investment decision is made, but before the order is released outside the firm? There are several.

Investment costs are mostly caused by unavoidable delays and frictions, and poorly designed or implemented internal processes. Strict attention and diligence will minimize investment costs, but they can be pervasive and deceptive. They include unavoidable inefficiencies; poor communication and integration between the investment committee, investment team, and traders; and cumbersome technology and its associated operating risk. The effect of these costs may be poorly formed or inaccurate objectives, execution errors, and excessive overhead in duplicated and ineffective procedures.

Delays and specification issues. Even highly effective and well-run organizations suffer from delays in transmitting investment ideas to their final security selection process and order generation (sometimes more so because of the company's size and complexity). Regardless of whether the firm uses a human trading desk or a predictive market model, market timing decisions and specifications must be made, including:

- **Selection** of the best "expression"[14] of security selection for the prevailing market conditions, fees, and costs. An investment with superior earning potential may be superseded by another due to market conditions, such as the all-in purchase price including fees and costs.[15]
- **Timing** by the investment team and trading desk may involve deliberation over best execution. However, this does not refer to intentional timing decisions that are based on the prediction of long-term market fluctuations. Trade timing may include the consideration of short-term supply and demand imbalances and minimizing transaction costs (including impact) balanced with alpha decay.
- **Periodicity** decisions are similar to timing decisions, but extend beyond a single investment or execution cycle and often involve trading gaps, like overnight periods, compromising risk controls, and complicating estimates.
- **Order types** other than market orders may cause delays in execution or worse execution prices, but also may improve control over the timing and price of the trade.
- **Routing decisions** can reduce costs and fees and improve execution but, if they are subjective or poorly informed, they can inject uncertainty, delays, and errors into the investment process.

Trading Costs (Execution)

Trading costs occur after the order is released outside the investment firm and do not include fees. Trading costs are the most conjectural and challenging part of the four models:

- **Predicting** transaction costs is the primary function of the *market model*.
- **Optimizing** costs and fees is the purpose of the *execution model*.
- **Simulating** costs and fees takes place in the *broker model*.
- **Analyzing** costs and fees occurs in the *attribution model*.

Of these models, the market and execution models can be lifted out and used for live trading. The attribution model can also be used "in production." The most difficult part of all four models involves predicting trading costs.

Box 16-1 Standard Terms

Standard terms and concepts that describe a stylized order execution process include:

- **The pre-trade** equilibrium region, which extends up to the decision time
- **Decision price,** which is the price of the asset when the decision is made by the investment team, which may be well before the order is placed
- **Arrival price,** which is the price when the order is available for execution
- **Execution price,** which is the price that the order realizes and may be a weighted average of several child orders spread over time, venues, and counter-parties
- **Implementation shortfall and slippage,** which represents the difference between the decision price and the execution price[77]
- **Temporary impact,** which is the price change that is incurred from the arrival time until the new steady state is reached and may

involve many cascading orders and effects that can attenuate and expand

- **Permanent impact,** which is measured by the price at the theoretical steady state equilibrium after the market settles down from the perturbations produced by the order

In addition to the various causes of slippage, the magnitude and direction of slippage can be affected by the type of order and the strategy that generates the order. For example, trend-following strategies that encounter execution delays are still more likely to enjoy price improvement because of the positive momentum of the trend, while mean-reverting strategies that encounter delays are more likely to suffer adverse price movement for the same reasons.

Idiosyncrasy and estimation error. For simplicity and convenience, market impact is usually discussed for categories of assets rather than specific investments. Investments are idiosyncratic, not just requiring different parameters but oftentimes entirely different market models; their estimation error due to noise and residual effects frequently outweighs their specificity.

Not only do market impact processes vary by asset class, sector, and size, they also vary by investment strategy and factors. In addition to their susceptibility to market forces, like momentum, investment strategies may require more precision and turnover (such as with arbitrage strategies). Their factors (like dividend income) may be more or less sensitive to the effects of market impact. Individual algorithms and traders can sometimes be identified by their “footprint” or “signature,” which may allow the market to anticipate their actions.

Transaction stages span several categories. It is difficult to estimate how to attribute market changes to these stages due to their ambiguous borders and definitions. *Pre-trade* conditions exist when the order is received, including the level (arrival price), direction, momentum, and imbalance of the investment, the bid-offered spread, and the relative price of alternative investment choices. It includes the order book and its momentum. Gauging the nature of the arrival state can be complex.

Temporary impact is the short-term effect of placing or executing an order in excess of the unperturbed market action and the long-term effect of the order. Crowding and market segmentation can make temporary forces asymmetric; for instance, crowded buy orders in the general direction of market momentum are usually more disruptive than sells, which provide the liquidity for that trend. Repeated but often decreasing episodes of temporary impact may occur before the *steady state* is reached, possibly due to multiple trades or delays in information dissemination.

Permanent impact is the lasting effect of an order in excess of what the market action would have been in the absence of the order. The two uncertain temporal boundaries complicate permanent impact; permanent impact begins where temporary impact ends and can transition slowly or attenuate asymptotically. It can also be truncated violently by market shocks like economic releases, making estimating a theoretically undisturbed permanent impact more difficult.

Pre-trade conditions include:

- **Uncertainty in the level** or *resting price* of the investment.
- **The bid-ask spread,** most commonly associated with the best bid and offer (BBO) of a consolidated price feed. This more accurately includes the entire order book.[17]
- **Imbalance** is related to the bid-offered spread and the depth and distribution of the order book, akin to the potential energy of the market, where drift and volatility are like kinetic energy.[18]
- **Liquidity** is affected by the depth of the order book and the associated imbalance.
- **Special event timing** can alter the pre-trade state in the absence of the order; it includes economic announcements, scheduled political events, and significant derivatives expirations.
- **The drift** (momentum, trend) may move the market in the absence of the order.
- **Volatility,** like drift, affects the future deviation of the arrival price in the absence of the order.[19]
- **Relationships** between investments may include spreads, correlations, and other statistical and deterministic co-movements.

Temporary market impact refers to the effect of an order. The forces that drive market impact are naturally loss-inducing, although there are exceptions like those involving trend-favoring momentum trades. Knowing that a buyer is interested in a product may tempt a seller to raise his price, or knowing that a seller is interested in unloading a product may compel a buyer to lower his bid. In addition, as liquidity is depleted, extreme prices may become the only ones available.

Market impact is determined primarily by liquidity and signaling but has many components: *Price shocks* are instantaneous changes (impulse functions) in price, often due to the consumption of liquidity. If a trade *sweeps* the order book and consumes all the liquidity on that side, or scares away the resting orders, the best bid or offer will jump to another level. *Dislocations* occur when some asset prices move in a way that violates existing relationships with other instruments. Term structures and *volatility surfaces* (or cubes) are common structures used for identifying dislocations.

The order may set the market in motion along a trajectory, or *drift*. Unlike momentum, drift connotates a less mercurial trend. *Seasonality* includes daily effects (such as the various exchange opens, closes, and lunchtimes in major trading cities), day of week, beginning and end of month rebalancing, and particular months of the year. *Volatility* can result from the attenuating reverberations of instantaneous shocks or the expanding effect of information dissemination. Volatility has been extensively studied, and its properties are well documented.

Liquidity can be a dominating feature. Many relationships, including the asset type traded and the market's state, condition the effect of liquidity.

Signaling, information theory, and game theory are intertwined in market analysis. Execution, including the attribution of market impact and the "chess game" of minimizing it in the presence of adaptive adversaries, involves the principles of signals intelligence (SIGINT).[20] *Interaction* between impact factors is complex. For instance, higher volatility can exacerbate low liquidity (by increasing volume with fear) and permitting temporary price fluctuations to trigger extreme bids and offers.

Opportunity cost results from the suboptimal execution of an order and can happen for many reasons, including sacrificing timing certainty (market order) in pursuit of price certainty (limit order) as the market "runs away" from the decision price. *Operational risk* is expected in high-pressure, fast-moving environments. “Fat fingers,”[21] ticker specification errors, handle errors, and many other otherwise avoidable errors are both embarrassing and commonplace.

Permanent market impact affects the market in a persistent way as information is absorbed and distributed:

- **A new pre-trade state** is established when a permanent impact creates a new price level that is stable until another shock perturbs it.
- **Regime shifts** result when the pre-trade state is altered so that the dislocations and other characteristics are persistent and form new lasting relationships.
- **Interactions** can affect permanent impact as well.
- **Opportunity cost and operational risk** can be temporary and “traded out of,” or permanent and lasting.

Trade-offs

An investor’s desire to minimize costs and fees require complex choices with unclear consequences. Compounding these difficulties are uncertainty, pressure, and urgency, which are all made more difficult by active and capable adversaries (competitors) that mislead and react to the investor’s efforts. They can also be compounded by passive adversaries (a financial ecosystem designed to generate and sometimes maximize costs and fees).

There is no *free lunch*.[22] There are some pure arbitrages, but they are infrequent and fleeting. Opportunities without barriers to entry and exit are sought-after and require risks, including the operational risk of rushing to complete the trade before an person or algorithm “snaps it up.” Structural advantages, like access to successful hedge funds and private equity, are frequently overvalued. Access is necessary, but much skill and effort are still required to select lucrative private investments. Some of these many trade-offs include:

Information content. Nearly everything has information content; every action signals adversaries, and every attempt at understanding information destroys some of that information (bias versus variance).

Observable versus hidden. Mitigation often shifts obvious and direct fees, costs, and risks to indirect, obscured, and bundled manifestations. “Costless” trades are not free; they just do not require up-front fees. Their opaqueness often increases their cost.

Opportunity cost. This can have a significant benefit of mitigating observable expenses. We discussed how using a limit order may reduce price uncertainty but increase the chance that the order would not be completed—not just because the market may not reach the specified level, but also because the broker may be “leaning” on the order, which may signal adversaries.

Scheduled (fixed) versus uncertain (variable). *Risk shifting*, both fixed and variable, is the foundation of insurance companies, market makers, and bookies. Minimizing uncertainty (other than through diversification) can be costly. The rational reason to buy insurance is either diversification or an inability to tolerate the potential loss. Preventing the risk of ruin has value beyond the expected loss.

Convenience versus opportunity. Certainty is comforting and enhances risk-adjusted rewards by reducing risk. But tolerance of discomfort and pain, relentlessness, vigilance, and formidability frequently are part of the cost of opportunity. A large portion of trading, investing, and life in general is “blocking and tackling” or “sweat equity”—doing the simple but challenging things that are inconvenient, unpleasant, and tedious. Many products have been invented to allow market participants to pay to reduce their risk and cognitive load. However, automation and process represent a better, more sustainable solution than paying for products.

Intentional versus unintentional risks. These are the essence of successful arbitrage. Arbitrages are rarely pure, involving many risks and factors. They may be cross-sectional, longitudinal, or time-varying. They may involve different strikes and high-order moments. They may involve different geographies, like buying gold on one exchange and selling it on another. Invariably, they require some timing and operation risk. An arbitrageur should isolate a finite set of risks and factors that she has an opinion or view on and distinguish those from the risks and factors for which she does not have an opinion. The former is the arbitrage set based on an informed view and is the *raison d'*être of the trade. The latter comprises the unreasonable and unwanted risks that should be balanced against the cost of eliminating them.

Let's take a look at the Fundamental Law of Active Management and the information criterion. Richard Grinold and Ronald Kahn taught that the benefit of skill, known as the *information ratio* (IR), is proportional to the skill itself, known as the *information coefficient* (IC), and the square root of the number of *independent* opportunities (*breadth*). The *information criterion* measures the ability to distinguish between predictive factors and models and spurious ones. With these basic concepts, we can describe and forecast the effect of timing on skill.

Alpha decay versus transaction costs. In this instance, the term *alpha* is misused to describe the information coefficient. Alpha decay is the dissipation of IR over time and is analogous to *theta*.[23] Ideally, alpha decay measures the natural decline in predictive ability rather than unpredictable idiosyncratic vicissitudes. *Signal half-life* refers to the time it takes for a predictive signal to lose half its predictive ability.

Information horizon. The IC after an *implementation lag* and the correlation between IC and the effective IC are valuable concepts for transaction cost analysis and cost forecasting.

Skill is largely prediction, although it can involve much more; as a result, part of IC is the ability to predict. Signal half-life helps determine the trade-off between patience in the pursuit of best execution and loss of value in the execution goal. This balance is similar to the conflict between variance and bias. High-bias strategies, like activism and deep value, have great persistence and long half-lives, while high-frequency arbitrage is fleeting.

Alpha decay versus the alpha model. It is difficult but ideal to adjust features to reflect their effective IC holistically and to incorporate decay in the meta-selection process and optimization. From a practical perspective, the uncertainty, complexity, and path-dependency of determining alpha decay tends to relegate this task to the backtesting stage. "In-the-loop" and post mortem attribution can provide empirical feedback, at least crudely in the form of an efficiency factor, to adjust ICs downward and prevent recidivism in overestimating predictive power. Alpha decay is relevant to the rebalancing decision not just because of transaction costs for each trade, but also for netting costs cross-sectionally and over time. The appropriate timing of rebalancing, addition, and withdrawal events can also benefit from market momentum and other time-sensitive fees and costs, like taxes.

Measuring Fees and Costs

Measuring fees and costs is essential for developing and calibrating high bias models and creating adaptable high variance models. Metrics are valuable for attributing investment acumen and rewarding participants. Traders are frequently rewarded using a formula based on their ability to fill orders at average price levels that exceed benchmarks.

Fees, including hurdles, high-water marks, lock-ups, and gates, are easy to determine but can be complicated to model. Shocking fees that result from ignorance or "edge effects" occur in the same way exogenous shocks affect markets. The difference is that those windfall fees can be modeled with some certainty even if the likelihood of triggers being activated cannot be anticipated as easily. But this is not always true; trigger definitions can have unintended and unanticipated consequences. Triggers and reference rates are legal concepts that are sometimes debatable or subject to unexpected forces and events.

Availability. Measuring costs is far more difficult than measuring fees. Costs and their components are rarely available in data sets. When these details are available, the data is typically unmanageable due to its complexity (e.g., including many tax lots for each client). Rough estimates of liquidation cost or estimates of market direction based on BBO momentum are poor substitutes for proprietary data and order book history.

Cost component breakdowns are uncommon. The order's decision and purpose (factors, thesis, or even timing and decision price) may be unavailable for externally managed investments. Even when the data is internal, these details may not be communicated to mid- and back-office analysts. Basic trade characteristics like size and liquidity may be challenging to determine if market data feeds are used without order book information. Segregating and attributing costs is an integral part of attribution, which we discuss in Part IV.

Noise makes measuring costs (and statistical transaction cost modeling) extremely uncertain, especially over long horizons and multiple periods that invite pollution by exogenous forces (like other trades by other parties). Imposing high-bias models, even linear models, is common because estimation error eclipses inaccurate bias.

Uses of measures. Implicit and unobservable effects can be estimated with persuasiveness but involve significant assumptions, interpretation, and nuance. For instance, an analyst could estimate that a lock-up delayed the repatriation of his funds for some time, leading to an appreciable decline in the investment during the delay.[24] It may be useful to consider the consumer of these measures and their willingness to accept assumptions when tempers run high during difficult decisions and compensation negotiations.

A complex and subtle disaggregation of transaction cost components can be illuminating, but simple, observable measures are important for purposes that require clarity and certainty. Imperfect measures can be gamed and may invite the agency problem of misalignment of interests.

Three broad categories of explicit cost estimation include *benchmarks* that are observable and unambiguous reference prices or rates. There is benefit to simple measures with a minimum of parameters, and many are used in practice.[25]

The drawback of indiscriminate measures is that they are inappropriate for specific situations. For instance, if the trade dominates its interval volume, it can skew performance measures. *Time shifting* is one method of alleviating this skew. Simple measures should be monitored for abuse, because they can be hedged and leaned on to increase performance statistics and compensation.

Relevance and investability are critical characteristics of reference prices. Markets can be segmented, and the costs of implementing benchmarks can be complex with the risk of estimation errors. Usually, these nuances are ignored, and benchmarks are impossible to replicate without risk. Conversely, execution often involves “seat value,”[26] like information flow and internal crossing networks that can provide low-risk or risk-free price improvement. Sometimes, traders are compensated beyond performance with salaries, fringe benefits, and severance packages. Tailoring benchmarks for specifics—like the permitted execution interval from the transmission of the investment decision to the trading desk to the maximum permitted transaction horizon—can make them fairer and more precise, but risks bias and less universality.

Shortfall methods require a benchmark but can account for arbitrarily complex fees and costs, including unobservable expenses like opportunity costs. The benefits of shortfall measures include increased precision and alignment, with the potential for granular attribution. A carefully designed shortfall incentive can align execution interests with investment interests by properly adjusting for execution effects on the profitability of the portfolio and the health and ongoing viability of the firm. Shortfall can be idiosyncratic and vary with different execution partners (such as brokers).

Models attempt to surpass benchmarks in the aspiration for increased complexity and realism. They are uncertain and difficult to design well but can be a strategic advantage in many stages of the investment process, including feature selection and forecasting, alpha and risk modeling, portfolio construction, execution, and attribution.

The Market Model (Prospective)

Models need goals. As with all the models we discuss in this text, specifying the intended use of the model's output will produce more pragmatic and suitable design. Market models for "alpha" forecasting are complex and can reflect many goals, including risk, impact, and decay. Risk forecasting overlaps considerably with return forecasting but emphasizes different aspects. Incentive forecasting is more specific and requires an aptitude for balancing sensitivity, consistency, and interpretability.

Many factors affect market impact. Generally, they are thought of in several categories. The *resting state* of the market includes investment type and market cap, volatility, average daily volatility (ADV) and volume "at time" (delta AVAT), volume distribution (such as front- or back-loaded) spread, order book depth and imbalances, drift and trend, and proximity to technical levels and imminent events. For example, the time of day and execution regime affects volume on a typical day. We discussed this briefly in Chapter 7 where we showed both trade frequency and volume are greater at the open and close of the day, and that futures contracts exhibit seasonality near their expiration. Most trades are executed at the close, but this has not always been true and is likely a result of algorithmic trading. Historically, the manual process of market makers opening exchanges every morning with full knowledge of the order book produced high volumes at the beginning of days, as well as at the close.

Many such regimes changes have occurred. For example, the decimalization of the US market in the 1990s altered the market structure.[27] These regimes emphasize the importance of training models based on relevant data and factors, rather than falling into the common trap of using as much market data as possible. More relevant data sets can employ simulated forward-looking data rather than outdated historical records that span regimes and confuse the model.

Order features include liquidity (such as order size as a fraction of market cap, volume, and order book), execution horizon, contingencies (like substitution), direction, and strategy. For example, small-cap stocks are more sensitive to liquidity. There are trade-offs with liquidity as well. Illiquid assets may incur higher transaction costs but reward investors by paying a liquidity premium over time. While it is easy to conceptualize the features and responses of simple-to-understand products like common shares, this exercise becomes complex when dealing with products like derivatives and structured products with nonlinear and non-smooth responses.

Execution strategy can include speed and aggressiveness, schedule, and participation rate. Execution quality and environment affect information diffusion and market reaction time. Algorithmic execution designed to mask intention is different from a retail brokerage selling order flow. Dark pools are different from open-outcry pits.

Interactions are relevant for market models. It is commonplace for perturbations and shocks in one asset to infect other assets ("contagion"), some of which may not be obviously related, like crude oil and US Treasuries. These interactions may be due to fundamentals, statistical co-movements, or temporary substitutions.

Variance versus bias. Conflicting and complex social forces imply that market models should be impenetrably complex models of behavior and cycles perturbated by exogenous shocks. Modern optimizers and other quantitative investing techniques can effectively incorporate nonlinear effects. For example, optimizers can evaluate nonlinear risks and transaction costs with quadratic programming (QP) or second-order cone programming (SOCP) methods. When extremely low latency and specialized computer methods (like FPGAs) are required, computational ease can add speed.

The overwhelming noise and the resulting estimation error can make complex, nonlinear market models difficult to fit, calibrate, and verify. A spartan high-bias model that considers these forces in a stylized way may have accuracy and predictive power that is indistinguishable from more sophisticated methods. Simpler, parsimonious models have the advantage of interpretability, explainability, and ease of execution (speed). In the extreme, piecewise linear models can combine well-understood linear models to produce a quasi-nonlinear response that can be calculated quickly, reducing implementation shortfall.

The Execution Model (Action)

The execution model creates an action plan based on the forward-looking predictions of the market model, including the trajectory of costs like liquidity and alpha decay. As part of the fund module, it is the automated execution engine that can be designed, trained, and tested in a safe backtesting “sandbox” and then “lifted out” and used to conduct actual trading. It may interface with aggregators, order routing algorithms, order management systems (OMS), and execution management systems (EMS).

In an institutional or highly structured wealth setting, permissible trading activities should be defined and documented in a trading policy:

- **Best execution** should align investor interests with execution practices.
- **Best efforts** describe what constitutes acceptable attempts at best execution.
- **Errors** occur frequently. The responsibility for correcting the error and acceptable remedies should be predetermined and documented.

- **Counterparties** should be listed and credit risk addressed for transactions that are not guaranteed by a third party, like an exchange.
- **Governance and oversight** of the policy should be explained, including conflicts of interest and restrictions for personal trades.

Understandably, the details of these algorithms are usually proprietary and a source of competitive advantage. Even if they are simple, obfuscating intentions and intended actions may be undertaken as a defensive tactic. In the extreme, these algorithms conduct high-frequency electronic warfare with each other, including predicting countermeasures (multiperiod prediction) and counterintelligence (spoofing). Execution models are not limited to highly secretive algorithmic trading hedge funds and investment banks. Off-the-shelf execution models and *application programming interfaces* (APIs) are also available through execution agents like brokers.

Execution schemes generally combine these broad categories:

- **Benchmark aware** algorithms try to beat or minimize tracking error with a metric like a benchmark.
- **Scheduled or triggered** algorithms are set in advance or are contingent on events.
- **Liquidity or risk-aware** schemes balance market forces to optimize execution.
- **Routing, crossing, substitution, and arbitrage** techniques involve multiple assets and venues and may allow risk shifting and time shifting.

When building and choosing execution models, many technical considerations need to be examined. Small technical decisions can be critical, including constraints like timing, quantity, and risk aversion. For example, execution of small orders can benefit from rapid execution (*front-loading*) to minimize uncertainty risk. The choice of benchmark, like VWAP, TWAP, arrival price, closing price, or shortfall, can dramatically affect profitability. Alpha sources (such as predictive versus uninformed, or high- versus low-decay) should be selected based on resources and goals.[28] The *instrument and venue* choice can also be critical, including asset class, pairs, baskets, substitutions, external (exchanges) or internal (crossing), and multiple venues requiring order-routing versus single venue. *Goals* like low risk, low shortfall, or low impact should align with the trading and investment strategy. Goals can involve optimizations and constraints. *Trade scheduling or urgency* can also be impactful, including start or end time; front-load, uniform, or back-load; fragmented or blocks; individual or basket; participation rate; liquidity makers versus takers; balanced versus unbalanced; and active versus passive.

Urgency balances the cost of execution with the potential for adverse market movement. Urgency can be adaptive based on estimating the unrealized market impact of the order balance. Adjusting the urgency and adaptivity can alter the expected cost distribution. Urgent execution can minimize uncertainty risk and is suitable for high-alpha decay, uninformed market models, or models that cannot accurately predict trends. More passive models can be effective for low-alpha decisions or in situations when market models can be relied upon to predict trends. Other variations sometimes create agency risk by "benchmark hugging" to reduce the risk of criticism and maximize incentives.

Fragmentation (legs, pairs, blocks, baskets) can be used to aid execution. It can also be used for obfuscation, for instance to confuse Trade Reporting and Compliance Engine (TRACE) and swap data repository (SDR) decoding. *Liquidity* can be improved by using single or multiple exchanges, direct or smart order routing (SOR), crossing, dark versus lit pools, or trading over-the-counter (OTC). A smart engine will dynamically route orders based on liquidity needs and availability. *Participation*, such as POV or dynamic scaling, can affect performance. *Adaptivity* (set and forget or dynamic), including adaption frequency and retraining cyclicity, can be important in competitive markets.

The Broker Model (Retrospective)

If backtesting is viewed as an "eyes wide open" evaluation of a strategy rather than a training exercise that requires validation, the broker module can be unchained from the validation process. Unlike the fund module (including the market and execution models), the broker model need not be burdened by avoiding look-ahead bias. Instead, it may take full advantage of perfect foresight. The broker model can estimate the impact as accurately as possible and take full advantage of data describing events after the analysis date. For instance, the broker model may simply use the interval POV or VWAP *after* the arrival time to determine the execution price.

Fee and Cost Attribution

"In-the-loop" TCA is important and useful for predicting and continually adapting investment decisions, including choosing the appropriate market and execution models in dynamic environments. The attribution model may contain similar code to the "in-the-loop" model, but it has a different purpose and likely a different design. It may be as simple as tallying the fund module's recorded predictions compared against the broker module's realization estimates.

Information is more easily destroyed than reconstructed. Detailed records saved during execution (including inputs, parameters, and confidence) can be unexpectedly valuable in a postmortem analysis long after the original simulation. Storage is inexpensive, while time is invaluable.

Time, resources, and technology limit the scientific process of modeling and portfolio construction. Backtesting is the imperfect technique used to add more realism and complexity. Chapters 16 continues our discussion with more backtesting nuances.

Each chapter of this book is a field of study in its own right. Figure 15.1 is offered to help outline and categorize fees and costs that may be considered, along with their observability. It can serve as a checklist for transaction models to help researchers consider details that can turn promising backtests into failed funds and paper profits into actual losses.

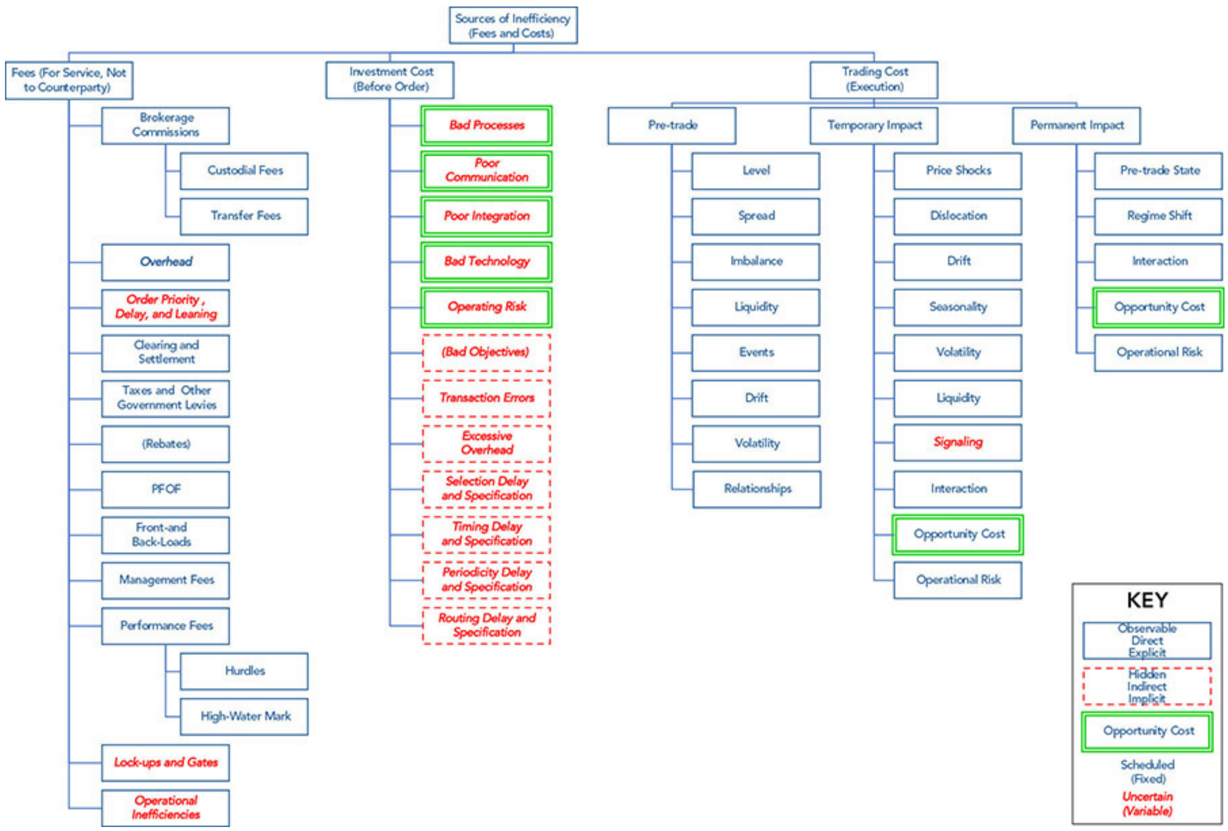


FIGURE 15-1 Transaction costs and fees

---

1. The EMSC object is MATLAB code that simulates the Bloomberg EMSX (Execution Management System) API. It is available on the book's website, www.QuantitativeAssetManagement.com.
2. The cost function and attribution will look ahead, but the prediction used to create the orders must not.
3. For example, the EMSC object is MATLAB code that simulates the Bloomberg EMSX (Execution Management System) API. It is available at the book's website, www.QuantitativeAssetManagement.com.
4. Regulated reports through the Trade Reporting and Compliance Engine (TRACE) or in swap data repositories (SDRs) allow traders to hide their intentions while providing limited transparency.
5. United States of America Congressional Record: Proceedings and Debates of the 90th Congress, First Session, Volume 113, Part 18, August 22, 1967, to August 31, 1967, pages 24155, 24714, and 24715.
6. As with the story of the two friends who are running from a tiger (one tells the other he doesn't need to outrun the tiger; he just needs to outrun the friend), proper governance and supervision can go a long way toward soothing concerned clients and regulators.
7. The objective function requires more tractable calculations than other parts of the research process. So, the transaction cost estimates used in the objective may need to be oversimplified. For instance, a complex schedule may be estimated by a piecewise linear curve.

8. The attribution model can only look forward if it is not intended to be used live or provide feedback to the decision-making functions; that would inject lookahead bias and invalidate the simulation.

9. One may argue about the details of hidden, indirect, or implicit fees and costs, but an objective analysis should make their existence apparent. For instance, one might argue what fraction of implementation shortfall is attributable to poor timing or poor execution, and the acceptable magnitude for each, but the data should reveal that shortfall did occur.

10. Technically, *costless* describes the practice of providing a product or service without an explicit exchange of money. Costless transactions and bundled fees are often expensive compared to explicit costs and fees. The bundled costs and fees may be prohibited or unattractive if their components are itemized and charged separately. For instance, a trader may pay a higher commission or accept a wider bid/ask spread in exchange for better service, use of office space, or subscription fees for research or equipment. Some geographies have passed laws to limit or prevent bundling, e.g., research must be charged for separately.

11. Many providers offer a schedule of costs and fees which may be complex but easy to encode in a function.

12. See Nassim Nicholas Taleb, *Antifragile: Things That Gain from Disorder* (Random House, 2012).

13. Frequently, a misguided attempt to create a closed-form solution or to “vectorize” code will reduce its practicality and extensibility. It is difficult to make financial analysis elegant. A notable exception is when an economist or statistician oversimplifies a relationship with the intent of gaining a general insight into a system rather than making a precise forecast or valuation.

14. The expression is the specific investments made to affect a trade idea, e.g., which forward rate on a yield curve best isolates an interest rate outlook or dislocation.

15. This change in security selection may be permanent or temporary. Temporary changes may result in purchasing a placeholder investment to gain effective exposure to the same or similar factors until the originally specified instrument can be purchased efficiently.

16. Some people make a distinction between the shortfall and slippage.

17. The spread depends on the size of the order and compensates for the risks of being passive (providing liquidity), including drift and informed (“toxic”) orders. The spread should be skewed from the neutral half-spread to account for anticipated price improvements and other forces.

18. “Hidden” liquidity in dark pools and other crossing networks can alter this imbalance.

19. Volatility (“vol”) varies with asset type. Equity volatility clusters and has extreme distributions and is correlated with returns. Equity vol is sensitive to leverage, while sovereign debt vol generally is not.

20. Competitive market orders can involve electronic information warfare in a real sense. Identifying an antagonist’s intentions in the presence of obfuscation (spoofing) is challenging in general, and more so in high-frequency operations where computational resources are limited by need for speed and the limitations of co-location.

21. A “fat finger” error is canonically a typographical error in a financial transaction order but can be produced by automated trading systems. Voice brokers may be more lenient in canceling bad trades than electronic systems, but once the trade is executed, the resulting series of offsetting trades and market reactions may be complex and difficult to reverse. For example, the London Metal Exchange’s (LME) suspended trading of nickel in March of 2022 and reversed some trades. The result favored some and damaged others. Many funds sued the exchange, including AQR Capital Management, who allegedly lost 80 million pounds as a result of the event. Britain’s Financial Conduct Authority (FCA) was not pleased with the LME’s decisions, either.

22. See D. H. Wolpert and W. G. Macready, “No Free Lunch Theorems for Optimization,” *IEEE Transactions on Evolutionary Computation* 1 (1997): 67.

23. Theta is change in value of an option as time passes and approaches the expiration date of the option.

24. This example was discussed in Box 1-4 in Chapter 1.

25. Common measures include arrival price, spread midpoint at arrival time, opening, comparison with open-high-low-close (OHLC) prices, percentage of volume (POV) including "volume inline," TWAP (evenly distributed in time, uniform, or uninformed), and VWAP (concentrated at busy times including front-loaded or L-shaped, back-loaded or J-shaped, or both, U-shaped) for the day.

26. The structural advantages of working in a particular occupation or company may afford a trader benefits that require no skill but only a "warm body in seat." Oftentimes, large companies may pay traders a lower rate because of the benefits of the brand name and resources of the employer make the trader's tasks easier.

27. Another regime change involves the importance of closing prices. Not long ago, it was common to run models overnight based on the previous day's close. Now, models are capable of managing larger data sets and executing their analysis quickly so they no longer need to rely on daily price data.

28. Strategies like VWAP are appropriate for uninformed execution.

16

Rebalancing and Taxes

The Cost of Doing Business

Backtesting is more than an accounting exercise. It simulates the type of action we can expect once our model "goes live." We simulate the impact of our decisions by attempting to predict transaction costs and taxes. With that knowledge, we design investment, risk, execution, rebalancing, withdrawal, and tax strategies to anticipate and react to events that affect them.

This chapter completes our three-chapter discussion of building our backtest. We have discussed predicting risk and performance, and explored methods to account for transaction costs and fees. In this chapter, we will focus on rebalancing, withdrawal, and some tax strategies.

As with costs and fees, these strategies should be integrated throughout the investment process. The complexity and difficulty of integrating costs and fees throughout the various stages of modeling is prohibitive, and so they are often relegated to the unscientific backtesting phase. Costs and fees affect the usefulness of signals and the effectiveness of allocation, security selection, execution, rebalancing, and withdrawals.

Rebalancing and withdrawal strategies are similar to execution strategies. While execution strategies may result in a multiperiod event by adapting to the execution environment, rebalancing and withdrawal strategies explicitly plan for these events and optimize their execution. *Tax strategies* can be a significant improvement on execution tactics. Like other costs and fees, tax management effects can outweigh performance, and their complexity can present significant challenges to the simulation.

Rebalancing and Withdrawal Strategies[1]

Rebalancing a portfolio or executing periodic withdrawals is necessary for the long-term success of a portfolio for a number of important reasons. Most investment portfolios contain assets that are marked to market or periodically appraised and vary in value (in proportion to other investments in the portfolio). Even capitalization-weighted portfolios must adjust for corporate actions like dividends and buybacks and are *reconstituted* to adjust for additions and delistings. Most portfolios have *constraints*, like a maximum concentration in a particular holding or a regulatory requirement that limits asset proportions based on characteristics that change with time, like credit ratings. Some portfolios have liabilities and spending goals that change in ways that differ from the asset mix designed to fund the liabilities. Periodic adjustments are often necessary to maintain the intended investment policy, prevent the investment thesis from being perverted by evolving market conditions and other externalities, and to keep the portfolio aligned with goals and liabilities that may evolve with time. Adjustments may also be required if the policy or the thesis changes. Even passive indices require periodic adjustments for reconstitution.

Rebalancing improves the drivers of the Fundamental Law of Active Management. Skillful rebalancing may decrease risk by realigning investments with the investment thesis; however, this may come at a cost of reduced exposure to momentum and compounding. It may also burden the portfolio with the frictions associated with transacting. Improving diversification reduces risk and decreases alpha decay.

Rebalancing moderates trends, including the belief in long-term appreciation in favor of balance and diversification. It also moderates uncertainty, the concern that drawdowns can be long relative to the investment horizon, and utility, the need to access and employ the capital for purposes external to the portfolio.

Liability Aware, Liability Driven, and Goal-Based Investing[2]

Institutions like banks, pensions, endowments, sovereign wealth funds, and foundations frequently have dynamic spending schedules and liabilities. These schedules and liabilities may require portfolio adjustments. Private wealth portfolios may be designed to accommodate future expenses like education, real estate purchases, healthcare, and retirement, which typically vary with inflation and other exogenous forces.

Later in this chapter, we will discuss how liabilities and goals are complicated by regulation, most notably taxation. Liabilities are sometimes managed with techniques like passive *immunization* and portfolio *insurance* schemes (Chapter 12). For example, cash flow matching can result in defeasance, eliminating some balance sheet liabilities. The *surplus* (assets less liabilities) can be managed proactively with *contingent immunization* or, more thoroughly, with *liability-driven investment* (LDI) policies.

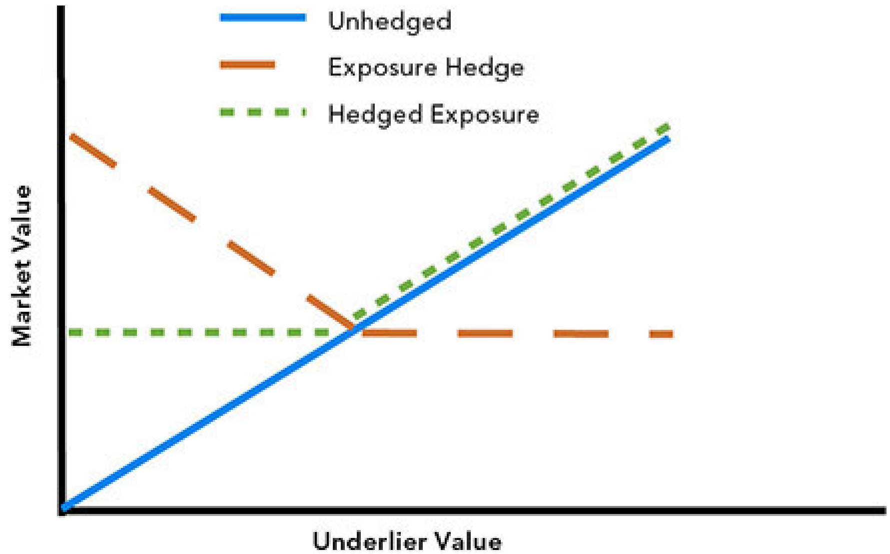
A casual observer might think that managing liabilities is a simple duration matching exercise, but the assets and liabilities can be sophisticated and nonlinear. Fixed-income assets frequently contain Byzantine features and embedded options making their price response highly nonlinear in yield. Liabilities can be formidable actuarial constructs that depend on many economic and demographic trends.

Treasury and other highly regulated operations. The confluence of the market, risk, regulation, and taxation can produce *perverse incentives*.[3] Regulations are burdensome and ever-changing. They influence how assets, liabilities, hedges, and impairment are classified, measured, and managed. Classification and measurement determine how these assets are taxed, reported, and protected (such as by setting aside regulatory capital).[4]

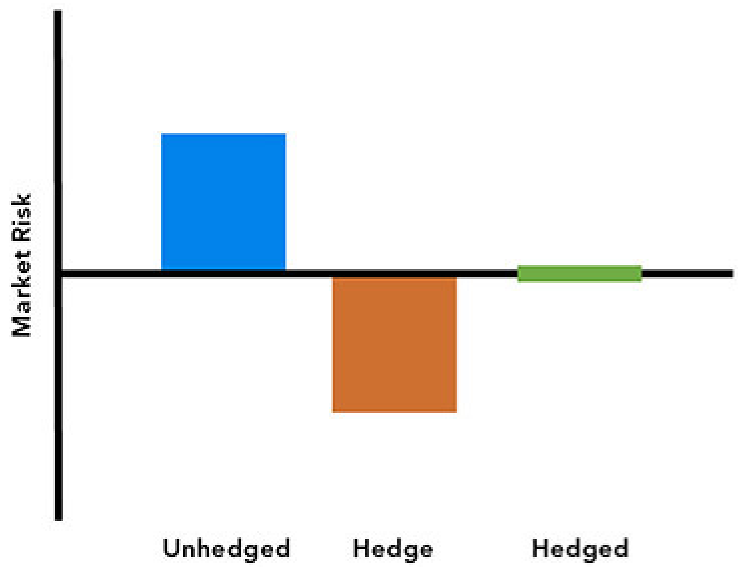
At the treasury level, accounting can alter an institution's investment decisions (even small institutions in the case of current expected credit loss, or CECL). Accounting also affects the amount of regulatory oversight (including setting aside regulatory capital) and tax burden created by trades. These rules may cause cash flows to be reclassified in the regulatory filings, such as transferred to (or taken off) the company balance sheet. The reclassification may even result in changes in the treasury's parent company's valuation by analysts who monitor these filings.[5] Many non-economic investments are made because of their regulatory and tax implications, affecting incentives as well.

Box 16-1 Unintended Consequences

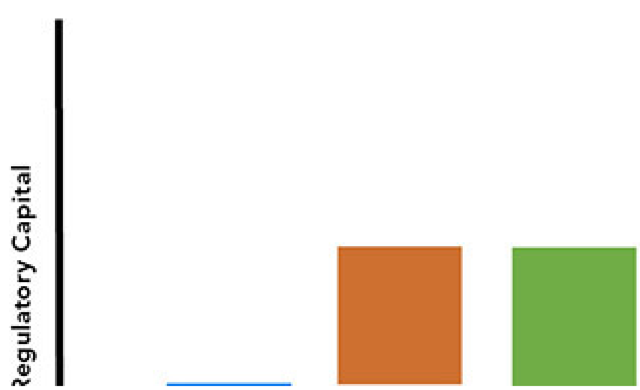
Hedging market exposure with a second counterparty can double the counterparty risk, reducing the economic benefit. It may be better to cancel the original trade (e.g., "tear up the contract" at a loss) rather than hedging with another counterparty.[6] In this example, it may be necessary to incentivize the counterparty to break a contract, resulting in a worse price than a hedging solution. The regulatory burden of a hedge may outweigh the benefit of capital protection (Figure 16-1). For example, a foreign bond may be hedged with currency forwards that require daily collateral adjustments.

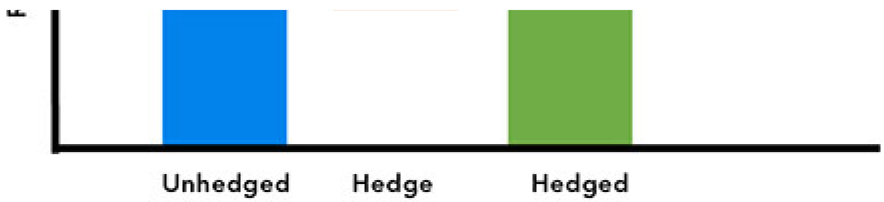


(A) Without concern for regulatory or tax implications, the risk of an investment can be hedged against market exposure.



(B) The risk can be diminished or neutralized.





(C) But the regulatory burden would be increased because of the hedge's counterparty risk.

FIGURE 16-1 Unintended consequences of hedging and rebalancing schemes

Rebalancing

Optimal rebalancing frequency, selection, and sizing are the primary rebalancing decisions. It is impractical to rebalance continuously, and it is sometimes inefficient to rebalance to optimal weights (but instead rebalance partially or to tolerances that are different than the optimal weights).

Evaluating rebalancing schemes is nuanced and can involve complex costs and tax implications. Many analyses can evaluate the relative benefits of different schemes and parameters, but a backtest is required to understand how rebalancing strategies apply to specific situations. Because most rebalancing schemes are sensitive to market momentum, trend, and reversals, the backtest will be sensitive to the scheme's specific sample periods and regimes. Rebalancing is related to *turnover* and must consider many details.

Rebalancing skilled investment decisions is beneficial to keep realizations aligned with those decisions. If these decisions have low information value or if the dispersions of potential outcomes are wide, rebalancing can be less effective or even detrimental (such as crystallizing a loss at the point of ruin). Avoiding the rebalancing of the constituents of some asset classes may resolve some troublesome investment category considerations, like estimation error. For investors with uninformed decisions, strategies like buy-and-hold, immunization, and portfolio insurance may be more suitable.

The benefit of rebalancing may be significant for effective predictions or nonexistent for passive indexers. We focus on rebalancing to limit alpha loss and maintain diversification, not to create performance. The skill required to identify the optimal rebalance frequency and produce excess returns is a timing skill and not the purpose of this discussion.

Risk management is the primary purpose of rebalancing. Risk management often varies with skill and effort, but diversification benefits from rebalancing and requires little or no skill.[7] Generally, rebalanced portfolios have lower risk than portfolios that drift from optimal allocations and greater expected returns. The return distributions of rebalanced portfolios will have thinner tails (including the right tail that benefits from letting winning positions run) by transforming luck into expectation.

Decreasing rebalance frequency reduces breadth and increases the variance of the investor’s skill, or lagged information coefficient (IC). The benefits of decisions dissipate as time passes. Decay may be high for strategies like arbitrage and low for others, like strategies employed by activists. Dynamic strategies are more sensitive to rebalancing than strategies like long-term indexing. A skillful investor’s compounded IC (*horizon IC*) will increase initially as his idea is realized but will become stale after the optimal *information horizon* is reached.

Transaction costs and fees include the expense of implementation. Higher costs will reduce the value of rebalancing by offsetting potential benefits with additional expenses. Expenses are related to the type and efficiency of the investment process and that of the assets being rebalanced. Rebalancing and *overlay* hedging with liquid and efficient assets, like indices and futures, rather than those specific to the security selection decision, may reduce costs while reducing broad exposures to categories like classes, geographies, and sectors.

Most fees and costs are realized quickly and create a temporal mismatch between the long-term benefits of rebalancing. If expenses are measured in a single-period analysis while benefits appear in future periods, rebalancing benefits may be underestimated. In multiperiod analysis, expenses are compounded, so they increase with time.

Turnover can lead to churn, which tends to be counterproductive. Without skill, turnover usually benefits only from rebalancing during extreme events and overextended downside drift. Unskilled rebalancing should rely on systematization, not timing.

Rebalancing is mean-reverting; it sells winners and buys losers (or trailing assets). Short-term momentum effects are more pronounced in some implementations (risk targeting and triggered rebalancing) than others (like fixed frequency rebalancing). Risk and drawdowns tend to even out over time, equalizing the benefits and drawbacks of various rebalancing targets. If the holdings have low correlations, rebalancing also increases diversification.

It is easy to dismiss the trade-off between expected long-term momentum and short-term dampening of return volatility (“you can’t eat risk-adjusted returns”), but an extended and deep drawdown can be painful when funds are needed to pay liabilities. Deviating from the target allocation by delaying rebalancing is a momentum timing decision that requires skill (with the tailwind of some reduced expenses, like fees, and the headwind of increased costs, like opportunity cost) rather than a passive strategy. When determining the effectiveness of the strategy, it may be more appropriate to compare a rebalancing strategy with idealized continuously rebalanced (constant allocation or constant risk-budgeted) portfolio rather than with a buy-and-hold (cap-weighted) portfolio. Each has its use, though a continually rebalanced portfolio is unrealistic.

Autocorrelation and “good runs” are a widely used argument against prompt rebalancing (“buying low and selling high”), but a catastrophe that persists for many periods (and regime change) is difficult to predict and time. Drawdowns can be intolerably deep and long; there is no definitive proof of poor timing for systematic rebalancing techniques.

Diversification can reshape the terminal wealth distribution so it exhibits lower risk. Diversification can reduce the tails (bad and good luck) and increase the likelihood of achieving the forecast (median return) and it does not require skill or timing, only a positive expected return.

Specific non-normal portfolios (like those with the negative convexity of many mortgage-backed securities, or MBS) may require special attention. Illiquidity can increase inefficiencies, which may increase costs (or provide opportunities for price improvement for savvy traders).

Unlike allocation and risk targeted rebalancing, dynamically adjusting leverage rebalances in time as well as cross-section. Leverage can seek a specific ratio or, more thoughtfully, scale to specific risk target. Leverage benefits from compounding, which in turn benefits from reinvestment and diminishes with withdrawals. Rebalancing leverage is dependent on the ability to forecast winners. Just as good forecasts, reinvestment, and recapitalization are magnified by leverage, leverage also magnifies the negative effects of withdrawals and poor investment choices.

An additional rebalancing logic is usually dedicated to governing leverage (and de-risking). Leverage is frequently employed as an *overlay* for simplicity and ease. It frequently invests in a liquid bellwether or index-based instruments (e.g., as an overlay) but is best integrated into the process at its core (factors, allocation, selection, implementation, and rebalancing). Leverage can be effective but can suffer from misalignments between the proxy instruments used to lever the portfolio and the holdings, as well as from additional turnover, related costs, and the potential for timing errors, including implementation shortfall.

Contributions and withdrawals can provide passive opportunities to rebalance with minimal additional expense. Systematic rebalancing can remove some emotion and subjectivity from the investment processes. In the extreme, derivatives, like straddles, range notes, and volatility products, can be used to enforce rebalancing discipline. Active decisions, including the approval of fully systematic rebalancing programs, can result in stressful periods of doubt when the implemented portfolio underperforms. In place of thoughtful rebalancing, simple method like fixed-frequency rebalancing is an active decision, although it carries less career risk than one that requires more ownership.

Weighting schemes include strategic, tactical, and trending (for example, opportunistic) rebalancing. Some strategies that are concerned with model error attempt only to choose which investments to make and whether to buy or sell them. More often, a substantial part of investment success relies on *how much* is allocated to each individual investment and how much is held in reserve. If the target weights are described by a static method—whether using fixed numbers or a triggering scheme—rebalancing may compensate for market movement. Heuristically allocating within categories (like asset classes or sectors) can allow weights to drift within a more diligently maintained meta-allocation. Tactical rebalancing shifts targets based on a dynamic tactical asset allocation (TAA). Allocations may incorporate drift, trend, or another systematic change, like the *glide slope* of *target-date funds*.[8]

Traditionally, allocations are defined in terms of percentages of capital. Strategic allocations and their associated bands are typically adjusted periodically ("Calendar Rebalancing," Figure 16-2A), and tactical changes to target allocations are adjusted more frequently within those bounds.

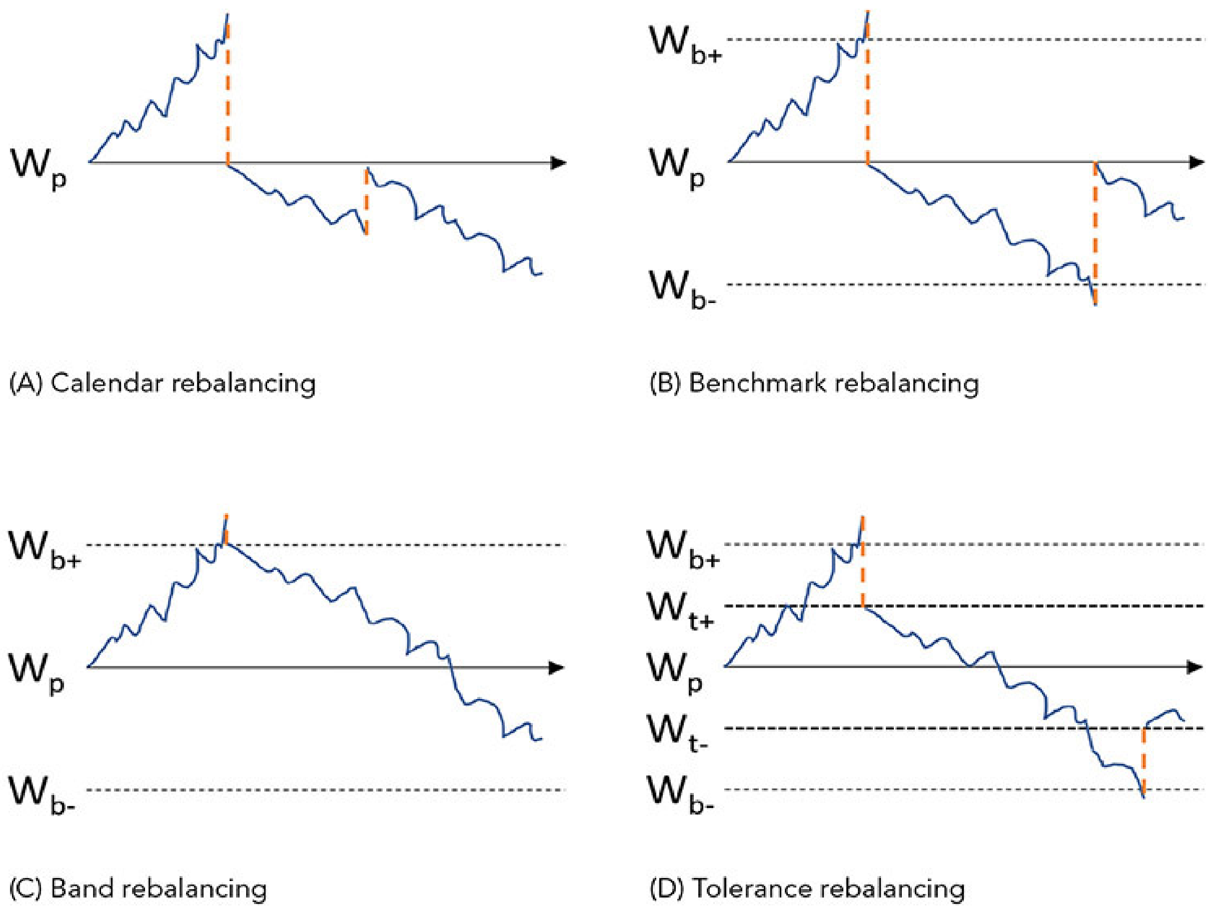


FIGURE 16-2 Rebalancing schemes

Risk budget percentages (*dynamic risk targeting*) can be more effective than allocation targeting by addressing the chief benefit of rebalancing (risk mitigation) directly and explicitly. Risk-based proportions introduce complexities, including the precise measurement and calculation of risk. Risk management requires ownership of the process and its consequences. Risk measures can change dramatically, quickly, and frequently, producing significant changes in allocation recommendations. While a board or committee may authorize a modest allocation band for an asset class, even a fixed-risk weighting may require large transactions to maintain risk tolerances.

Using derivatives, indices, and other vehicles can substitute liquid (but often blunt) investments that can maintain tolerances without buying and selling individual holdings. This is also true of schemes like portfolio insurance and risk transfers. These adjustments can lower costs, increase liquidity, and impose discipline. While they can cause a deviation from the policy benchmark, they may also increase alignment with alpha and risk factors (if the proxy instrument is better aligned than the security selection). Examples of products whose performance can align with their risk include some “smart beta” products or well-engineered structured products.

One problem with precise rebalancing is that it is impossible to maintain. For instance, by the time the trade confirmation is returned to the investment team, the weightings will have deviated from the ideal. Managers are often judged explicitly by their policy benchmark and heuristically (consciously or not) by the layman’s benchmarks (such as those announced by newscasters)—an impossible *dual mandate*.

Similarly, a manager’s stewardship is often judged by second-guessing his rebalancing activity, punishing responsibility with the wrong-way risk of political repercussions. Underperformance due to wide bands may instigate criticisms of indecisiveness, while tight bands may provoke cries of overtrading, high costs, and lack of skill. Debating these frustrating contradictions may be tempting, but political capital needs to be reserved for the most important discussions, like holding onto illiquid investments or resisting pressure to deviate from the investment process.

The most straightforward weighting scheme rebalances assets to their precise policy benchmark weights (“Benchmark Rebalancing” in Figure 16-2B) after deviating outside an acceptable range or *band*. Sizing can be important, and any balancing that differs from policy can violate delicate hedges, such as those for pairs trading.

Band or range weighting schemes (“Band Rebalancing” in Figure 16-2C) rebalance to the triggering threshold that is some level away from the policy benchmark weight (such as plus or minus several percent). If a portfolio is rebalanced to its band, it will need to be rebalanced again if it continues to drift in the same direction. *Tolerance* bands relax this restriction. Volatility can churn the portfolio and increase expenses. Bands may expand or narrow around transient events to avoid being triggered by noise. Bollinger bands are an example of triggers that adjust with trailing volatility.

Tolerance weighting ("Tolerance Rebalancing" in Figure 16-2D) is triggered by a band level but rebalances to a different, tighter tolerance band. This decreases the churn of frequently adjusting a portfolio back to its bands as it drifts.

Portfolio rebalancing can reset all classes back to their specified weights, bands, or tolerances during every rebalancing, even if only one of the classes triggers rebalancing. This can be relaxed by aligning only the offending class and minimally adjusting the others to provide the slack necessary.[9] The portfolio need not be rebalanced as a whole. Individual asset classes that need rebalancing can be brought back to their weight, band, or tolerance levels with a minimum of trading in other classes.

Assets can be located in different accounts or across different legal entities (*householding*) but treated as a single portfolio for allocation and rebalancing purposes. Taxable portfolios incur tax-related costs and may benefit from lower or higher rebalancing frequencies. (We will discuss tax implications in more detail later in the chapter.) The effect of locating higher returning assets in qualified accounts may be more psychological for the client than impactful in terms of after-tax returns. Attempting to distribute assets in qualified and unqualified accounts can make investors feel like they are making an impact. Tax-loss harvesting and similar techniques can make trades in non-qualified accounts tax-efficient; skillful householding can be relevant and significant.[10]

Rebalancing triggers include fixed time intervals or "calendar" rebalancing (Figure 16-2A), which may take place after monthly or annual investment committee meetings or a similar event. Uniform (*uninformed*) periodic rebalancing, especially if infrequent, may miss significant movements (*gap risk*), like the buying opportunity at the beginning of 2020, but also reduce the costs of transacting frequently.

Bands or ranges are defined by the investment policy to allow for some market drift between rebalancing. Without bands, allocation or risk-budget rebalancing would need to be continuous. A *matched book* (*book making* or *market making*) and high-frequency arbitrage portfolios are close to continuous but deviate with *flow* (supply and demand) and other forces.

Event triggers are more precise than bands but are less transparent and may seem more arbitrary. Autocorrelation and cross-correlation of triggers can reduce explainability. As with lagged ICs, the delay between trigger and action can be a source of inefficiency.

Flows, like dividends, benchmark additions and delistings, fund subscriptions, redemptions, inflows, and withdrawals, offer "free"[11] opportunities to bring a portfolio closer to alignment by allocating to holdings that are underweight rather than *pari passu*. When flows are known in advance, agreements between the fund and the broker can limit impact. Since flows are difficult to forecast, they are also difficult to model, other than through sensitivity analysis.

Fixed periods can be combined with bands so that rebalancing decisions are considered on rebalance dates (such as investment committee meetings) and only executed if required on those dates, but not in between (like a *Bermudan* option). Emergency or exceptional rebalancings, due to unexpected events are possible, too. *Opportunistic* or *lazy* rebalancing[12] reallocates asset classes individually when they exceed band thresholds and not entire portfolios. Adjusting individual classes rather than entire portfolios can result in allocations tuned to more compressed *stress correlations* than those suitable for long-term relationships (including the environment that follows the trigger after a shock dissipates). Opportunistic rebalancing can also use flows and limits trades to adjustments that are large enough to "move the needle," further reducing costs and fees. Opportunistic rebalancing is tolerant of deviations and is most suitable when investment skill is low and least appropriate when alpha decay is high or precise hedges are needed.

Holdings constraints restrictions may or may not affect frequency and triggers, but they may complicate the process. Constraints are a "lawyer's approach" to the agency problem and risk management in general.

Constraints can create perverse situations. Many investors were caught flat-footed during the Great Financial Crisis of 2007–2008 (GFC) when some external managers sold their most liquid investments rather than their least desired holdings. This had the unpleasant consequence of leaving a suboptimal portfolio for the investors who did not redeem in the first wave. When the portfolio is in distress, gates and lock-ups can become advantages instead of hindrances by imposing discipline on investors and preventing "fast money" from forcing managers' to make decisions that damage the portfolio and hurt the remaining investors.

Some issues may complicate constraints. For instance, policy, client, or discretionary restrictions like concentration caps (maximum position size) may constrain decisions. Prior (legacy) or illiquid holdings and separately managed accounts or SMAs (including legacy portfolios or exposures to private investments like core businesses for companies or family businesses) may force individual clients to retain holdings and cause their performance to deviate from the model ("off model"). Frequently, some assets will be *held away*, and the account will only hold the *completion portfolio*.

Asset illiquidity (like finding a buyer for real estate) can require multiple implementation frequencies for different asset types. Managing asset-specific frequencies can be messy, but the investment team can also exploit it.

Instruments like total return swaps (TRSs) can alter the economic exposure of some of these awkward holdings but will not reduce the estimation risk of their characteristics (like risk, return, valuation, and factor exposures). External timing constraints of execution decisions can be delayed by externalities other than market forces. These delays should be incorporated into the transaction cost model. *Policy exceptions* may require approval by boards or clients. *External funds* may restrict purchases by being *hard-* or *soft-closed*. They may restrict outflows with lock-ups, notice periods and frequency restrictions, gates, payment periods, penalties, and other common or bespoke conditions.

The form of redemption may be complicated by SMAs, side letters, and side pockets. Payment may be made in cash or *payment-in-kind* (PIK). If PIK, then the assets received as payment must be incorporated in the portfolio or disposed of. Hedges and other engineered trades may require precise sizing, fixing relative allocations to a prescribed relationship. Tax lots, asset location, and householding may also constrain decisions.

Withdrawals

It is natural to focus on rebalancing to maintain the efficacy of our decisions and to use flows to move the portfolio closer to the ideal allocation (the policy portfolio). The investment team may overlook liability planning and optimal withdrawal strategies (except for organizations that focus specifically on that task, like pensions and treasury departments) beyond considering wrong-way risk, risk tolerance, investment horizon, and asset location.

A holistic, liability and goal-aware, multistage investment process can produce better outcomes. To do so, apply the concepts and tools we have discussed (like cash flow modeling and agent-based models) to adapt goals and horizons to optimize withdrawals. A significant, often dominant concern for withdrawal scheduling and planning is the tax strategy.

Tax Strategies

Taxation is an intricate and constantly evolving topic, and we will only touch on it briefly and only from the US perspective. We will then explore an application of tax-aware investing: *tax-loss harvesting* (TLH). Tax *evasion* is illegal,[13] and tax *avoidance* must be motivated by investment decisions rather than simply a desire to not pay taxes.[14] Tax-*efficient* investing using tax-*aware* methods is common and effective. Remedies involve reduction and deferment.

Complexity. Taxes are multifaceted. They are usually *progressive* and applied in stages, and may be charged by federal, state, local, and foreign governments and municipalities. They may be general (such as income-based) or use-based (like road tolls). Many deductions and exclusions exist for various situations and entities (like trusts) and individuals (like estates and gifts). Rates and rules change frequently. Experts like accountants, tax attorneys, and estate planners can be valuable for understanding and planning.

Complexity represents an opportunity for a quant, creating niches to explore and exploit that competitors may miss. Complexity need not be difficult if modeled in layers with scalable technology (not spreadsheets!) and addressed iteratively by:

- Aggregation of cash flows (such as *bottom-up*)
- Successive refinement (such as *top-down*)
- Highest *impact* and lowest tax *efficiency*

Designing the Investment Process

Identify the purpose and goal of the investment effort, the assets and what they generate over time (such as appreciation and income), location by account or entity (such as allocating between qualified and non-qualified accounts and *householding* across entities), and their tax status. Identify withdrawal and liability needs and their penalties and taxes. *Choose* what tax effects are most compelling. In most cases, a comprehensive analysis is neither feasible nor necessary. *Measure* after-tax value for the appropriate location (account or entity), including *embedded unrealized taxable* value, such as a low-*cost basis*. Uncertainty, such as for multiple clients with different tax situations, may require a probability distribution-based analysis, such as Monte Carlo simulation. *Model* the taxable effects to allow for changes in parameters and assumptions. The *portfolio factory* approach (Chapter 12) for automating and scaling customizing solutions for themes can be applied effectively to tax situations and create tailored portfolios while managing operation risk, increasing organizational brand value, and leveraging the competitive advantages of the fund's employees. The downside to the write-it-once, fix-it-once benefit of leveraging technology is that any error will flow through many products.

Restrictions on assets and liabilities, timing and location, domicile and control, growth and income, and retention and distribution may be required for some entities to maintain tax status and must be reflected in the simulation. *Analyze* after-tax costs and the effect on profitability and risk. *Create* strategies, including location, efficient vehicles, and timing to optimize projected or potential after-tax effects. *Plan* and protect against estimation and model error, including the inevitable change in tax rates and market returns.

Managing Tax Effects

There are a variety of purposes for managing tax effects:

- **Maximizing** economic value (after-tax appreciation and income) is the most immediate and short-term goal. It is also the most scalable and applicable to the broadest audience (such as for funds open to outside investors).

- **Risk mitigation.**
- **Goals and liabilities** for long-term tax plans, including education and housing.
- **Freeing up capital** and equity monetization from cumbersome exposures (such as restricted or low-cost-basis founder stock, illiquid assets like real estate, or concentrated risks like privately owned business). Creative solutions include hedging and borrowing against the asset and swapping or replicating desired cash flows synthetically and zero-cost collars.
- **Exit** strategies including spend down, legacy, and periodic or terminal gifting.

The form of legal entity that transacts or holds the investment can determine the regulatory benefits and restrictions that are applied to it. Institutions and corporations may be exempt from taxes if they constrain their activities as required by regulations. Cross-border activities and other complex financial actions and relationships force some institutions to consider tax effects. Product offerings that offer tax efficiency include ETFs, sovereign and municipal bonds, swaps, pass-through investments, structured tax trades, variable prepaid forward contracts, exchange funds, and qualified opportunity zone funds.

Professional investor arrangements (funds and dealers) and partnerships are often designed with their tax treatment in mind. Onshore hedge funds and private equity funds employ partnership accounting (*schedule K-1*[15]). *Unrelated business taxable income* (UBTI) can violate the nontaxable status of some retirement plans, trusts, charities, and other structures. *Payment-in-kind* (PIK) can help lower taxes for these funds. Hedge fund expenses may be offset if the fund can prove *trader status*.[16] Otherwise, expenses flow through as miscellaneous on the K-1 and will be wasted. Some clients may not be able to participate in some deductions, such as those who have a large *alternative minimum tax* (AMT) burden or those who do not have enough income to deduct management expenses.

Tax-aware investing by managers can include a qualified dividend allowance for a long/short fund that is long for at least 61 days and short 46 days, or when a merger's target is held for at least 61 days. Exemptions may apply for holding longs and shorts in the same fixed-income instrument.

Offshore funds can be corporations that are not exposed to UBTI. Offshore funds may be less attractive to US citizens if they trigger *passive foreign investment company* (PFIC)[17] status.

Mutual funds, ETFs, CTAs, and other entities have different tax treatments. We discussed this at some length in Chapter 4. Taxation of transactions and corporate events is often considered when determining which structure to use.

Wealth and individuals. Taxes are a significant, sometimes dominant cost for investors that are not exempt. Taxes are usually asymmetrically negative (though there are exceptions, like *franking credits* in Australia). While institutions are rationally tax-averse, private clients often irrationally avoid taxes, diminishing their opportunity set while myopically focusing on reducing their tax burden. Conversely, complex tax situations can be challenging to explain, understand, and report. The difficulties in explaining the benefit (or lack thereof) favor the salesman over the quant. Wrap fees, exit barriers (like the cost of changing managers), and other hidden expenses, as well as high-fee products, can burden captive clients who seek "free" services like tax mitigation. Wealth entities are numerous and include: multigenerational and perpetual family offices, single-generation family offices, individual investments and separately managed accounts (SMAs) including qualified and non-qualified accounts and trusts, and external funds.

Modeling and measurement of tax efficiency have not been standardized. "Fit-for-purpose" formulations need to be chosen and require compromises in all but the simplest situations. Though it is impossible to discuss all tax issues here, overlooking any can be dire.

Income, appreciation, and income reinvestment can be a constraint imposed by the investor. This may include treatment of unrealized gains and losses, interest, dividends, amortization, and embedded liability (such as potential capital gains exposure [PCGE] and the capital gains realization rate [CGRR]). *Required distributions* and penalties for early withdrawal may need to be considered. Accounting treatments like *marked-to-market*, *hold-for-investment*, and amortization may be relevant. The investor may have a preference for pre-liquidation or post-liquidation over different horizons (short-term, long-term, estate, charitable) and holding period, including those for estate planning.

Buying and selling measures, which may be asymmetric, can determine performance and fees, commissions, and costs charged for tax-efficiency services are often considered important by investors. *Lot specific* tax-basis and selection—like last in, first out (LIFO); first in, first out (FIFO); and highest in, first out (HIFO)—and potential for future liability or mitigation (such as with estate step-up or charitable gifts), may have a significant effect on tax efficiency. *Timing details* like wash sales, short/long term, five-year capital gains, section 1256 (futures and options) and 998 (currencies), and *vehicle specific details* like original issue discount (OID) or premium, qualified or non-qualified dividends, and phantom income from stripped zeros can play a significant role.

Equating qualified and non-qualified investments (for example, the difference in the benefit of using tax-exempt bonds versus taxable bonds for multiple clients of a fund) is more complex that it may seem.

Many considerations include *progressive and ever-changing tax tables* (effective versus statutory tax rate), *geography-specific* effects for tax havens, locally taxed geographies, and globally taxed populations, including foreign withholding and resident/nonresident treatments, the trajectory and shape of future tax regulation, and unobservable effects like alpha decay and reputation risk (such as being portrayed as unpatriotic for avoiding taxes).

Tax-Loss Harvesting

Tax-loss harvesting (TLH) is an attractive, dynamic arbitrage strategy that is an example of how even straightforward tax strategies can be subtle to model and analyze. Sensitivity analysis and stress tests are essential for any complex analysis that suffers estimation error or relies on uncertain futures.

In its simplest form, TLH is an active strategy that involves selling investments to realize offsetting losses and replacing them with similar investments to maintain factor exposures. Substitution accomplishes the benefit of offsetting or deferring taxes on gains while maintaining the investment thesis of the portfolio. Deferred taxes benefit from compounded reinvestment. Gains are not limited to appreciation, but may be collected as a result of reconstitutions and corporate actions.

Short-term losses, taxed at higher rates, can be harvested to offset long-term gains, taxed at lower rates. Losses realized during high-income, high–tax bracket years can be deferred and may offset gains during lower income, low–tax bracket years (such as during retirement). They may also be deferred indefinitely using vehicles like the estate tax step-up, gifts, or other methods.

Tax-loss harvesting is easy to explain and may seem simple, but much of the apparent benefit is ineffective. A thorough simulation is not a trivial exercise. The complexities and specifics of tax analysis require simplification or complex rules.

Market conditions affect the breadth of losses available for harvesting. Combining market conditions with tax lots across households is similar to using *vintages* and *cohorts* to analyze loan pools and structures. Market conditions create opportunities after the profits of a legacy portfolio are consumed. Cooperating markets are uncommon, and many years provide no substantial TLH profits.

Be cautious about *market direction* and the sequence of alternating directional trends. Decreasing prices for limited periods provide more losses from higher *cost bases*. In the extreme adverse case, *portfolio lock-up*, or *ossification*, can prevent harvesting if a portfolio has appreciated so much that the cost basis has become too low to permit losses (*burnout*).

Low short-term correlations facilitate TLH in trending markets by allowing losses for harvesting while others appreciate. *High long-term correlations* can be an indicator of TLH opportunity breadth because they permits similar investments to be substituted for those that are harvested. *Low long-term correlations* can identify investment picking regimes for use by strategies that seek to replace losing investments with better investments, rather than merely substituting them with similar investments.

Volatility can offer TLH opportunities. Like cross-sectional variations, time-series variations present more opportunities with more short-term losses. Transaction costs and technology affect the minimum lot size (trade frequency) and cost of transacting frequently. Operational costs and risk are significant for any active strategy. *Realization is more about deferral than profit*. Since the benchmark portfolio is not harvested, the TLH portfolio appears to outperform on an after-tax basis because TLH losses are realized, but this ignores the *embedded* (unharvested) losses that would have remained in the portfolio that would be realized if its assets were sold later.

The embedded profits effect is as real as carryforwards are. *Carryforwards* create the potential for future taxes that deteriorate the benefits of TLH. For example, tax rates may increase in the future, making carryforwards insufficient to offset future taxes. Tax rates were much higher in the past, and they may be again. Realization of tax losses is *time shifting* and can result in the portfolio being burdened by higher tax brackets in the future. When future tax rates are expected to be higher, it may be better to harvest *gains* than losses.

Deferred taxes can be realized through a loss of control, such as in passive investments, an estate (*step-up*), or charitable giving (avoiding *liquidation tax*). More sophisticated methods can involve structures including surrendering control to a trust and borrowing against the trust value.

The cost basis of the legacy portfolio and inflows (used to buy higher-cost investments) and dividends (qualified after 61 days) provide high purchase prices to offset low-cost bases and are major determinants of profitability. As losses are harvested, substitute investments are bought at low prices (cost bases) and make future harvests less profitable (*tax-alpha decay*).

The IRS allows applying tax losses against similar gains (short-term versus short-term and long-term versus long-term). The tax losses that cannot be matched this way are then permitted to be used to offset other gains (short-term versus long-term). Tax losses that are still not matched can be used against a small amount of ordinary income. Finally, the rest can be carried forward until the investment horizon (such as death).

Loss carryforwards may accumulate without profits to offset them, and TLH benefits dissipate nonlinearly to a low asymptote. Without significant inflows and favorable market conditions, portfolios lose the benefits of TLH in a couple of years (burnout).

A TLH marketing paper by Wealthfront[18] estimated that their TLH method only produced significant *tax-alpha* in 5 years of a 13-year period beginning immediately before the dot-com crash in 2000 (including the first four years and the GFC of 2007–2008) and a negative value in two years (before and after the GFC).

The time value and economic benefit of TLH are far lower than the losses harvested. Two contradictory forces are at work. Realized (embedded) benefits are harvested quickly and dissipate after a few years. The economic benefit of harvesting those losses only accumulates over long periods as the *tax-alpha* is invested and compounds. The economic benefit of deferral comes from the reinvestment of the harvested capital and not in the harvest itself.

Consider a stock purchased for $10 and sold at $20 ten years later. $10 of capital gains would be taxed. If the same stock were purchased and sold in five years for $9, and a substitute bought for $9 and sold in the remaining five years, a total of $10 would still be taxed, perhaps at a higher rate. However, a tax loss resulting from the $1 decrease in basis would be accumulated. If the capital gains tax for the transaction is 15 percent, the $1 could be used to reduce 15 cents of taxes *if an offsetting transaction occurs*. The TLH benefit results from the reinvestment of that 15 cents over the remaining five years, and not the $1 or even the 15 cents itself. The benefit is a fraction of the tax alpha. If harvesting produces 15 cents of tax alpha, and reinvestment produces 8 percent annually over the next five years, the economic benefit is a meager total of just seven-tenths of a percent over five years—a little more than an eighth of a percent annualized over five years:

$$0.14\% = \sqrt[5]{100\% + \frac{(\$10 - \$9) \times 15\% \times (100\% - 8\%)^5 - 100\%}{\$10}} - 100\%$$

In addition, the analysis ignores the impact of future tax rates on the lower cost basis.

Skill-based alpha losses due to substitution (*tracking error*) are related to how specific the investment thesis is to the original investment. Less precise strategies (like index replication or portfolios that use geographically broad or thematic instruments) can use sampling to create a reservoir of substitute investments, but other strategies are less forgiving. The *wash sale rule*[19] is designed to prevent perfect substitution (such as buying or selling investments that are substantially the same within 30 days across *all household* accounts), creating basis risk for highly skilled security selection.

Complete realism is rarely achieved in simulations due to complexity. Critical items to consider are international withholding and tax treaties; federal, state, and local taxes; tax brackets based on income; short-term and long-term capital gains; ordinary income; qualified dividends; alternative minimum tax; other transaction costs and fees (including the bid-offered spread and unobservable costs); market access; changes in suitability and preferential treatment of asset types; changes in tax law like the loss of the estate step-up. Tax tables change with time and need to be reflected in historical analysis, just as reconstitution needs to be incorporated in index analysis. Tax tables and independent variables (like client income and domicile) must be forecasted.

The effort, expense, and uncertainty of TLH usually result in some positive benefit. TLH must be balanced against the strategic goal of life planning and risk management. It is best applied against passive, low-alpha, and suboptimal portfolios. It benefits some styles over others, like value over momentum. Depending on an investor’s skill, breadth, risk tolerance, and resources, pursuing other solutions may be a more effective use of time and resources.

The loss of alpha due to substitution is uncertain, as are future tax rates. As with optimization, restricting leverage and a *long-only constraint* eliminates a significant fraction of potential TLH benefits. The flexibility provided by short positions (such as with a 130/30 portfolio) creates a greater breadth of opportunity to offset short-term gains with short-term losses. In an upward-trending market, shorts can provide losses that longs may not. As discussed earlier in the book, shorting adds significant challenges to realistic simulation and actual investing.

■ ■ ■

Rebalancing is an important consideration for many portfolios and is not a simple trade-off but rather a complex analysis. Ideally, it would be incorporated throughout the investment process, including the selection of features, alpha, and risk, but is frequently relegated to the backtest to make the solution tractable.

Taxes can be the most significant controllable factor in some portfolios and can have a psychological impact that outweighs its financial significance. Like costs and fees, these complexities are frequently simplified or even ignored. Far better, tax strategies should be modeled with reasonable efficacy and improved over time so that they become formidable competitive advantages.

---

1. Strategies for deploying contributions (incoming cash) are important, too. Unlike many withdrawals, additional cash may be unplanned. Also, while some withdrawals are small, withdrawals from investment portfolios can often be large since they can be used to fund life goals or events, like financing education or buying a house.

2. *Portfolio immunization* matches the duration of assets and liabilities in an investment portfolio to manage interest rate exposure. *Portfolio insurance* schemes seek to hedge exposure to a declining market without reducing equity holdings, e.g., by selling futures as part of an overlay to reduce the percentage of equity in the portfolio. *Contingent immunization* strategies are triggered, e.g., by a predetermined return (large loss) or wealth level. *Liability-driven investing* matches some assets to specific future liability needs and invests some or all of the remainder in risk assets. Similar to LDI, *goal-based investing* matches assets to fund future goals and measures performance based on the ability to meet those goals, rather than comparing results to a benchmark or absolute hurdle.

3. *Perverse incentives* are enticements that work contrary to the intended purpose. Rules and constraints can often encourage the same behaviors they were designed to prevent, like wanton risk taking during periods of drawdown.

4. Aaron Brown, author of *Red Blooded Risk*, *The Poker Face of Wall Street*, and *A World of Chance*, and then chief risk officer of AQR Capital Management, told me something like “if the trade doesn’t make economic sense, it’s probably a tax [or regulatory] trade.”

5. In a separate conversation with the chief investment officer at a Fortune 500 company, he somewhat sarcastically described the company as a “small . . . company with a large pension liability.”

6. For example, a firm may create a “bad bank” to isolate non-performing assets from the parent’s balance sheet and then “write off” the loss. The portfolio manager in charge of “winding down” the portfolio may be benchmarked against the write-off valuation. Since the write-off level may have been determined by an outside accounting or consulting firm, it may be low. The consultant would not want to be embarrassed by overestimating the value of the assets and causing the bank to take an even greater loss against that valuation when the assets are disposed of. This would make the benchmark easier to outperform, so the portfolio manager may be more concerned with disposing of the assets quickly than fighting for a better price. If he combines this with the burdens of *additional* counterparty risk created by expensive hedges, he may prefer to merely break the trades at a loss (when compared to the purchase price), which may be fair value or better when compared to the write-off valuation.

7. Like many investment methods described in this book, effective rebalancing is a structural advantage (a “tailwind”).

8. *Target date* funds are generally designed to transition from a risky portfolio (e.g., suitable for a young person) to a less risky portfolio (e.g., suitable for a retiree) gradually from the purchase date to a specific target (retirement) date. The *glide slope* or *glide path* describes the gradual de-risking

adjustments made to the allocations of the fund over time and are named that way to evoke the image of a smooth landing.

9. Once a holding "breaks the band" and is rebalanced, the trader needs to decide what to do with the other holdings of the portfolio. Rebalancing the offending holding will change the cash balance of the fund and may even put it in deficit. Altering the other positions should be guided by the desire to minimize rebalancing costs and reduce the chance of a subsequent rebalancing. This may involve changing the other holdings to bring them closer to policy. A more thoughtful suggestion might involve analyzing the expected return and volatility of the other positions to determine which is most likely to cause a future rebalancing. The trader may reduce the exposure to risky positions to simultaneously mitigate the chance of another rebalancing event and manage the cash balance.

10. Householding can be difficult if some of the clients' assets are inaccessible. For example, a client's retirement account might be managed by his employer at another company ("held away").

11. Small flows may be more burdensome than large ones, so the trading team may set limits on minimum transaction sizes. Costs, fees, and operational risk can make "free" adjustments expensive.

12. Gobind Daryanani, "Opportunistic Rebalancing: A New Paradigm for Wealth Managers," *FPA Journal*, 2008.

13. U.S. Supreme Court, *Gregory v. Helvering*, 293 U.S. 465, 1935.

14. Paul Kiel and Jeff Ernsthausen, "How the Wealthy Save Billions in Taxes by Skirting a Century-Old Law," *ProPublica*, February 9, 2023.

15. Schedule K-1 (Form 1065) is an IRS tax form used for reporting partners' income, gains and losses, and dividends for partnerships (and for shareholders of S corporations).

16. Topic No. 429 Traders in Securities (Information for Form 1040 or 1040-SR Filers), United States Internal Revenue Service, October 22, 2022.

17. The PFIC test is at least 75 percent of gross income is derived from non-business operations and at least 50 percent of assets shed passive income. See IRS Form 8621.

18. Wealthfront Tax-Loss Harvesting, https://research.wealthfront.com/whitepapers/tax-loss-harvesting/.

19. Internal Revenue Service, "Publication 550: Investment Income and Expenses," https://www.irs.gov/pub/irs-prior/p550--2020.pdf.

17

Time Series and Cross-Sectional Analysis for Financial Markets for Part III

What Works

We concluded Part II with a discussion of algorithms relevant to that part. In that chapter, we described techniques used in preparation for prediction and simulation. We will now address algorithms related to the topics covered in Part III. Because we are focused on the investment process rather than the technical application, the chapter is organized in the same general order as Part III.

Prediction

Predicting future valuations and factors using the features and objectives we studied in previous chapters helps identify the potential risk and reward of our investments, so that we can select and allocate them appropriately. When forecasting “alpha” and risk, it is essential that we include realities like transaction costs and implementation strategies, as well as identify adverse interactions and overly salient signals that may skew decisions.

Alpha Models

Alpha models are important in improving predictive power and confidence in the factors, objectives, allocations, and sizing, as we discussed in Chapter 11. Alpha models are used in a variety of applications related to tracking trends and predicting performance, such as:

- Picking stocks using methods including least absolute shrinkage and selection operator (LASSO), boosting and bagging, or classifying them using non-negative matrix factorization (NMF or NNMF), and K-means
- Picking debt using random forest, neural networks (NN), support vector machines (SVMs), K-means[1]
- Trend following and mean reversion using LASSO and state-space models
- Index tracking and hedge fund replication using LASSO, boosting and bagging, and sparse Kalman
- Relative price movements often use Granger causality and cointegration, one-vs.-rest classification, impulse response, affinity propagation, manifold embedding[2]
- Regime identification can be used for tactical allocations security selection, or a risk overlay[3]

Signals are used to identify opportunities—also referred to as *anomalies* by those who believe in efficient markets. While signals research is easy to find in the literature, signals are much more difficult to exploit as a business model. Signals are used to predict responses that can be numerical, such as returns, and may benefit from regression, or the use of categories, like “buy,” “sell,” or “hold.” Signals can be combined, either cross-sectionally, such as averaging the results of different models, or in tandem, such as using a classifier to identify features and using those features in a regression to predict a response.

Numerical signals are commonly used in data analysis. Penalized regressions that use shrinkage to deflate correlations to make correlation matrix inversion possible, like LASSO (to prune predictors), ridge (to decrease collinearity), and elastic nets (combining both) can outperform traditional statistical methods like ordinary least squares (OLS) regressions. Trees that are restricted to a few layers, including gradient-boosted regression trees, are also effective. More complex multifactor signals with interactions can benefit from SVMs and artificial neural networks (ANN)—especially long short-term memory (LSTM) applications—but these methods hinder performance attribution and may make investor meetings cumbersome. There are many interesting applications of esoteric techniques for signal generation, especially those incorporating lagged signals, hidden or latent layers, and risk scaling.

When the number of predictors is large relative to the number of observations (Big-p, Little-n), then trees, ANNs, SVMs, and K-NN can be effective. Methods like ordinary least squares and logistic regressions are not generally effective in these situations.

Classifiers can be more effective in predicting returns. It may be natural to want to predict the returns themselves, but it may be helpful to simplify the task by predicting classes such as wins and losses. Predicting returns using classifiers may employ methods such as SVMs (with various kernels), logistic regressions (logit), linear discriminant analysis, classification and regression trees (CART), passive aggressive classifier (PAC), stochastic gradient descent (SGD), and trees. Other ingenious classification methods include Birch hierarchical clustering for big data, affinity propagation for small clusters and low-noise, and hierarchical density-based spatial clustering (HDBSCAN); these techniques manage the influence of outliers. However, unsupervised classification does not work as well with returns data as it does with factors (such as identifying shift, steepness, and convexity as drivers of the yield curve), dimension reduction, independence, and importance.

Ensembles, boosting, and stacking can be much more effective than single complex methods (large models) when predicting returns because returns are complicated by trends (autocorrelation) and other confounding characteristics that make prediction challenging. Frequently, it is easier and more effective to divide a complex task into smaller tasks and then average the results.[4]

Boosting combines a series of weak prediction algorithms or learners, using techniques like extreme gradient boosting (XGBoost) and adaptive boosting (ADABoost) by using the output of a previous model as an input for a subsequent model to improve predictive power. Boosting, stacking, and bagging are frequently applied to tree models but need not be limited to them.

Bootstrap aggregating, or *bagging*, including various forest algorithms, combines models through a voting technique, like a weighted average. Bagging is a powerful tool for combining multiple models. Bagging often dominates complex individual models, but tends to be more effective in reducing variance than increasing predictive power. Boosting is not just joining independent algorithms. it is a process for combining them while considering their interactions.

Stacking or *meta-labeling* is similar to boosting and bagging. But rather than voting on a prior tier of models with a simple formula like an average, stacking uses a model to weight the outputs of the models in the previous layer to minimize variance and improve predictive power. Stacking is a natural and useful method for combining the work of multiple teams or investment managers.

Decision trees, like common flow charts or causal graphs, are especially easy to interpret and explain. For instance, the first branch of a tree may select stocks based on return on equity (ROE) and then, conditioned on the ROE split, may select stocks based on cash flow at the next level, and so on. However, because they use thresholds for splitting branches, individual trees can be unstable and overfit, like portfolio optimizations. Bagging techniques, like bootstrap sampling, can make trees more robust, but like robust optimization for portfolios, the sampling can take a long time to train. Trees are natural classifiers but can also be used for regressions. For example, they may be used to fit a piecewise line to a single factor tree, or a piecewise surface to a two-factor tree.

Expected returns, like buy and sell classes, are a natural goals for alpha models. Expected returns can be estimated by simply calculating a historical mean or OLS regression.

Fundamental estimates involve fundamental data, such as balance sheet and income statement ratios. Many attempts have been made to apply machine learning (ML) to fundamental analysis to predict returns. In the academic literature, SVMs and ANNs were used early on, followed by natural language processing (NLP), ensemble methods, cluster analysis, decision trees, and LASSO. Radial bias function (RBF) kernels are commonly used to allow SVMs to perform large dimensional nonlinear classifications. NLP analysis of text data, including news and social media for content, context, and sentiment, can provide alpha signals and can be used for *meta-learning* to select predictive factors.

Time series. Deep learning is not as powerful and successful for time series analysis as in other applications. Financial time series are far more difficult to predict than most other kinds of data, like production line quality data that have well-defined distributions of control statistics. Deep learning requires large amounts of data and can be suitable for high frequency trading and alternative data—like satellite images, sequence analysis, and textual analysis. Specialized models and oversampling techniques, such as those using generative adversarial models (GAMs), offer some promising results.

Recurrent neural networks (RNNs) is a deep learning method that can manage autocorrelation, similar to the traditional autoregressive integrated moving average (ARIMA) family of techniques commonly used by econometricians. Long short-term memory techniques (which remember the recent hidden state and then forget it) are common. Gated recurrent unit (GRU) methods are like LSTM applications but more efficient and are often used with RNNs. Recurrent skip layers can be used to deal with seasonality.

Convolutional neural nets (CNNs), which are commonly used to analyze images, successively find new features in a manner similar to boosting. For time series applications, this means a CNN may find seasonality (without needing to choose a lag) and then trend. A novel use of CNNs involves identifying classic technical analysis patterns like *head-and-shoulders* or *wedges*, transforming time series to images and then analyzing the images. Critically, human traders and market technicians usually employ significant discretion when choosing which interpretations to believe since there are often many technical indicators that describe the same price action and they often contradict each other. A weakness of CNNs is that they treat all spaces with the same emphasis, while periods in time series can have different importance; causal padding can be used to add a temporal dimension to CNNs.

Deep reinforcement learning combines deep learning with relentless reinforcement learning (in contrast to human learning, which stops when the learner becomes tired). This technique has produced stunning results in areas such as command systems for autonomous aircraft. Notably, reinforcement learning combines delayed gratification with attribution. These models may replace human choice in limited circumstances, like order execution strategy, but processing speed can be a limitation in this application.

Agent-based simulations in finance are often less sophisticated than reinforcement learning models because they do not allow the agents to alter the situation, but merely react. *Online learning* allows algorithms to dynamically process new data as it is created, while *transfer learning* allows a model's knowledge to be used by another model. For instance, if a model is trained on general price movements, that learning may be used by another model and applied to a specific idiosyncratic investment. It can adapt general knowledge to the peculiarities of the specific application. Algorithms that are based on covariance, like OLS, LASSO, or Ridge, are well suited for online learning by using rank-one updates. Gaussian, gradient, and some tree methods can also be adapted to online learning.

Text-as-data is a specialized ML field and deserves a special treatment. *Natural language processing* (NLP) does not yet excel at traditional human investment analysis,[5] but it is effective at other tasks like detecting sentiment and syntax complexity in news, internet search data, regulatory filings, analyst calls, and comment letters. NLP may also be useful in predicting risk and return. Metadata and alpha-capture technology can identify authors who are influential or accurate. SVMs and RNNs are popular, but specialized techniques like bidirectional encoder representations from transformers (BERT) are dominant.

Pricing of complex and nonlinear instruments with high-bias models, like the Black-Scholes equation, can involve difficult calculations and restrictive limitations. ANNs can relax these limitations through nonparametric estimates.

Shorting is a specialized skill but provides tremendous breadth. Without shorts, an investor is limited to expressing a negative view and may be "locked" in trending markets. Simple measures like short interest, short-to-float, and days-to-cover can be used in alpha and risk modes. Shorting presents a variety of challenges:

- **Hard-to-borrow** shorts we most want to make may not be available to us or called away form us when we want them most.
- **Negative carry** imposes extra borrowing costs. Shorts must pay dividends to longs.

- **Competition for shorts** is more informed. Sophisticated professional investors are more likely to short than retail investors, with the exception of "meme" traders.
- **Interventions** by governments and events like mergers and acquisitions can favor longs.
- **Trends** tend to favor longs, such as bull markets in equities.
- **Unlimited downside** is possible for most short positions, while zero value is the maximum loss for most longs.
- **Negative convexity** in price momentum is a wrong-way risk phenomenon of shorting (as longs lose value, they become less significant allocations; as shorts lose money, they become larger positions).
- **Regulations** and accounting rules may limit or prohibit shorting.
- **Zombie companies** can cause negative carry to bleed short positions.

Risk Models

Risk models are more diverse than alpha models and typically serve a different purpose. Alpha models are generally focused on improving the predictive power and confidence of the factors, objectives, allocations, and sizing. Risk models address more varied issues. Crucially, risk models are not adversarial like alpha models—there is no electronic warfare going on, at least not in the first order.

Market risk involves general movements of markets (prices) on a macro scale (investment horizon) and micro scale (market impact). ML can help examine and forecast measurements of risk, including incorporating qualitative and unstructured data, like social media feeds. In assessing market risk, there are several key factors to consider.

Volatility is the most basic and ubiquitous measure of market risk. Volatility has the benefit of predictable characteristics, such as mean reversion, clustering, and downward momentum (leverage effect). Volatility has been studied for many years, and effective statistical and econometric models are in wide use, including stochastic models and generalized autoregressive conditional heteroskedasticity (GARCH) and its extensions. Nonlinear machine learning techniques have improved on traditional market risk forecasting techniques, both alone and in conjunction with traditional methods (such as by stacking). Applying SVMs and ANNs to pricing data, especially using sequence learning (such as LSTM), can be effective but is slow and less effective with small data sets. Mining text data with SVM applications from sources such as news feeds and social media can also work well for assessing content and sentiment. *Nonlinear estimations* of market risk for large portfolios of assets (such as for banks or pension funds) are computationally intensive and can be performed with Monte Carlo analysis. A shortcut involves using experimental design (clustering, ranking, fitting) to choose a subset of assets and scenarios, followed by models to estimate the higher-order risk parameters of the portfolio. Multiple models can be selected and combined by a meta-model.

Risk management applied to hedge and control risk using ML is an active area of research, but explainability for critical decisions like hedging can be a concern. Common alternatives to traditional methods, which may be more effective, include hierarchical clustering, penalized regressions (such as LASSO and elastic net), and ANNs (which provide added flexibility for nonlinear risk).

Regime and shock risk can be different from market risk, which may refer to more frequent and expected losses (like those measured by *value at risk,* or VaR). Regime changes and shocks are rare and involve broad and extensive interactions (contagion and systematic risk). Imbalanced data can be a significant problem when forecasting rare events like market crashes.

Asset-specific risk, including risk forecast by credit models, can require complex and idiosyncratic modeling that combines deterministic effects (such as cash flow waterfalls) and conditions (such as covenants, guarantees, and insurance) with stochastic processes. Credit risk modeling for classification, scoring, and prediction is a well-understood example of asset-specific risk and used widely in finance. Credit modeling is commonly used in areas such as distress and bankruptcy of businesses (corporate bonds); distress, bankruptcy, and prepayment by individuals (residential mortgage-backed securities [RMBS], asset-backed securities [ABS], etc.); as well as consumer spending. Traditional models have long been used, but ML models have also been effectively applied, e.g., high-variance models can process high volume data. Credit data may include trends like spending habits monitored in real time through credit card accounts, and driving habits monitored through inertia measuring devices in vehicles.

Traditional statistical and behavioral methods perform well for credit and other asset-specific risk models. They can be as simple as a strengths, weaknesses, opportunities, and threat (SWOT) analysis, weight of evidence, multiple linear regression (MLR), logistic regression, proportional hazard, LDA (and Z-score), or probit. More advanced techniques, including expert and fuzzy techniques, are also popular.

The sophistication of ML models and their ability to manage many factors and large data sets are what sets them apart from traditional models. Dynamic techniques, including online, sequential learning, and anomaly detection, are especially important for credit risks that involve changes in behavior, like fraud, and the timing of credit transitions.

Machine learning models can be problematic in consumer credit estimates due to biases and a lack of explainability, resulting in legal and regulatory risk.[6] Once a loan is made and traded in the secondary market, these concerns can become less impactful. Genetic programming (for small data sets), SVMs (which have the benefit of being resistant to overfitting and are excellent at estimating loss given default [LGD]), KNNs, and probabilistic neural nets are all commonly employed.

Liability and inventory risks include claims, underwriting, and loss reserves for insurance companies, and retirement and mortality forecasts for pension beneficiaries. Traditional methods like bivariate regressions can handle the interdependence of timing and the magnitude of events. ARIMA is also commonly applied (e.g., the Lee-Carter actuarial model). Newer methods are more capable, including agent-based models for demographics, Bayesian models, GRUs, and ANNs in general. ANNs are capable of more complex analysis than traditional methods, including managing multiple populations.

Operational risk, loosely including model risk, legal and regulatory risk, fraud, and reputation risk can also benefit from a wide variety of modern techniques. Some common methods include text analysis for document analysis and communication monitoring, binary classification, and anomaly detection.

Transaction Cost Models

Transaction cost analysis (TCA) models can deterministically analyze scheduled fees and costs, but modeling dynamic market reactions requires sophisticated analysis. For large transactions, market impact can be the dominant cost and can easily transform a profitable strategy into a loser. Two important types of TCA models are cost models and order book models.

Cost models are difficult to design because of large estimation errors, such as market noise, and due to hidden states and values, such as temporary versus permanent impact. Knowledge of an competitor’s model can help an adversary design countermeasures like spoofing techniques, which is one reason why models can be trade secrets and publicly available research is limited. Intuition and experience can help design high-bias models that use order book data or models that specify features like direction, size, liquidity, and imbalance. Even intuition and experience may miss mechanisms and interactions that are unanticipated or are too fleeting to be captured by the model. LASSO regressions can be effective. Bayesian networks can also help develop intuition from a model. ANNs are less reliable but can be effective. For market dynamics involving vast amounts of data, piecewise linear approximations can offer better tractability and speed than more complex forms. The benefit of using more complex formulas is often lost because estimation error can eclipse prediction error.

Limit order book (LOB) models are more behavioral and less economically intuitive than cost models. Complex, high-variance models appear frequently, including RNNs and CNNs, and even automated ML tools that border on data mining—data mining, like p-hacking, is generally frowned upon in the financial community—are often used in an attempt to overcome the voluminous, complex, and noisy data.

Many models can perform reasonably well for short-term LOB prediction but lack the ability to forecast reliably over longer time periods. Special modifications, like self-attenuation, are commonly used to increase performance. CNN-LSTM offers predictive results over multiple time frames.

Implementation and Rebalancing

The actual buying and selling of investments—implementation—can be simple but often requires planning and tactics to achieve the intended price. Most portfolios, including buy-and-hold portfolios, require adjustments that result in new purchases and sales due to corporate actions, subscriptions, and redemptions.

Rebalancing requires all the execution planning of an initial investment, along with added strategies (such as optimal frequency and tax management). Because execution may extend over multiple periods, *dynamic programming* is a traditional method of approaching some of these decisions (e.g., transition management) in theory, but these methods can be impractical for portfolios with many assets or over many periods.

ML methods, like Bayesian methods, are highly adaptable and capable of managing the large data sets needed for high-frequency transaction analysis. Other methods like ANNs, SVMs, and Gaussian process regressions can manage nonlinear solutions and outperform more traditional techniques. Cluster analysis and penalized regressions, like LASSO, can manage large feature sets.

Reinforcement learning is more flexible than traditional vector autoregressive (VAR) methods for both cost prediction and execution strategy design and can incorporate important nuances like hidden liquidity. Fragility and interactions between market actors are important because large orders can destroy equilibrium and start a cascade.

Integrated Models

There are some drawbacks to using different models for alpha, risk, allocation, selection, TCA, implementation, and attributions. Mixing disparate models and assumptions can create incompatible forecasts, misaligning the stages of the investment process.

Integrating the models can help ensure that interaction effects are managed properly, such as by including transaction costs during allocation. Using a top-down approach with multiple non-integrated models is like taking the intersection of the solution sets—keeping only those solutions that satisfy all models—while integrated modeling may produce a different, more accurate solution space which accounts for the overlap and interaction between models. Integration is difficult in many ways. For example, different signals may have different decay rates, making them difficult to combine.

Ordering, Selecting, and Weighting Groups

We will use three broad categories of techniques to order, select, and weight groups of holdings:

- **Cross-sectional** techniques including fixed proportion (heuristics), portfolio sorts, and simple rules like ranking by signal strength
- **Optimization,** which is a cross-sectional technique that we treated separately because of its complexity and importance
- **Time-series** techniques like risk overlays that adjust leverage based on triggers

High variance tools, like many machine learning techniques, are promising, but it is still not clear that they offer the best solutions for strategic allocation decisions. They are also useful for tactical decisions that suffer from a lack of data and the need to peer far into the future to uncover patterns that "rhyme, not repeat."

In generating categories using classification, returns are difficult to use as predictors. Noisy data and contagion allow parametric models like logistic regression to perform better than high-variance techniques like SVM, GB, or NB.

For most portfolios, optimization is the "cleanest dirty shirt"[7] of asset allocation. Optimization allows portfolio managers to determine the best portfolio weights under specified conditions and objectives. Nihilism is common in asset allocation, where many feel the futility of overcoming the limits of technology. Some of these concerns may have been justified in the past (portfolio optimization was introduced in 1952, a long time ago), but have since been addressed through improvements (such as Jorion's use of

Bayes-Stein, Black-Litterman's views, Chow's tracking error, and Michaud's resampling).

Ignorance and misuse of the methods are also rampant, including the use of historical data instead of beliefs—against Markowitz's original design. There are four major components of portfolio optimization: expected returns, expected risk, constraints, and investment horizon.

Expected returns, which are often misspecified as average historical returns, can be more creatively designed. They are frequently estimated with statistical techniques like OLS but also by using techniques such as LASSO regressions and ANNs (as we discussed earlier in the section on alpha models).

Expected risk is most frequently misspecified as historical volatility, but can be made more relevant to investor experience, e.g., by using conditional value-at-risk (CVaR) or shortfall. More sophisticated measures may require more intricate optimization techniques[8] and employ involved estimation methods like SVMs and ANNs. Risk is often easier to estimate than return. *Expected deviations* in time and cross-section are tricky. Covariances are time-varying and inconveniently distributed, requiring techniques like shrinkage to make them more appropriate for sensitive optimization techniques. Alternatives to covariances, including hierarchical trees and other clustering methods, have been proven to be effective. *Optimization* techniques are limited, and compromises must be made. The *objective function* can take several shapes. Canonically the objective function is quadratic, but may also be more complex or more difficult to solve (such as penalizing for VaR or CVaR, which require convex functions).

Constraints that seem simple can be difficult to solve and require specific solutions. Complicating constraints (such as cardinality holding constraints and nonlinear transaction cost functions) may require different techniques. These may include conic, mixed-integer, or second-order cone programming (SOCP) solutions. Full-scale optimization may seem like the ultimate solution, but it still cannot overcome difficulties (like estimation error) without great effort.

The *investment horizon* choice is a particularly difficult topic. Optimization problems are best expressed as having several decision points for goals or implementation. Single-period optimizations are ubiquitous because multiperiod optimization is difficult. Applying multiperiod optimization to dedicated portfolios with horizon matching is especially appropriate in certain situations. For instance, it can be helpful in hedging liability-driven portfolios rather than merely immunizing the duration and convexity. *Deterministic and stochastic methods* are both common, as are techniques to address uncertainty and instability, such as robust optimizations.

Machine learning can offer some solutions to optimization problems. For instance, if the goal is to find the best portfolio, why not determine that using ML instead of traditional optimizers? Traditional techniques, like mean-variance optimization (MVO), have many caveats and limitations, but creative solutions like multi-objective evolutionary algorithms (MOEAs) and ANNs have been effective in relaxing some constraints and allowing for more realistic and complex optimizations. Meta-learning can be used to combine signals, including the use of reinforcement learning. Index replication with a small number of holdings using autoencoders is an area of active research. Techniques like hierarchical clustering are also proving effective.

Ordering, Selecting, and Weighting Holdings

A number of popular strategies have proven effective for security selection. *Macro and thematic strategies* invest based on broad-based factors like economic trends and social investing scores. Like strategic asset allocation, some of these schemes rely on long-term predictions and sparse data sets. Economic regime prediction is a good example. These models can use monthly and quarterly data that are frequently revised to predict long cycles. High-bias statistical models seem to perform well and have the added advantages of economic intuition. They tend to be well understood, and offer the benefit of having good company if predictions should prove wrong (not “wrong and alone”).

A distinction should be made for academic models whose purpose is to understand and validate theories. These are often poor predictors; their purpose is insight into the mechanics of a system, not prediction. Practical biases based on intuition often perform well, especially when overfitted models cannot recognize that some changes in trend do not indicate a new regime. Predictive models predict, but they are not always designed to help the analyst understand; they do not answer a question like high bias models might. They may provide erratic but accurate results that do not convey the trend and may be brittle. To be fair, ML methods may outperform because of how they are being used, not because of what they are. As research makes progress with causal inference, including the use of counterfactuals, ML may be able to outperform statistical models “on their own turf.”

Cross-sectional strategies (some of which are *relative value* strategies), like fixed-income arbitrage, may have complex deterministic transformations and nonlinearities. These mechanisms are the “bread and butter” of arbitrageurs. Failure to apply domain knowledge thoughtfully can be catastrophic. Arbitrage is a loser’s game with many rules and subtleties.

Multi-strategy funds and *funds-of-funds* (FoFs), which are often treated with statistical methods because of the opaqueness[9] of their holdings and strategies, can be too diverse and untenable to analyze in other ways. Meta-learning can be applied to these strategies, but care must be taken to identify regime shifts like style drift. This is especially important when the data represents the performance of nonsystematic strategies. *Equity-based strategies*, like long/short funds, are similar to cross-sectional strategies, but the drivers are often less tractable and the transformations less mechanistic, though some managers use advanced techniques. *Event-driven* strategies, like merger arbitrage, require the most cognition and are less likely to be fully automated. Automation is frequently used to screen investments to a manageable number, but a nuanced and deep understanding of the event or situation is usually considered the best practice.

Simulation

Simulations are commonly used to analyze strategies and tactics, and differ from prediction. They benefit from using hypothetical or historical data to test and evaluate predictions in a safe environment before they are used for “production.” There are a great many ways to use simulation in investing. The *limit order book* data can be used to reconstruct and simulate markets to test execution strategies. Techniques like multi-agent-based simulations and random forests have been used to create realistic environments to test execution strategies. *Transaction costs*, including market impact and tax rates, are frequently simulated with high-bias models due to the noisiness and complexity of high-frequency data. Many other methods are being used with less effect. Various *execution strategies* and rules can be simulated using agent-based models and more complex methods. Agent-based models have the considerable advantage of intuitiveness, explainability, interpretability, and modularity.

Liabilities and goals are usually the domain of actuaries, but actuarial estimates can be influenced by political realities. Investment managers may wish to simulate liability drivers and effects like population transitions (such as from working to retired) and may use methods like agent-based simulations of these populations and the resulting withdrawal strategies.

Attribution, Monitoring, Feedback, and Adjustment

Chapters 18 and 19 will discuss ex post performance and risk attribution, including techniques for the classification of the sources of risk and performance, and disaggregation of risk and performance by factor and investment decision. Many investors, both retail and professional, fail to measure and attribute performance, causing adverse behavioral biases. For instance, it is common for investment advisors to surrender hard-earned gains made through selection and trade entry when they exit positions because their exit decisions are frequently driven by the desire to raise capital, rather than because of a violation of their investment thesis.[10]

Attribution is necessary when applying the scientific method to investing and informs the final stage of the investment process cycle: performance and risk management. This final stage applies skill where ideal and uses knowledge of changing environmental conditions to improve the strategy and adjust the portfolio.

Sophisticated ML techniques can be used to classify and track external managers. They can also be used to detect fraud and style drift, as well as reverse engineering strategies for analysis and replication. For strategies that are more transparent, more deterministic models like Brinson attribution are commonly used.

1. Older structural models, like Moody's credit rating model that used to impose a linear continuous constraint (like logit or probit models), have been greatly improved by more flexible models like random forest.

2. Probabilities can be extracted from trees, including decision trees and random forests.

3. For example, economic indicators can be used to create models for business cycle probabilities (a model for expansion probability, recession probability, etc.), and a classifier can be used to identify which model is closest to the current state of the economy. Dynamic Bayesian state-space models like Kalman filters, particle filtering, and hidden Markov models (HMM) are popular for regime identification. Hidden states can be inferred from observable states like volatility, rallies, or routs. Techniques like the supremum augmented Dickey-Fuller (SADF) and the generalized SADF (GSADF) have been used to try to identify bubbles (but have difficulty forecasting when they will pop). Other indicators like the Hurst exponent have been used to classify trending versus mean-reverting behavior.

4. Those who are familiar with regular expressions know it is far easier, faster to develop, and faster to execute a sequence of simple filters than to design a single, elegant, and efficient expression to perform the same task.

5. Transformer-based deep learning neural network architectures, like GPT-n models, have impressed many people but have dangerous limitations for use in investing without human oversight.

6. Some models may capture or transmit sensitive information for training. Analysts need to be mindful of how their data is used and whether it will be provided to their competitors in raw or processed form. It is also possible to identify clients, patients, and other entities that should remain anonymous through the clever processing of data.

7. Kris Kristofferson, "Sunday Mornin' Comin' Down," 1969. The phrase "cleanest dirty shirt" was re-popularized by Bill Gross when he described the dubious alure of US assets in his July 2012 PIMCO investment outlook.

8. Many advanced portfolio optimization techniques are available and easy to use; e.g., they are part of the MATLAB Financial Toolbox and the Python library pypfopt.

9. Oftentimes, fund-of-funds do not have "transparency" or "lookthrough" into the holdings of the funds they invest in. They may only offer access to the monthly returns of the funds that make up the FoF, or even the FoF itself. Multi-strategy funds that employ their portfolio managers "in house" usually have full transparency in real time. Without knowing the precise assets the funds hold, it is difficult to apply advanced analytical techniques effectively.

10. For more on this, read our discussion of heuristics in Chapter 18.

PART IV

WORKING OUR PLAN

Now that we have built our investment tools, we must operate our business. No system will work indefinitely without guidance and support. Despite promising out-of-sample backtests, most systematic strategies fail in practice.

The fragility of most systems should not be a concern; above all things except raising and keeping capital, an investment manager must be a good operator. He should be able to profit on both failed trades and good ones through the diligent application of governance and process. Cash management, risk management, and model management are essential to perform and to overcome periods of distress.

The final two chapters will lay out the key advantages and shortcomings of risk management. We will also offer a blueprint for instituting a long-term strategy aimed at avoiding costly mistakes, honing the risk management process, and systematically learning from our experiences to improve our overall results. Two of the keys to a successful long-term plan are performance and risk assessment (Chapter 18), and measurement and testing (Chapter 19).

18

Performance and Risk Measurement

Tracking Our Progress

This book is written from the perspective of a culture of risk management that is woven into a holistic investment process. Most firms are governed by a performance culture or driven by sales. These firms may not pay as much attention to a risk analyst as they might pay to an investment manager or regulator. Good luck may be centric to a compensation negotiation, but a good track record that can demonstrate skill carries a great deal of weight, even if accompanied by volatility and risk.

During critical periods, mundane concerns can supersede more important tasks; operational details can be minimized and ignored on an average day but take center stage when the traders cannot “turn on the lights.”

In this chapter we will focus on the concerns of investment managers. We characterize many of these risks as uncompensated risk premia—merely part of the “loser’s game” that punishes errors without rewarding competence. Operational risk, business risk, and other non-investment concerns encompass many areas beyond just execution: compliance; legal, regulatory, and reputational risk; external crime like cyberattacks; and internal crime like fraud. All models, not just backtesters, contain model risk. Even data feed handlers can inject a systematic bias. Unwelcome events like spikes and drops can convey information that will be missed if filtered out. We will discuss managing risk and operating an investment business in the next chapter.

In a risk-aware culture, the portfolio’s market risk should be intentional, acceptable, undiversifiable, and idiosyncratic, generally due to:

- **Equity, rates, inflation, credit,** and related asset class-specific risks
- **Investment choice,** including risk of external managers and their styles
- **Liquidity and leverage**

To begin, we *measure* and attribute performance and risk by estimating various quantities from several viewpoints. Different measures expose different aspects of the portfolio's risk and may be symmetric or asymmetric, absolute or relative, performance-based or risk-adjusted, asset-based or liability-based (cash flow), or designed to capture a salient feature like drawdowns.

Measurements are best *decomposed* into factors using data from returns, holdings, or transactions. Creating informative, explainable, interpretable, relevant, and independent factors is difficult, and some compromises must be made.

It is vital to *attribute* performance and risk to classes, factors, styles, and, critically, to decisions and decision-makers. Attribution will fine-tune the analysis and offer more detail for evaluation and opportunities for corrective action. Attribution is best performed with an inside knowledge of the decision-making process but can be performed by outsiders, e.g., using a Brinson analysis.

Evaluating performance and risk attribution will allow us to learn from, react to, and anticipate successes and failures. Evaluation may be a nuanced and difficult task, especially when attribution is imprecise.

Interrogating risk and performance with stress and sensitivity tests can be more important than the evaluation. A strategy that performs well may attract attention when evaluated, but one that is fragile should "set off alarms." Attribution and evaluation explain the current state and potential reactions to future scenarios, but stress and sensitivity tests expose vulnerabilities.

Monitoring risks, exposures, returns and their moments, and operational performance (execution, technology, etc.) is the most important of all these steps. Any study can be flawed or myopic. Vigilance is necessary to guard against catastrophe. As Ernest Hemmingway wrote, " 'How did you go bankrupt?' . . . 'Two ways. . . . Gradually and then suddenly.' "[11]

Compounding is powerful and easily ignored. In many situations, data from previous periods is unavailable, or the methods needed to model multiple periods are too cumbersome and slow. Fortunately, profits compound faster than losses, and this benefit may provide the motivation to solve the multiperiod problem. Conversely, long samples may have to be divided into decision periods for attribution.

Common workarounds for multiperiod research include using multiple evaluation periods (such as monthly and yearly), seasonal comparisons and adjustments (such as comparing end-of-year holiday seasonality from one year to the previous), and *rolling windows* such as (monthly holding periods).

Build or Buy?

A simple system that knows is better than a sophisticated system that guesses. Many excellent commercial risk software products may be less appropriate than a straightforward system designed for a specific purpose. Commercial software must accommodate a variety of clients and may make corresponding compromises to do so.

Transparency is important for an effective system. While commercial software may provide the latest and most sophisticated analysis, it may be deficient in analyzing the intents and actions of managers and strategies. Internal analysts who have access to the decision-making process can be more insightful than those who must reverse-engineer the process. Homegrown attribution systems can identify benefits, risks, and opportunities and attribute them to specific departments or employees to foster more targeted learning and improvement. Even a long-only index-aware manager's implementation relies on an investment policy, a benchmark portfolio, a target portfolio, a rebalancing frequency, and many other critical factors that may not be revealed to their investors or even to their internal department who measures their performance, or to their employer.

The rules vacuum. We continue our discussion of the case study in Box 5-5. We described a research project conducted one summer by some graduate students at Columbia University. They reverse engineered the investment policy and rebalancing rules of 101 of the largest robo-advisors in the United States.

They learned that, despite the simplicity and rule-based management of these portfolios, determining the essential parameters of their investment decisions was not easy. A five-minute honest conversation with a decision-maker would have obviated this entire exercise by revealing the true intentions of the decision makers. Several shortcomings were encountered in the research project:

> **Oversimplification** is often a threat when modeling with incomplete information. Underneath the technical complexity of many risk systems are simplistic assumptions that treat most or all assets similarly in order to permit calculations to be tractable and scalable. Some of the largest asset managers may use these systems to manage portfolios that are astronomical in breadth and complexity. As systems are simplified, they can miss the most basic intuitive relationships between drivers and outcomes.
>
> **Idiosyncratic and causal relationships** can be lost through aggregation and simplification. Arbitrages that exploit deterministic relationships are a good example of a meaningful relationship that a comprehensive risk system might miss. The cash-and-carry bond basis trade converges at expiration according to a complex relationship between bonds or notes in a “basket of deliverables” and with five options embedded in the futures contract. A general-purpose analysis that evaluates the assets held and not the basis relationship would likely miss the embedded optionality, convergence forces, expectations of changes in the underlier as it becomes expensive to the “basket,” and the potential for short squeezes.
>
> **Asymmetric and nonlinear** behaviors can be lost as well, especially for fixed-income instruments and investments with optionality. The basis may be substituted for nonlinear behavior, but its behavior is simplistic compared to many derivatives, structured products, and pooled collateral.

Many investment classes are *difficult to compare* to each other, particularly illiquid assets like private equity and real estate. Data frequency, reliability, reporting lags, hidden fees and costs, and many other irritants can make comparison between illiquid assets and liquid ones difficult. An easy, though controversial, workaround involves substituting a risk proxy or basket of liquid assets that mimic the illiquid assets. Although illiquids frequently compose a large portion of the investor's risk, adjustments and assumptions can make illiquids challenging to measure no matter what method we use.

Obsolesce. Systems inevitably have a life cycle and become *unmanageable*. Commercial off-the-shelf (COTS) software is often modified, creating a bespoke development path, or "fork," that diverges from and does not benefit from the software enhancements and updates provided by the manufacturer.

Attributing skill. The build-or-buy decision is particularly important for attribution. Measuring risk helps managers understand and react to dangers; in some cases, analyzing risk can be predictive. But knowing who and what actions contributed to that risk and comparing their risk contribution to the resulting effect is more actionable. Attributing skill is difficult and can be politically sensitive since it often influences the hiring, firing, and compensation of investment teams.

"The rearview mirror." If performance-based pay is discretionary, it is based on future performance. An employee will be paid their contractual obligation plus an incentive to remain in the company and perform during the next period. Forecasting risk is more straightforward than forecasting returns, and so the focus is often on the denominator of risk-adjusted measures when the numerator is far more salient.

"The value of the seat." Analyzing risks can expose opportunities, but for skill to be used effectively, it requires breadth and frequency. When the company's reputation, deal flow, or other benefits make profits easier to capture, an investment manager's performance may be discounted due to these tailwinds. Conversely, a highly uncorrelated and reliable return stream is often well compensated. The value of the seat must be considered when attributing skill.

“Taking one for the team.” Wrap fees, product focus, or client focus can constrain a manager and cause her to mute her returns to benefit other parties’ interests. When these incentives encourage fear of being “wrong and alone,” they can be counterproductive. Volatility control may not protect capital (unless capital falls so low that the firm cannot recover), but it protects investors and attenuates redemptions. Ideally, an investment manager should be credited with forgone opportunities that benefit his employer at his own expense. Often, “possession is nine-tenths of the law,” and these sacrifices are not always rewarded.

“A rising tide lifts all boats.” Intuition can be overrated, and many successful and intimidating managers can be replaced by a combination of smart beta, leverage, risk, and luck. It is common to assume good performance is due to skill and bad performance due to luck. An honest assessment usually reveals that a mixture of luck and skill is responsible. Nonetheless, preparation and practice can maximize luck, and that discipline and effort deserve reward.

“Head of the snake.” Luck matters greatly, and regime analysis should be considered heavily in performance attribution. An investment manager’s performance is framed by initial fortune and timing, like an optimization that starts in the wrong place and gets stuck at a local minimum. Mitigating a manager’s desire to close and restart a fund is one of the motivations behind rolling returns and resetting high-water marks.

Measurement (Statistics)

Measuring risk, performance, and persistence can be easy for an internal department that monitors a liquid portfolio, but nearly impossible for an external entity watching illiquid holdings and opaque managed funds. The variety of measures that have been invented are evidence of the difficulty in distilling a complex concept into a manageable number.

As with machine learning and statistics in general, most situations require domain knowledge and wisdom to interpret the measurements. More than 100 metrics are used for risk and performance attribution, and they are all flawed. They must be used together to form a more complete and balanced picture (a “mosaic”).

Symmetric measures, one-sided measures, and heuristics are often useful and can be appropriate for many purposes. *Symmetric measures* that consider both positive and negative values equally—such as volatility, correlation, returns, the Sharpe ratio and its variations, active share, and tracking error (TE)—are not necessarily naive in their design. A properly managed portfolio should not experience surprises, even if those surprises are windfalls—unless the strategy is designed for that purpose (for exposure to randomness with a positive reward/risk payout). An example of a strategy that is designed to capture windfalls is a venture capital fund that seeds many risky investments with the hopes that one winner (“unicorn”) will make up for the losses of all the others. For strategies that are not designed to harvest low-probability events, large profits may indicate poor management or deviations from a prescribed mandate. Happy accidents may happen from time to time, but they are no guarantee against catastrophe—more likely, even beneficial accidents indicate a propensity for ruin.

Distributions and their moments are used to measure risk and performance. These may include summary statistics, with or without adjustments like Winsorization or the Cornish-Fisher expansion. They are often calculated relative to a benchmark or to a time series, such as moving averages or Bollinger bands. If a well-behaved and tractable distribution cannot be employed effectively, robust statistics like mean absolute deviation (MAD) can be substituted for more common statistics like standard deviation (SD). For example, robustness might be needed if investment returns fail the Jarque-Bera test. Measures of deviations from well-understood and well-behaved distributions, such as the bias ratio, are also used to understand the caveats of applying a well-understood model to poorly behaved data.

Asymmetric and nonlinear measures, such as up- and down-capture, semi-deviation, the Sortino ratio, the omega ratio, and value-at-risk (VaR) measure only one side of the scale in an attempt to differentiate the good from the bad. They are more appropriate behavioral measures (losses feel more intense than gains) but are not as intuitive as they may seem. Symmetrical measures properly retain the lack of causality inherent in unintended events and statistical analysis. A windfall may be more warning of the potential for a large move (up *or* down) than of a good strategy. It may be unintuitive to redeem money from a strategy because it enjoyed unexpected profits, but allocators who ignore these warnings can be blamed for principal-agency conflicts.

Heuristic and behavioral measures take asymmetric metrics a step further; drawdown and shortfall characteristics like depth, length, and recovery time are common. Gain/loss ratios, conditional value-at-risk (CVaR), the Rachev ratio, the Calmar ratio, and the Sterling ratio are also good examples. Other less-common techniques, like the Ulcer ratio, Gini ratio, and minimax ratio, are also used.

Absolute and relative risk and return are distinctions that apply to benchmarks as well as other things. The Modigliani and Modigliani M^2 measures performance against systemic risk. Graham and Harvey and, later, Cantaluppi and Hug, measure relative to the efficient frontier. Multifactor models, such as those that are measured against Fama and French's three factors, Carhart's four factors, or macroeconomic factors, are often used to measure risk and performance, notably by Barra.

Utility-based methods are similar to heuristic methods and may incorporate risk-seeking and risk-aversion tendencies, limiting their usefulness to the population they are modeling. Many academic versions are based on various utility functions and prospect theory: Muralidhar's M^3, Kaplan's lambda, Roy's measure, and the prospect ratio. A notable exception is Morningstar's risk-adjusted return, which uses a power utility function.

Conditioning, classifying, and clustering can be used on return streams. Once returns are organized, specialized models can be applied through bagging, boosting, or stacking. These models can be effective for disparate strategies so long as the characteristics are easily identifiable (to avoid data mining and spurious relationships), such as:

- **Manager strategies,** including fixed-income arbitrage and equity long-short
- **Market conditions,** such as bull versus bear markets, volatile versus quiet markets, or periods of high or low correlation among investment opportunities (breadth)
- **Asset types and characteristics,** including stocks versus FICC, long versus short duration, good versus bad credit, and growth versus value

Holdings-based analyses focus more on the covariances of portfolio holdings than the risk and performance of the overall portfolio. Holdings analysis differs from a transaction-based analysis; it may not consider the market frictions embedded in actual performance returns, and may be incomplete and stale (e.g., relying on infrequent 13F filings).

> **Popular methods** may not be the most sophisticated or appropriate, but they hold great value by enabling effective communication. For instance, *strategy characteristics* are often used to help understand the fundamental nature of a backtest or "live" fund.
>
> **Sample parameters** describe critical features of most historical analyses. A sample period can dominate an analysis by isolating a particular regime or scenario or for evaluating an entire cycle (*through-the-cycle* analysis).
>
> **Activity** increases the breadth of opportunities and reduces alpha decay, but it also increases costs and is not beneficial without opportunities to exploit. When measuring activity, it is important not to confuse *turnover* (trading frequency) with position changes (or *bet frequency*). Both trading and positions are essential for analyzing strategies and require different dimensions of breadth. Examining investment horizons and holding periods can also help decipher if there is a mismatch between intention and action, indicating a misalignment of incentives.

“Bite-size” (amount transacted) is important but not always an effective measure, since strategies may scale poorly. Some strategies perform best when investments are small and then suffer from a lack of breadth when a manager exceeds the capital capacity of his strategy or requires liquidity to implement an illiquid strategy. These strategies, like small-cap funds with large assets under management (AUM), may be *capacity constrained* because of a short supply of attractive opportunities. It is common for investment characteristics like illiquidity and transaction costs to adversely increase with fund size rather than benefit from economies of scale. AUM statistics as a percentage of volume (POV) help explain capacity demands. Shorts are usually not netted with but are added to longs when evaluating capacity. Similar statistics applied to positions are also useful, such as minimum, maximum, and mean position size.

Holdings summary statistics are helpful to explain the constituents and types of opportunities exploited by the strategy. Holdings statistics are particularly important since external managers may not provide data for individual holdings. These may include pairwise correlation and cointegration, dispersion and diversification, correlation to a benchmark, long or short bias, funding ratio, or leverage. Sometimes information “aggregators” are trusted with holdings data and provide summary statistics for their clients to protect the privacy of the funds that provide the data. Privacy is valuable for many reasons. For instance, keeping investment holdings secret can prevent competitors from executing a short squeeze.

Performance and risk statistics are ubiquitous. Like the implied volatility calculated by the Black-Scholes equation, these measures are far from perfect but form a common language for communicating with counterparties. More sophisticated attempts can be more accurate or precise but incomparable.

Absolute performance is often top of mind when trying to understand a new strategy; an investor will check the distribution of returns and cumulative profit and loss for the portfolio. Categories like longs and shorts, and conditions like upside and downside capture, are also important. These figures should include transaction costs and are sometimes measured by liquidation value. It is common to measure absolute performance for different holding periods (one year, three years, five years, and since inception) and on a rolling basis. These measures allow investors to understand historical performance based on investment horizon.

Win/loss statistics are also standard, including hit ratio and statistics on profitable and losing trades and positions, such as average return from longs or shorts. Statistics based on intention are less common, but can measure trading ability without requiring forecasting skills (such as average return in cases where the investment thesis is wrong).

Risk statistics are many and varied, but the most salient involve both predictable business losses (such as skewness and fat tails) and unpredictable worst-case losses (long tails). In this sense, VaR measures neither a frequent nor an unpredictable risk statistic, since its primary purpose is to measure infrequent but predictable losses.[2]

Emotionally charged measures, like drawdown magnitude and length, resonate with investors and help them anticipate potential outcomes. For comparison purposes, these measures may be scaled by the rate of return, but the unscaled measurements are most arresting. As with absolute performance, it is common to measure risk over different periods and regimes (such as measuring performance "since inception") using methods such as maximum drawdown and maximum time "under water."

Risk-adjusted performance can be used to "check the box" or allocate assets, but often fails to resonate with investors. After all, we cannot "pay the rent" with risk-adjusted returns. The Sharpe ratio is one of the most common measures of risk-adjusted return. Like the Sortino ratio and similar risk formulas, it is simple to calculate and understand. Jensen's alpha and the Treynor measure are also popular despite requiring an elusively accurate and manipulatable benchmark. These statistics assume the theory underlying the capital asset pricing model (CAPM) and security market line (SML) are valid. The Fundamental Law of Asset Management (FLAM) and its components, information ratio (IR; closely related to the t-statistic), and information coefficient (IC; which measures skill), are simplistic but take the concept behind the Sharpe ratio a little further. FLAM attempts to measure skill by honing in on the active part of risk-adjusted return. While it benefits from skill, conviction, and breadth of opportunities, its simplicity necessitates unrealistic assumptions, including assigning the same standard deviation of returns for all strategies or investments. One way to reduce the effect of these assumptions is to rank order IRs and ignore their magnitudes. It is useful to monitor and compare the ex ante and ex post tracking error over several investment periods.

Efficiency measures help anticipate the unavoidable influence of fees and costs (returns are not guaranteed). Fees and costs, like slippage, can be measured explicitly relative to turnover or as a fraction of returns. Performance relative to execution without costs and fees can be a helpful number.

Uncertainty and the "Primacy of Doubt"

Some time ago, an Israeli bank changed its investment performance reporting frequency from daily to monthly for wealth clients. This resulted in a noticeable decrease in their clients' trading and improved client outcomes by discouraging clients from being distracted by uninformative volatility.

Similarly, Tim Palmer and his team embraced uncertainty in climate forecasting in a fascinating shift from deterministic "best guess" estimation to a fuzzier ensemble prediction methodology.[3] Building on the work of physicists before—employing Lorenz's insights into chaos and Feynman's primacy of doubt—he perturbed stochastic parameters and fed them into superensembles of different models. The result was a "spread" of forecasts that both enforce the mindset of uncertainty and measure the reliability of prediction (the spread-skill relationship). With this newly applied framework, he and his team were able to address uncertainty in the initial state, model risk, and the explainability and interpretability of the output. The improvement was dramatic.

As our discussion progresses from measurements to attribution, evaluation, and testing, it is instructive to recall the importance of working with probabilistic distributions rather than point estimates like population statistics. This important step helps infuse our process with the reminder that our predictions are inherently unstable and uncertain. While it is common to try to apply all the computing power we can wrangle to produce the most precise deterministic values possible, this mindset can be dangerous if we are not aware of the limitations. Performance measurement uncertainty can propagate and compound. Performance data frequently contains confounding imperfections that make deterministic evaluation difficult, including:

- **Survivorship** bias that overweights successful performance
- **Estimation error** and inaccurate proxy metrics
- **Style drift** and misleading mandates
- **Bayes error** in assuming a process or distribution
- **Classifications** that are intuitive but often turn out to be challenging, such as identifying a sector by clustering return features
- **Inconsistent and fractionated data**[4]
- **Self-reporting**, which is commonplace in performance data and can cause many problems besides survivorship bias
- **Window dressing,**[5] delayed reporting of profits, smoothing of losses, and other techniques to mislead
- **Incompatible measures** like internal rate of return (IRR) and the effect of cash flows
- **Non-coherent measurements,** most notably a lack of *subadditivity*, which can make measures such as VaR difficult to work with even without comparing them to other measures
- **Illiquid and appraised holdings,** which present many data problems
- **Structural and management regime changes,** like the management changes at Fidelity Magellan that we discussed in Chapter 2 (Figure 2-4)
- **Market structure and market regime changes,** which may invalidate investment theses, such as those that led to the failure of

Long-Term Capital Management (LTCM) in 1998

- **Capacity constraints** and other unobservable limitations
- **Infrequent, sporadic, or indeterminate delays in reporting,** which may be due to loose controls, such as self-reporting and leeway in regulatory disclosure rules
- **Opacity,** including missing or partial holdings disclosure
- **Hidden risks** such as leverage, contagion, shorts that can be "pulled," and funding that can be rescinded
- **Autocorrelation** and *multicollinearity*
- **Nonlinearity,** which invalidates many techniques

Decomposition, Attribution, and Performance Evaluation

Risk and return measurements are essential in nearly any investment process, but they become more actionable when we decompose them to in order identify their drivers and then attribute those factors to decisions. This ex post analysis can be complemented with ex ante perturbation, stress, and sensitivity studies. These methods can be used to reveal opportunities for improvement and characterize decision-makers by identifying skill and luck, danger and opportunity, and mismanagement and style drift.

Let's start with some definitions. The terms *decomposition*, *contribution*, and *attribution* are frequently interchanged, but each carries its own precise and technical meaning. For clarity, we use them in a more colloquial manner. *Contributions* refer to the effect of specific forces or decisions. A number of direct and indirect measurements and calculations are used to determine and quantify these values. Many people use this term to mean the risk or return of absolute-return (non-benchmarked) managers and assets, but we will relax that definition to allow both relative and isolated causes of risk and return.

Decomposition involves separating risk and returns in a way similar to bootstrapping risk premia, or zero-coupon rates. It may involve distinguishing between the benchmark's return and the excess return due to active choices, the effect of specific sectors or asset classes, and those due to particular holdings or factors. *Attribution* may identify the mechanism that produces risk and return, such as which risk and return contributions result from investment decisions. Decisions are frequently separated by stages in the investment process, such as asset allocation or security selection. Many use attribution to refer to the contributions of benchmarked managers and assets. Importantly, attribution involves the interactions between combinations of contributions.

Timing, holdings, and returns were used by Brinson, Hood, and Beebower[6] to decompose returns and attribute them to allocation and security selection decisions (as well as interactions and timing decisions). Active return analysis distinguishes *time-in-factor*, the amount of time a portfolio is exposed to a factor as measured by a 1/n portfolio, and *factor timing*, the decision to change factor exposures.

Timing attribution is generally assessed separately or by combining multiple observation periods over time. Most practical analysis focuses on selection, while attributing timing as a residual. Timing measures like those of Treynor, Mazuy, Henriksson, and Merton are dated and not widely used in practice. Regressions-based methods have not been effective. Variations on Merton's option-based work are commonly researched though not as frequently used. Other time series analysis approaches, such as using the Hurst index (or Hurst exponent), are actively explored but not as standardized or popular as the selection methods that occupy the bulk of our discussion.

Many computer programs analyze performance data and produce reports, but what they often cannot do is provide insights. There are some fundamental issues with most attribution methods:

- **Imprecision:** All data is in-sample.
- **Orthogonality:** Each asset must often belong to only one category.
- **Mislabeling:** Models that attribute allocation, selection, and interaction effects often consider residual performance "alpha" even when residuals only truly capture unexplained performance, not skill.
- **Linearity:** These effects are not linear in allocation weights leading to misspecification and incorrect alpha approximations.

Performance data is frequently too limited in scope, as well as in the type of attribution that can be applied. We will discuss common types of attribution, along with some of the strengths and limitations of each.

Returns-based (style) attribution dates to the original works of William F. Sharpe in 1988. Returns are poor substitutes for better information, but in many cases may be the only useful data available. Some of the challenges of using returns include *accounting treatments*, such as accrued interest and declared dividends, that can distort returns. Accounting measures are designed for a different purpose and may not be appropriate.

Reporting standards, including the appropriate use of time-weighted or money-weighted calculations, may similarly affect performance results. Illiquid and appraised valuations, which are money-weighted, are much more difficult to use than time-weighted liquid asset returns. A good rule of thumb is to use money-weighted calculations when the manager controls the flows and time-weighted measures when the flows are out of his control. *Reporting frequency* is usually out of the analysts' control and dictated by the valuation method, such as mark-to-market for liquid assets or appraisal for illiquid investments. When investment decisions such as purchases, sales, and rebalancings are out of sync with the frequency used in the analysis, the effect of investment decisions may be unexplained. When the analysis is too infrequent, performance may not be attributed when it is deserved. When the analysis is too frequent and captures market effects, attribution may be spurious.

Factor-based attribution can identify and isolate the drivers of returns for a more precise attribution of cause-and-effect. Factors often employ separate specialized models for asset classes, sectors, or geographies. For instance, a factor model may be able to separately isolate the effect of an oil shock, currency appreciation, inflation, and recession. If an investment performed well due to one of these factors, but the factor is unlikely to be important in the future, the performance should be discounted when deciding on future allocations. Similarly, a poorly performing asset may be retained if the poor performance is due to a factor that is no longer relevant.

Orthogonality is nearly always a concern when using factors to describe risk, and categories often project onto each other. Actual investments are characterized by many factors, making them difficult to map and produce results that are not subadditive. Subadditivity is convenient, but factors can be used to calculate performance metrics even without it. Subdividing classes, such as categorizing by verticals within industries, can help differentiate factors.

It is possible to use indices to represent factors and compare the performance of a market-weighted portfolio of indices to one weighted in the proportions of the investment portfolio. In addition to many indices that focus on sectors or other asset characteristics, factor mimicking indices, such as smart-beta or environmental, social, and governance (ESG) indices, can make this method straightforward. Many statistical methods can make factors more orthogonal, such as principal component analysis (PCA) and Attilio Meucci's minimum-torsion method.

Holdings-based (style) attribution is superior to returns based methods, but with caveats. For instance, *fewer months of data are required*. Among the many shortcomings of returns-based methods is the need for long histories. About 36 to 60 months of monthly return data are needed for returns-based analyses—if not more—while holdings-based studies can sometimes be more predictive than returns-based methods with only a few months' of data. While returns-based analysis lacks detail and is backward-looking, holdings-based analysis is *forward-looking*. Forecasting methods like vector autoregressions (VAR) can sometimes predict adequately but rely exclusively on past returns. Holdings-based data describes the present state in far more detail. Access to the individual investment details in the portfolio (look-through) makes it possible to understand the specific investment's factors, interactions, and nonlinearities. Even without opinions regarding the future performance of factors, modeling the individual *nonlinear and conditional* investments and their interactions is far more accurate and precise than analyzing historical data (especially for nonlinear investments involving options and structures, such as contingent cash flows).

Caveats are numerous. Holdings data are often incomplete and infrequent, such as client-reported holdings or those garnered from regulatory filings. Incomplete data may be misleading or manipulated to deceive. For instance, if long positions are reported without offsetting shorts or options, portfolio exposures can be misestimated and entirely wrong.

Even innocent aggregation and netting policies may inject unwanted biases into the analysis.

Using *transactions-based data* can provide a more precise and accurate picture by adding information about the purchase and sale price to holdings data. Transaction prices affect returns and may alter taxes. They also provide insights into investment decisions and allow more granular attribution to assess implementation.

Transaction costs and taxes can make up the bulk of the difference between what the investor realizes and what is predicted by her holdings. In addition to transactions made by the investor, managed funds may generate tax events due to their transactions. Assessing the cost of recognizing taxes earlier than necessary and not deferring them (and earning income on the tax amount during the deferral period) is more nuanced.

Identifying transaction costs and tax effects may lead to a strategic change in investment strategy, such as a change in rebalance frequency or a shift from active to passive management. Householding and tax-loss harvesting are less drastic solutions.

High-frequency attribution is more likely to measure confounding noise or execution performance than investment decisions, while lower frequencies may obscure timing effects and prevent implementation attribution. A variety of attribution methods can help refine the process:

- **Attribution by stage of investment process:** It is challenging to attribute performance to specific decisions and categories of decisions; building an in-house attribution system can be more precise and easy to operate than expensive COTS systems.
- **Mapping attributions can illuminate areas of concern**, such as non-observable factors and key-rate exposures (*risk bucketing*), and determine risk bets, trends, counterparty risk, liquidity, and forecasted cash flows.

- **Relative performance** measures such as *tracking error* (TE) and *active weight* can evaluate marginal and excess contributions.
- **Multiperiod analyses** can be built by chaining together single-period methods and can provide a more realistic and appropriate attribution;

the added complexity is averaged over long periods.[7]

- **Machine learning methods** like classification and regression trees (CARTs) can be effective in minimizing the total residual of multi-asset models and are better than Brinson models.

Benchmarks help to separate active decisions from policy mandates. There are a variety of types of benchmarks, such as those used to measure policy, discretion, or performance against commonly known but inappropriate standards. Here are some of the challenges of using benchmarks in our process.

Benevolent benchmarks: A benchmark that accurately represents a passive version of the manager's portfolio is not always convenient. Often, published benchmarks are "straw men" designed to be easily outperformed, and other times they are chosen for their ubiquity rather than their appropriateness.

Policy portfolio vs. structural bias: While the manager may be legally or procedurally bound to a particular investment policy, a commonsense or incentive-driven persistent bias may make the baseline portfolio deviate from the policy or benchmark. For instance, the portfolio objective may be "inflation plus 6 percent," but the implementation may be 60 percent domestic equities.[8] Identifying the actionable and accurate objective is significant because active decisions need to be measured against an appropriate passive unskilled alternative.

Reverse-engineering benchmarks and policy portfolios from returns, holdings, or transactions can be difficult. Even static, passive portfolios may be challenging to understand due to complications such as investor cash flows and rebalancing rules that employ tolerances. (See Box 5-5 in Chapter 5.)

Smart beta creep: As passive and low-cost alternatives become more capable of replicating strategies that were formerly only accessible through active skill-based management, the definition of skill has become narrower. Rules-based strategies can be used to create a “smart benchmark” to identify irreplicable alpha from smart beta.

Adverse weightings: Thematic funds may be exposed to unavoidable biases, and may be better matched with a benchmark with similar biases, *so long as those biases are out of a manager’s control*—such as a Sharia fund that is prohibited from owning shares of casinos or a global equity fund forced to invest heavily in an underperforming geography.

Heuristics

Performance evaluation must consider performance persistence and reproducibility, as well as the various behavioral aspects of active management, like incentives, skill, and capacity. An example was documented by Akepanidtaworn, Di Mascio, Imas, and Schmidt.[9] Due to incentives and biases, advisors systematically and destructively base their selling decisions on choices that are unrelated to investment benefits. Whether due to incentives that are realized from purchases, the attention and excitement that a new investment provides, or the desire to show value to their clients through active management, advisors’ sell decisions are often based more on the need to raise capital for new purchases (the “shiny new thing”) and significantly less on the risk-adjusted return potential of their current holdings.

Fixed Income, Currencies, and Commodities (FICC)

Fixed income, currencies, and commodities (FICC) investments require a different treatment than other asset classes due to the variety of instruments, the variations of each instrument, frequent nonlinearity, contingent cash flows, the factors, and factor interactions required to describe them.

There is significant contagion and interaction between FICC factors and those of other asset classes, but the dominance of these factors for FICC markets makes them special. Similarly, relative factors such as spreads, carry, and class-specific transformations are not exclusive to FICC markets, but they are first-order importance for FICC markets rather than being of marginal importance for other markets. Here are some of the key factors in FICC markets.

Macro factors, interest rates, and inflation are front of mind for most FICC investments. They tend to be more important than idiosyncratic characteristics like company risk. Moreover, FICC investments usually require a more nuanced analysis involving term structures, spreads, rolls, and key points (knots), and they are often dependent on non-observable measures like forward rates and par curves. Demographics and other macro factors are primary and secondary drivers of FICC returns and are far less important for many other asset classes. Because of their design, specific fixed-income sectors, such as inflation-linked instruments (linkers) or municipal bonds (munis), are more sensitive to specific factors, such as inflation or tax rates.

Implied volatility is a common factor present in many markets, but it takes center stage for the many FICC investments—particularly those with explicit, embedded, or "real" optionality and indices that are composed of them. Option-adjusted spread duration (OASD) and other more esoteric measures are the lingua franca of many of these instruments. The importance of these complex features can make weightings to coarse categories like geography and industry, or blunt measures like duration, far less important than in other assets classes.

Nonlinear exposures, such as prepayment and default risk, can be beneficial (such as the interest rate convexity of straight bonds) or adverse (as for the negative convexity of many mortgage products), requiring dynamic hedging or other costly techniques. As mentioned above, nonlinear exposures need to be modeled since historical data may not reveal their nonlinear behavior.

Dimensionality is far more significant for FICC factors than for equities due to the increased precision and determinism used to evaluate them. Maturities that can contract and extend, vintages, contingent cash flows and claims, and other instrument features increase the complexity of FICC simulations and valuation. The proliferation of complex derivatives requires the use of high-order factors such as the "five-year, five-year forward breakeven inflation rate" and swaption volatility cubes as primary factors.

Credit, liquidity, and other idiosyncratic risks are more pronounced in some fixed-income markets than many other asset classes. The limited lifetime of most products, call features, and stripping can make individual investments more perilous than most investments in other markets. Specialized methods such as duration times spread (DTS) and optimal risk budgeting with skill (ORBS), both pioneered by Barclays, can help compensate for the idiosyncrasies of fixed-income attribution.

Benchmarks face many of the same challenges for FICC investments as they do for the other applications. Index construction may involve adverse rules, such as weighting by issuance. A company that issues substantial debt may be heavily weighted in an index because of its prolific issuance. That same high debt burden may make the company a poor credit risk, diminishing the portfolio's value.

Adverse selection effects like this entice active managers to deviate from "core" portfolios that hug the index and to create "core plus" products. Since these deviations may be commonsense adjustments, core-plus benchmarks may be more appropriate than the index itself. Even when replicating the core portfolio, the number of investments required may be impractical. In these cases, approximations like stratified sampling or tracking error minimizing optimizations may be used.

Fixed-income benchmarks can be complex, and their fundamental characteristics, such as duration, may be time-varying. Fixed-income indices may have a considerable number of constituent holdings, and many holdings may be illiquid, esoteric, or otherwise difficult to own. For these reasons, benchmarking fixed-income portfolios can be more complex and nuanced than for other asset classes. Many of these techniques are used in other asset classes but are particularly important in fixed income.

Custom benchmarks can be designed, and the performance reporting of these bespoke indices can be outsourced to minimize agency risk. Custom equity indices are often just compositions of more targeted indices to better match the active portfolio's characteristics like geographic, style, or sector exposures. Custom fixed-income benchmarks may require exposures that are not represented in existing products. Derivative instruments such as swaps can be used to create synthetic exposures that are better suited to the investor's mandate.

An example of a portfolio that may deviate from common benchmarks in a passive but meaningful way involves the tax treatment of loans that are held for investment (HFI) and reported on the balance sheet on an amortized cost basis. Unlike traded loans, or held for sale (HFS; also known as available for sale, or AFS) portfolios, HFI portfolios are measured by accounting rules and may be benchmarked similarly.

Bucketing can "neutralize" biases. For instance, a thematic ESG fund may maintain the same sector weights of a broader index while selecting investments with the highest ESG ratings within each sector. Measuring risk and discretizing key rates into duration buckets is often used to manage first-order exposure to the shape of the yield curve, but cannot address higher-order relationships and risks that are not exclusively related to interest rates, like credit risk, unless they are controlled explicitly.

Capping (limiting) exposures of a particular characteristic or concentration is common. Equity indices often have a dominant asset whose proportion is untenable for investors. For instance, not long ago, Naspers consumed nearly one-quarter of the South African JSE Top 40 Index.[10] For fixed-income portfolios, caps on issuer concentrations and credit ratings may help the problem of issuer weightings, though these will likely create other biases. It is common to bag or boost capping models, such as equity caps that involve asset classes and sectors or fixed-income models that involve duration and credit rating. Credit risk is can be viewed in this way, with central banks or other regulatory oversight considered as a backstop, followed by business risk, capital reserves and profits, quality (including conservative underwriting), liquidity, and access to capital.

Immunizing benchmarks against drifts in maturity, duration, composition, or other characteristics can mitigate the changing exposures of bond indices. Many features and nonlinearities in fixed-income investments cause risks to change and are expensive to hedge. From 2012 to 2021, the duration of the Bloomberg US Aggregate Bond Index (Agg) drifted from about three and a half years to nearly six and three-quarters years. Using a more stable benchmark may be more appropriate for the investor's mandate, such as matching the liabilities of an insurance company. Tolerance bands are also typically used to exploit situations like the fallen angel anomaly. A manager may hold onto (or purchase) debt whose credit rating has fallen below investment grade while less-flexible investors may be forced to sell the same debt merely because the rating dropped by one increment.

Consequences for multi-asset models. Many multi-asset models use statistical measures such as correlation and cointegration to analyze and simulate portfolios for stress tests, scenario analyses, and sensitivity analysis. The nonlinearity of many FICC instruments and the dramatic evolution of FICC markets make historical assumptions more specious and unacceptably simplistic than for most other markets. Complex and burdensome calculations are required for many products when extrapolation is inaccurate.

Stress, Scenario, and Sensitivity

Up until this point, this chapter discussed ex post analyses. Now we will discuss ex ante "what if" investigations to uncover hidden risks and opportunities from an investor's perspective. Risk management and regulatory analyses, including regulatory capital tests, are far more comprehensive and prescribed than what we are discussing here to assess and manage portfolios by "front-office" investors and researchers.

Perturbation and sensitivity analysis probe a model for its response to input changes. A simple example might be increasing and decreasing a model's input values, such as key rates, one at a time and in combinations, and measuring the change in the output. Even a simple exercise like this can provide many benefits:

- **Uncertainty,** such as the effect of estimation error, can be assessed, and the need for precise and accurate estimates can be determined.

- **Parameter estimation** sensitivity can be quantified, allowing analysts to budget their resources appropriately and focus on forecasting only those inputs that significantly affect the model output.[11]
- **Model validation and stability** relative to those inputs can be determined with perturbation and sensitivity analysis.[12]

Stress and shock testing are used to determine both the fragility of the model and the limits of the relationship the model is predicting. Some of the weaknesses—and insights—this testing may uncover include:

- **Exposing model weaknesses:** Most models are only effective over a range of inputs and fail, sometimes spectacularly, when extrapolated beyond their operating range.
- **Causal dependency** may be more robust than statistical tendencies, which can break down under various circumstances. Causal models and structural designs can better align predictive processes with the reality that they are analogizing.
- **Revealing valuable insights:** If the model is prodded and tortured without succumbing, then we may have discovered a “pearl of wisdom” in the unexpected response it provides.
- **Data generation** can use pseudorandom inputs drawn from distributions or some other synthetic method; when these numbers have an economic purpose, such as simulating a specific event (from the past or hypothesized), they form a scenario.

The distinction between stress, shock, and scenario analyses is mostly in the magnitude of the input and purpose of the analysis. *Stress tests* generally imply an attempt to break the model or understand how the correctly modeled behavior performs under plausible circumstances. These events may evolve rapidly or slowly; they may cascade and form a complex web of interactions; they may also be shocking. *Shocks* usually refer to impulsive inputs. They may result from extreme stress scenarios, such as a “flash crash” or “taper tantrum,” or to large perturbations like currency devaluations and downgrades, such as when Russia invaded Ukraine in 2022.

Scenarios analysis is more general than stress testing and shock testing. Scenarios may be synthetic and unrealistic but imply an attempt to model the results of a conceivable event. Common scenario analysis methods include a *threat matrix*, or *tactical outlook* (recall Figure 2-3 in Chapter 2), which can be created by exploring various potential scenarios. Frequently, probabilities are assigned to a range of possible future circumstances. In creating these matrices, a manager can plan for foreseeable circumstances. She can use them to decide how the fund should respond to future events or to anticipate those possibilities and hedge. A threat matrix and scenario matrix are particularly helpful when the outlook is contingent on an anticipated decision point with an uncertain outcome, such as an election or the Brexit vote. *Reverse scenario analysis* is used to identify the assumptions necessary to achieve a response. Unlike "p-hacking," the goal should not be to find inputs that validate a "pet theory" but to identify what conditions are required to produce the response. Once identified, a scenario may be deemed likely or ruled out. An intuitive scenario may be invalidated by an input determined to be unrealistic using this method. Reverse scenario analysis is similar to backward propagation or determining the views implied by a portfolio using the Black-Litterman model. *Simulations* using those scenarios can be performed using a backtester to generate cash flows and model portfolio performance (see Box 1-4 in Chapter 1).

Distributions of predictors used to generate distributions of responses are far better than using point estimates because distributions deliver a much clearer description of potential outcomes and their characteristics. Distributions can be determined by theory or by inspection, as shown in Figure 5-3 in Chapter 5.

Interactions and regime shifts are common when simulating interesting scenarios. An analyst must be vigilant when selecting and generating data for these simulations. "Contagion" occurs when uncorrelated returns become correlated, often during extreme situations.

Scenario generation can be accomplished many ways. *Economic and forecasted scenarios* (such as capital markets assumptions) can be the basis for generating scenarios for simulation, but it is often impractical to speculate on all the parameters that the simulation requires Some combination of the following techniques is usually required.

Historical returns scenarios can be generated by using asset return streams from the past to generate a backtest of a particular period, such as a crash. The ease of using this method makes it popular, but it is limited.

Historical factors are much better but more difficult to use. If the factors that formed the past events are determined, they can be applied to current conditions. For instance, if a crash happened when interest rates were high, and we would like to simulate how a similar crash might affect our portfolio, our simulation may need to be adjusted for today's current interest rates. This added abstraction and bias can make our simulation more applicable for estimating our potential future but may make the simulation less explainable or believable. The argument "historical data is unrealistic because rates are now different" may be countered by "historically inspired factors are unrealistic because they are unobservable, and I don't trust them."

Simulated scenarios can be based on historical factors and returns or entirely fabricated. Many methods are used in practice, and new methods are invented frequently. Simulations allow for a more scientific and hypothetical experiment with many trials, but care must be taken when creating synthetic data.

Parametric simulation methods use well-understood and well-behaved distributions to model values like returns. They can significantly improve the speed of calculations, especially when many holdings or simulations are required. When predicting future events that will not follow past experience with precision, biased parametric models can be superior but sensitive to underlying assumptions. Some commonly used distributions include the normal (Gaussian) distribution, the extreme value distribution (EVD) to simulate large values, the Pareto distribution for tails, and the Weibull (Rayleigh) distribution. Each has its bias and theoretical foundation. *Copulas* can be used to model joint distributions.

Nonparametric methods exploit the variance part of the bias-variance trade-off and fit models to a sample but may overfit. Nonparametric methods have the advantage of employing more complex and realistic distributions.

Monte Carlo simulations are convenient, easily implemented, and often "embarrassingly" parallelizable. Though they use many simulated trials, they can usually be performed efficiently. The relationship between return streams or factors is a critical and challenging process. Cholesky decomposition is a quick and convenient method for simulating price or rate evolutions using Brownian motion by randomly drawing from a normal distribution while preserving the correlations between the simulated holdings. One variation on this method is the *exact* method, which produces realizations with the same mean and covariances as the inputs and produces similar terminal values. Another variation is the *expected* method that uses a statistically equivalent mean and covariances and produces terminal values that are not similar. *Cointegration* relaxes correlation's assumption that asset prices move in lockstep at each increment, but rather evolve in the same direction over time and may diverge over smaller time increments. Tests like the *Engle-Granger* test are used to identify these relationships. *Machine learning* techniques like time series GAN are promising and evolving.

Complexities and inconveniences are sometimes confounding, but more often provide opportunities for innovation and improvement. Some of the challenges in scenario generation involve multiple interacting events and evaluation periods, time lags, cascades and feedback, interactions that only present themselves after a threshold is breached, and regime and phase shifts. These difficulties are often not theoretical, but computational—like the dynamic programming solutions to multiperiod optimal rebalancing that we discussed in Chapter 12.

■ ■ ■

Measuring and attributing cause and effect, skill and luck are necessary before we can punish, reward, expand, or improve our processes. A genuine pursuit of excellence requires a commitment to understanding the origin of decisions, the weaknesses and strengths of processes, and their effects. Decision makers hide their true reasons from others and from themselves. Understanding how and why they perform is difficult but essential to building a successful and scalable process or approach that adds long-term value and even alpha.

---

1. Ernest Hemingway, *The Sun Also Rises*, 1926.
2. CVaR is a slightly better choice since it measures the center of mass in the tail.

3. Tim Palmer, "The Primacy of Doubt," *Journal of Advances in Modeling Earth Systems 9* (2), 730–734, April 20, 2017.

4. A 2009 paper by Agarwal, Fos, and Jiang reported that HFR, Eureka, MSCI, TASS, and CISDM indexes barely overlap and contain less than 1 percent of each other's data.

5. Window dressing involves non-economic transactions designed to "clean up" the balance sheet or performance records to improve the "optics" of official filings or statements.

6. Gary P. Brinson, L. Randolph Hood, and Gilbert L. Beebower, "Determinants of Portfolio Performance," *Financial Analysts Journal* 42, no. 4 (July-August 1986), 39–44.

7. Other techniques involving backward propagation or continuous variables can be employed but are rarely used in practice.

8. In this example, the returns of the "CPI+6" portfolio may be similar to the "60/40" portfolio at times, but the composition is drastically different. CPI+6 implies steady returns adjusted for inflation (which increases consistently, not to be confused with the change in CPI that is often reported), while the 60/40 is a risky portfolio whose returns vary greatly.

9. Akepanidtaworn, Di Mascio, Imas, and Schmidt, *Selling Fast and Buying Slow: Heuristics and Trading Performance of Institutional Investors*, NBER Working Paper No. 29076, July 26, 2021, revised November 18, 2021.

10. A strategy that used the JSE Top 40 as a benchmark but also had exposure caps would suffer from a large tracking error due to the outperformance or underperformance of Naspers versus a capped benchmark.

11. *Overindexed* data refers to parts of the data set that produce a stronger signal than the overall data set. Sensitive models produce a magnified response to changes in signals—even ones that are not overindexed.

22. A model can be *unstable* and respond violently and magnify small changes in inputs, *metastable* and respond to only large changes but perhaps in a brittle and catastrophic way, or *neutrally stable* and respond to changes without magnification.

19

Investment, Risk, and Cash Management

Improving Our Process

Strategic plans are fragile, but that does not diminish the value of planning. A plan's virtue is not its face value, but rather the license it gives us to be decisive and effective when confronted with opposition. Whether we encounter a structured and directed competitive attack or simply the incomprehensible chaos of market forces, a well-conceived plan can help us navigate the choppy waters more confidently and competently.

Plan for chaos and work that plan. Proper day-to-day risk management of ordinary risks should permit a business to operate smoothly under routine conditions. If normal risks are controlled habitually, extraordinary threats to capital and endurance can be addressed quickly, objectively, and effectively by executing predetermined plans for assessment and action ("train like you fight, fight like you train"[1]). As Daniel Kahneman wrote, even simple rules can be better than expert judgments, especially in emotional and critical situations.

Here are some fundamental risk management concepts that can help guide us from initial plan through ongoing execution.

> **First, be adaptive and aggressive.** Just as those who deride all active management with an argument based on averages, critics of planning often attack the "straw man" of static and reactive management. A poorly performing investment strategy often results from not being systematic and rational enough—strategies fail when they lapse into arbitrary judgment, overtrading, or poor prediction. "Your first loss is your best loss."

Behavioral biases are inevitable. Systematic traders might not think subjective management is effective, but doubting the robustness of a strategy is inevitable during periods of distress. It may be appropriate to “pull the plug” if the model is outside of its operating range or otherwise misbehaving.

Risk control is the primary “dial.” Good investors can profit from bad situations, and bad traders can “snatch defeat from the jaws of victory.” Increasing terminal wealth and avoiding ruin are the critical outcomes of most investment programs. Keeping what we earn is a crucial factor for success. Execution, risk management, and money management are often more important than accurate analysis and decisive investing. Notably, an emphasis on risk rather than potential profit helps control greed and euphoria. Risk is also easier to estimate than future returns, so it is a more practical goal.

Risk is meant to be spent, not minimized. Risk is best viewed as a currency whose highest use is sustainable prosperity. Just as fees and costs can be well spent or carelessly squandered, risk must be spent judiciously to produce and retain long-term gains. An investment manager does not minimize risk but instead maximizes risk-adjusted returns.

Uncompensated or unintentional risk is wasted risk. Risk should be removed if the cost of removing it is acceptable. Other risks must be weighed against their costs, but also against the potential benefit of spending part of the risk budget on them. Self-insuring, hedging, or removing compensated risk should be based on predictability and skill, house views versus priced-in market expectations, risk tolerance, and cost.

Here are some guidelines for which risks are compensated and which are not:

- **Predictable exogenous risks** are usually compensated with a risk premium, including realized volatility, credit risk, and inflation.
- **Unpredictable exogenous risks** are not compensated (e.g., certain currencies, and catastrophes like natural disasters, war, and famine).

Some measure of unpredictability (such as implied volatility) may be compensated, but long or fat tails may not be.

- **Bounded endogenous risks** are compensated, such as illiquidity of investments, or external manager constraints and implied performance.
- **Unbounded endogenous risks** are not well compensated, such as style drift, breaches of constraints and covenants, and *key-man risk*.
- **Diversifiable and operational risks** are usually not compensated, including concentrations in categories or high-conviction trades, leverage, having too few prime brokers, counterparty risk, fraud, errors, poor infrastructure, and bad governance.

Everyday risk and risk of ruin. We should reveal which risks are knowable and which are uncertain, so that we can address those we know and understand the limitations of our efforts. This will help ensure that the risk budget draws from revenue and not capital, and minimizes the chance that the retained risk poses an existential threat to the continuous and effective operation of the business.

Reporting risk is necessary but not sufficient. Reporting serves many purposes for stakeholders. Our focus is on investing, so the reporting needed to manage risk can be categorized functionally:

- **Required**, such as regulatory reporting and counterparty credit, which includes capital adequacy and affects reserve requirements and collateral management
- **Procedural**, such as dashboards, warnings, and key risk indicators (KRIs)
- **Managerial**, such as action items and managerial reports
- **Portfolio**, such as client reports
- **Specific**, such as individual securities and factors that are deemed a risk. An example of specific risk includes investments whose revenue relies on assets rather than cash flow from operations, such as project finance.

A framework to manage risk. Throughout this book, we have laid out a framework for identifying and managing risk through a holistic process.

Governance, *leadership*, and a *culture of risk* represent the foundation of an effective framework. While these processes cannot replace sound judgment, they can help enforce thoroughness, discipline, and objectivity. *Data management*, *model design*, *research*, and *testing* methods are used to minimize biases and errors. The advantage of quantitative investing is the ability to do this consistently and well.

Asset allocation should diversify away unwanted risks (like systemic risk) and impose constraints to prevent excesses (like concentrated investments, undesired biases, and other risks). Many asset allocation decisions (including global macroeconomic theses) spend much of the risk budget on asset allocation. *Security selection*, including due diligence and “deep dives,” is another activity where the risk budget would be well spent. Idiosyncratic, undiversifiable selection is a good use of risk.

Periodic monitoring and *rebalancing* should be an integral part of managing the fund, rather than done ad hoc. It should be studied and systematized. *Constant monitoring*, to attribute performance and evaluate the effective use of risk, is best overseen by people other than the investment team (*separation of duties*). Compartmentalization can reduce bias, principle-agent problems, and conflicts of interest.

Portfolio risk management is the “teeth” behind the monitoring effort and should involve objective and unmistakable processes (including prompt flags, limits, and triggers). Risk processes should be set by the investment or governance committee to mitigate, control, and direct the business’s risk without reliance on externalities. *Learning* to analyze, attribute, improve performance, and maximize objectives (e.g., performance, risk, and constraints) are the true outcome we desire. Some treat this exercise as a mundane reporting function and forgo most of the value. “Going through the motions” will not achieve the goal of protecting the portfolio, the clients, and the fund.

Governance and Process

Many legendary investors emphasize risk management and discipline over profits and optimism. Windfalls are rare, but ruin is not.[2] Greed, performance incentives, fund structure, short-termism, and behavioral biases contribute to an environment that makes proper risk-aware investing one of the most challenging and rewarding careers. Many investing titans (including Ray Dalio and Paul Tudor Jones) have created a culture of risk management and introspection; it is their hallmark. Pivotal issues of governance and risk management can trip up firms, while helping others excel.

“A seat at the table.” In a sales or product oriented culture, there is sometimes a divide between sales and products staff, who garner fees and most directly generate revenue for their company, and investment managers (and risk managers), whose risk adjusted returns are the building blocks of products and are a step removed from revenue and influence. In addition, investment managers and salesmen translate intentions directly into profitable actions, giving them political clout, but risk managers often must rely on weaker hypothetical narratives to justify their arguments. Political expediency can overwhelm prudent management “just this one time.”

Agency problems. Risk concerns often lack impact because of the urgency of social and business goals.[3] Performance incentives play a significant role in diminishing the influence of risk management.

Risk management, though often not urgent, protects the business from long-term and theoretical events. When salient needs confront contestable concerns, they create agency problems. Powerful and successful investment managers—along with partners, advisors, and sales staff—can sometimes defang prudent risk efforts. Zillow’s spectacular failure in trading houses in 2021–2022 is one of many examples of business goals overriding prudent risk management.

Skill versus luck. Well-meaning investment managers and executives are often tempted to bend the rules to achieve short-term success or “gamble their way out of a hole.” Risk-managed incentive schemes diminish the role of luck and breadth, excitement, and windfall. Understandably, a manager who is incentivized to produce short-term profits—a bumpy pursuit for which occasional bouts of bad luck are expected and forgiven—would want to maximize her exposure to opportunity in hopes of exploiting good luck. If she loses her investor’s money, she can find a new job while retaining most of her past reward (a “real” option and a perverse incentive).[4]

Many executives and investment managers have "rolled the dice" to escape severe or existential losses. A desire to "break even" has been the origin of many rogue trading conspiracies and Ponzi schemes, resulting in spectacular bankruptcies and incarceration. The fall of Barings Bank in 1995 is one of many examples of poor governance and processes. The difference between aggressive, calculated risk-taking and irresponsible gambling may need to be confronted at the worst times. Governance and process exist to manage difficult times when systematic methods are questionable.

Product success versus investment excellence. Product success can be toxic to company success (*reputational risk*). As we have discussed previously, VXX[5] is a wildly successful product and a terrible long-term investment (Figure 4-2 in Chapter 4). The prioritization of product and sales goals may distract well-meaning companies from their pursuit of investment stewardship. Sustainable success as a prudent manager requires a determined and focused commitment to performance.

Sisyphus and compartmentalization. Action, not measurement, is the purpose of risk management. Many risk managers effectively implement risk management, especially those in highly regulated businesses (like banks and insurance companies). Many more are frustrated when they are overruled in critical situations. They may feel that their advice may prevent catastrophe and may not understand the larger business goals. Communication and inclusiveness can foster teamwork and alignment; "mushroom"[6] management does not.

The arc of responsibility and "playing the long game." It is essential to empower skilled and attentive professionals to own the consequences of the risk they manage. When decisions become contentious, profit-centric arguments like

> For an EM fund like ours, divesting from Russian and China will weaken our core narrative and immediately reduce sales by 20 percent.

can be more impactful than responsible counterarguments like this:

> There is a 65 percent chance that our returns may drop 6.5 percent if Russia invades Ukraine (a highly uncertain but foreseeable scenario), translating into client redemptions (but it is not easy to know by how much).

The proper incentives and culture to support prudent actions that conflict with more immediate business goals were discussed throughout Part I and are the solutions to this challenge.

Bridging the gap—a culture of risk and profitability. Commitment to cause, empowerment, and taking personal ownership of the business's success cannot guarantee effectiveness. Conflicting motivations and incentives will make healthy confrontations and disagreements between investment managers and determined risk managers inevitable. It is critical to socialize the true purpose of risk management—to make the business sustainable and productive. Governance through enlightenment is more effective and pleasant than by force and coercion.

A Culture of Responsibility

Building a responsible culture, or transforming a culture to be more responsible, is as challenging as it is important. Most will agree with the need for a good culture, but winning hearts is more difficult than convincing minds. Profit incentives often conflict with existential risks to the company, and to the welfare of clients. Sometimes hearts and minds are unattainable.

Process. A sustainable and effective investment process requires:

- **Verified and audited** research performance to prevent spurious or irreproducible results
- **Insightful and actionable oversigh**t of investment quality at the portfolio and company level
- **Actionable recommendations** for improvement

Implementing this process involves:

- **A structure** of investment tasks to provide a framework for reproducible excellence
- **Key performance indicators** (KPIs) to provide actionable goals and to allow for monitoring
- **Expected outcomes** as milestones, motivation, and progress indicators
- **Performance and risk measurements** to provide feedback
- **Attribution** and assessments to determine the causes of performance and underperformance, and to indicate actions to reinforce or remedy those causes
- **Risk warnings** and triggers to quickly identify problems and provide a clear escalation path
- **Plans for mitigation** and correction to show a clear path forward for action

Successful outcomes include:

- **Investment performance** that is measurable, attributable, and improvable in a scalable way
- **Sustainability** through difficult times and consistency to grow as new and distracting products are built
- **Scalability** in terms of new clients and target companies, as well as new and divergent business lines
- **Accountability** for the companywide impact of intense product-level efforts
- **Perspective** to serve as the "tenth man"[7] for discouraging groupthink

Cultural reform requires:

- **Identifying** stakeholder goals and company investment goals to establish a shared vision and an alignment of targets
- **Ensuring "buy-in" and alignment** of collective interests across the organization to ensure investment research and management excellence, and long-term success for the firm while simultaneously addressing the need for timely solutions required by the business units

- **Leveraging investment practices** across silos to form a companywide and collaborative investment process
- **Delivering long-term risk-adjusted performance** and sustainability, in addition to short-term profit-driven client outcomes
- **Building a meticulous** companywide research and investment process based on best practices

Money and Crisis Management

We will discuss the operation of a robust and effective investment process in response to uncertainty and outside shock. Events like the suspension of the London Stock Exchange's nickel trading and the cancellation of executed and hedged trades during the 2022 invasion of Ukraine are unpredictable, but disruptive events occur frequently enough that any successful strategy will encounter them many times.

Here are some of the qualities that make for a sound investment process:

- **Skill.** "Never willingly accept a fair fight." It is essential to exploit competitive advantages and systematically avoid situations that do not offer a substantial edge.
- **Breadth.** We can increase our chances of success by exposing niche strategies to as many high-confidence opportunities as possible, such as by using multiple investment vehicles chosen from a broad universe rather than just common and liquid ones.
- **Discipline.** Many discussions stop with systematic planning and forecasting, but sales, raising capital, and limiting redemptions are usually more important to profitability and continuity.[8]
- **Preparation.** Systematic risk management (including leverage, hedging, risk transfer, and risk triggers) allow a fund to operate in an anticipated way, even during crises, to preserve "mental capital" and prevent confusion and "paralysis." Risk triggers can define events and the subsequent action items—"a way forward."
- **Decisiveness.** Risk triggers should be elaborated with action plans to add structure during chaos (a "playbook"). The investment

committee should evaluate the investment process and take rapid and deliberate measures to continue, correct, or terminate a systematic strategy during a crisis. Prior to the crisis, they should be prepared with a thorough understanding of the strategy's assumptions, operating ranges, strengths, and weaknesses, including evaluating and accepting the biases embedded in the process.[9]

Systematic Risk Management

Understandably, systematic investors are inclined to reduce risk management to a mechanical exercise, but good planning may not be adaptive enough to survive "contact with the enemy."[10] Each systematic risk management strategy implies a belief in market behavior. Dollar-cost averaging, rebalancing strategies, martingale systems, variations on the Kelly criterion, and variations on Markowitz optimization imply strong views on how the market will work in the future and are likely to be right or wrong depending on various conditions, such as regime changes.

Systems may have exposure to market behavior that is not obvious, but willful ignorance is a poor decision and does not absolve a manager from the responsibility of choosing. Let's take a quick look at some of the most common risk management strategies:

- **Diversification** relies on assumptions and forecasts of market structure, including dynamic and conditional correlations between assets, and the relevance of the constraints and objectives that were employed (such as everyday volatility versus CVaR or "tail risk").
- **Symmetric and asymmetric** strategies make far-reaching assumptions about characteristics like stability, operating ranges, and linearity.
- **Mean reversion** can motivate concave strategies such as constant mix schemes that rebalance with the intention of buying low and selling high.
- **Trend regime,** momentum, or convex strategies and tactics, include techniques like constant proportion portfolio insurance (CPPI) or stop-losses.

- **Time-invariant** methods minimize the importance of timing through such strategies as through-the-cycle techniques, volatility targeting, and time-invariant clustering.
- **Timing** is notoriously difficult in most applications other than arbitrage techniques. The failure of drawdown control strategies, including systematic deleveraging (sizing), are commonplace due to overfitting and other pitfalls.

Methods for controlling risk should be integral and incorporated at every stage of the investment process. *Opportunity choice and product design* should balance the need for liquidity, sustainability, and breadth with investor outcomes (like profitability, variance, and the capacity to scale the investment strategy without "reaching" for low-confidence opportunities).

For example, structural complexities (such as mimicking commodities exposure or stable values using dynamic strategies and financial engineering) can create exciting and popular products. However, since the similarity between the intended exposures and their actual mechanizations are only "skin deep," unsophisticated audiences may misuse these products, and the "blowback" may result in reputational damage. The *process design and problem specification* should be thoughtful about their goals and the implementation of them. A common fatal error is answering the wrong question.

The value of leadership and governance may involve rules and organization, but their value is in "winning hearts and minds" to promote compliance and allow improvisation that embodies the spirit of investment excellence when specific examples are not available. *Research and development* involve more than predicting profits, even on a risk-adjusted basis. Features, predictors, responses, and objectives should be designed, interrogated, and validated with vigilance and earnestness. *Strategy design, backtesting, sensitivity, shock, and scenario analysis* should focus on the risk component of investing, including model risk. *Execution and implementation* (such as rebalancing, scaling, and timing trades) should be designed, backtested, verified, and performed with appropriate risk-taking in mind and proper incentives to prevent deviation or principal-agent problems.

Mitigation techniques (including hedging, risk-management overlays, and risk triggers) should be thoroughly vetted for unintended consequences.

It is common for these methods to have the opposite effect of their intention. As an example, a delevering overlay may reduce market exposure with a lag, selling at the bottom, relevering with a lag, and buying after the recovery. Properly designed mitigation strategies should limit adverse selection (such as selling good liquid investments to "prop up" poorly performing illiquid ones).

Communication up, down, and across the business and with external stakeholders, like prime brokerage, investors, and parent companies can avoid existential problems. Avoiding communication and "circling the wagons" may be a temporary fix that may prevent problems with the help of luck, but honest and effective communication can allow us to survive if things worsen. Rogue trading and Ponzi schemes can recover small losses with gambles or result in jail time and permanent loss of income. Restructuring a debt proactively is a better strategy than hoping to win the lottery and defaulting instead.

Hedging and risk transfer can be deployed to manage risk. Usually, the cost of these methods exceeds their economic value but may provide utility. Outsourcing distracting risks and potential operational errors can allow an investment team to focus on its core competencies. Throughout this book, we have discussed selection-level and portfolio-level hedging and some execution and implementation hedges. *Risk transfer* agreements can outsource portfolio risks in whole or in part, including beneficiary liabilities for a pension, capital requirements for a bank, or future project funding for an endowment.

Some pension plan–level contingent claims can be transferred, but other considerations are specific to the funding commitments of these plans. The *funded status* (the critical risk of a plan itself) may drop for many reasons, including asset values not appreciating as liabilities grow; liabilities may be adjusted for inflation, shortfall, and contributions. Liabilities can be guaranteed explicitly or conditionally. The *shortfall* (the risk of an employer failing to keep the plan funded) must be managed. And *contributions* can vary with time, demographic, and employee status changes (such as from contributing to the plan to retirement and drawing cash). The plan's actuarial assumptions regarding contributions may assume a steady flow. This gap in expected and realized contributions must be managed.

Box 19-1 Risk Transfer

Risk transfer does not always work out as expected, such as Credit Suisse Group AG's 2011 Partner Asset Facility (FAF2) operational risk note or the World Bank (International Bank for Reconstruction and Development) Pandemic Emergency Financing Facility (PEF) pandemic bonds, which did not trigger during the Ebola outbreak in the Democratic Republic of Congo in 2019 but did during the COVID-19 pandemic in 2020.

Overlays are investments that adjust exposures in addition to the portfolio design without altering the original holdings. They can be used to hedge the overall portfolio, asset classes, sectors, or even individual investment risks like prepayment or default risk. They are especially useful for management at the treasury or corporate level.

More specialized investment managers can be left to their competency unfettered, and the chief investment office (CIO) or treasury can overlay a hedge so that the institution will benefit from the manager's skill while not being exposed to the broader risk of his discipline. For instance, a technology stock portfolio manager may buy and sell tech stocks, and his CIO may hedge out technology sector exposure so the company will only be exposed to the performance of the technology stocks that were purchased *less* the performance of the technology sector overall—in other words, market neutral.

Overlays can be affected in several ways. *Liquid overlays* rely on liquid and inexpensive scalable investments like index funds, options, futures, and forwards. *Precision overlays* employ more targeted but illiquid and expensive solutions including exotic options and bespoke swaps. *Gap risk* can be filled by temporarily purchasing or selling liquid assets to provide the exposure intended for a future, more permanent, illiquid investment or sale. As the desired exposure is realized, the temporary asset exposure is removed (transition management).

Integral hedges are part of the portfolio strategy or design process and are most likely managed by the individual portfolio manager or trader. They may include derivatives or systematic strategies but are not "bolted on" to a portfolio's design. *Basis risk* refers to the lack of precision of a hedge. Since hedging often involves balancing several different investments, the potential divergence in price between the intended exposure and realized exposure is the basis.

Box 19-2 Costless Products

An example of a costless product is a difference between an exchange-traded American S&P 500 option (which can be exercised at any time during its tenor and includes dividends) and a "costless" European barrier note on the S&P 500 Index (which can only be executed on the last day of its life and does not include dividends). The exchange-traded American option may seem expensive compared the costless European option, which requires no payment up front. However, the value of the dividends and the ability to exercise the American option at any time makes the exchange-traded American option more valuable and may justify the cost. Other significant differences include the ability to buy and sell the exchange-traded American option at will and the counterparty risk of the European note.

Other risks may be complex and esoteric since many hedges are approximate and require a thorough and nuanced knowledge of the risks and of the hedging product (Box 4-1 in Chapter 4). An example is pin risk[11] around the expiration of a derivative.

Even hedges that involve the same product can have risk—for example, hedging an over-the-counter (OTC) derivative with a different counterparty creates counterparty risk (the risk of the profitable counterparty holding up the agreement and the other defaulting)—see Figure 16-1 in Chapter 16.

Hedging risk occurs across all dimensions. Another form of gap risk involves timing differences between payments. If an underlying investment is hedged precisely with a derivative, the payment delay may cause solvency problems. For instance, hedging a same-day settlement futures contract with a next-day settled underlying security (and its corresponding financing) may suffer from the difference in settlement dates. Even if all the dollar amounts offset each other, the fund may need the cash during the gap period for other purposes (for instance, to pay margin calls on other investments).

Embedded versus explicit risk requires investment managers to be careful in their choice and vigilant in their use of these tools. Many risks in hedges may be embedded in the structure and basis of the trade, rather than being obvious and explicit. There is also the possibility that the exposures being hedged are unwanted because they are poorly understood (either by chance of obfuscated by complexity).

Dynamic and static hedging may be a choice or a necessity. Some risks have no simple or inexpensive hedge and may require active management, including employing a trader and accepting execution risk. Bespoke derivatives may be deployed to manage these risks but can be expensive, illiquid, and may carry their own unique risks.

Preferences and constraints (such as tolerances to certain risks and not others, correlations, liquidity, sizes, product types and costs, and trading choices—such as intraday or net asset value [NAV] pricing) can limit hedge product choices. Similarly, market conditions like direction and volatility may favor some products over others.

Risk triggers and policy escalations are critical tools to control behavioral biases and give investors and other stakeholders confidence. Risk triggers may be categorized by focus and severity and may include:

- **Benchmark** level triggers and hedges that address the risk of the benchmark but not the deviation of the policy from the benchmark
- **Portfolio and policy** breaches, such as exceeding constraints and tracking error
- **Allocation** breaches, such as exceeding allocation bands
- **Selection** breaches, such as fallen angels
- **Counterparty exposure and securities lending/repo,** which may seem unimportant but can be an existential risk
- **Liquidity and capacity constraints,** which must be monitored
- **Proxy** securities or allocations, which are often used to simulate difficult-to-value investments
- **Performance and risk** breaches, such as value at risk (VaR), trends, and pairwise correlations
- **Active versus passive investments,** which can refer to managed versus indexed investments

- **Active versus passive violations,** such as willful deviations versus violations due to exogenous forces, like price action or credit downgrades
- **Structural** breaches, such as flow of funds, concentration, herding, leverage, liquidity, and fraud
- **Operational** violations, such as "fat fingers," information security, and client outflows
- **Limited control** over redemptions, which may hinder risk management

Redemptions are a difficult problem. For the manager of a liquid fund, client outflows may be challenging to slow down. For investors in illiquid funds, dealing with gates and lock-ups can be a problem, as we discussed in a previous case study (Box 1-4 in Chapter 1).

Frequently, outflows are contagious ("throwing out the baby with the bathwater") a fund manager's response may seem perverse (liquidating good investments to cover margin calls and losses in bad investments). The manager's response may be outside of his control or responsibility. Even though these situations may not be controllable at the source, decisive action may prevent excessive loss. During liquidity events, such as squeezes and flash crashes, first-mover advantage ("your first loss is our best loss") or patience ("this, too, will pass") may avoid panic and the resulting high transaction costs. Proper risk management and capital reserves can make patience palatable.

In situations where exit barriers exist, well-intentioned actions may have adverse effects. For instance, warning a manager that they are on a watch list may incentivize them to take excessive risk.

Systematic and Qualitative Heuristics

Systemic biases occur in many places in the investment process. They present themselves mostly in four categories:

- **Unfamiliarity,** the bias introduced by unintuitive decision-making, such as that by a machine learning algorithm

- **Opportunity** identification through the analysis and exploitation of competitor's behavioral biases
- **Built-in adverse** behavioral heuristics implemented in the system design, intentionally to align with manager expectations (like shortfall) or unintentionally from insufficient introspection
- **Oversight,** which can be swayed by behavioral biases that affect the operation of those systems, such as decisions to deviate from the system's recommendations

Myopia and automation. In Chapter 3, we discussed agency risk and the free-rider problem[12] and introduced the "arc of responsibility" (Figure 3-1). Principal-agent conflicts can be mitigated or amplified through automation by attempting to remove immediate responsibility from the decision-making process and replacing it with what may appear to be procedural decisions. The investment team retains responsibility but may injudiciously abdicate it. This can lead to a variety of possible outcomes.

> **Unintended consequences, the ever-present concern of interpretability, and explainability** are critical issues in machine learning and systematic investing. The implications of deliberate anticipations and reactions of simple systems may seem obvious, but the results may be inaccurate, especially when combined with other systems. More complex systems may appear simple and rational, but the mechanics that form their decision patterns may be unintuitive. The models may produce decisions that mimic their designers' intentions, but the route they take, such as those using combinatorial or probabilistic reasoning, may respond to critical decisions in ways that would be inconceivable to a layperson.

> **Aggravation in the pursuit of mitigation.** The *moral hazard*[13] of the principal-agent problem is not removed from investment risk but compounded by distilling intentions—malfeasant or otherwise—into imperfect rules or allowing unnatural systems to discover them. These systems can replace an intelligent but limited agent with a powerful agent that does not understand its purpose. Behavioral heuristics should be seen as a *call to action* for care and vigilance in systematization. By transferring responsibility from rational

decision-making to automated systems, we become vulnerable to new, unfamiliar, and dangerous risks that requires creative vigilance and conviction of purpose.[14]

Opportunity identification based on the behavior of crowds or individuals is the justification for many "tape reading" or technical techniques and is the focus of much of market microstructure research. Some examples include:

- **Expectations:** Differences between market expectations and the portfolio manager's views, performance chasing, or other potentially irrational motivations.
- **Irrationality:** Irrational motivations, such as meme stocks and glamor trades.
- **Insensitivity:** Price-insensitive agents and their capacity limitations, including sector and class rotations, rebalancings, index changes, and other value-insensitive trades.
- **Information rate:** The rate and extent of information distribution and absorption, including missing or misleading information, capability asymmetries, barriers to entry and exit, or the inability to process it effectively.
- **Fragility:** The elasticity or fragility of relationships between markets and vehicles.
- **Inaccessible opportunities:** Specific to the actors, such as tax trades, or opportunities that are accessible broadly, such as a dislocations.
- **Indirect vehicle:** Related to the action of a different vehicle, e.g., hedging.
- **Indirect decision:** Discretion unrelated to the actor, e.g., by a third party.
- **Flow:** Flow or other motivation unrelated to supply and demand.
- **Fragmentation:** Fragmentation, isolation, or inaccessibility of markets and vehicles.
- **Contagion:** Correlations or other herding behaviors, including cross-market linkages.
- **Exuberance** in excess of prudence.
- **Capitulation** or other forced actions.

The adverse implementation of behavioral biases as part of a system is commonplace and insidious, due primarily to the lack of introspection by its designers. There is value in intentionally addressing the saliency of objectives and constraints (like risk-adjusted returns and shortfall risk) over purely rational goals (like terminal wealth) but these concessions usually prevent the optimal achievement of rational outcomes. A system designed to keep losses palatable may be less likely to maximize wealth for a captive client ("gun shy"), but a wealth-maximizing strategy may lose investors before achieving a reward.

Oversight of systems is intrinsically infected by behavioral biases. Unanticipated and stressful circumstances lead to a lack of confidence in the robustness of a systematic strategy. The allure of Bayesian updates, stakeholder pressure, the temptation to conclude that "this time is different," and more conflicted temptations (like the desire to gamble out of a losing trade) impose human decision-making into the system.

Hirshleifer[15] offers a helpful taxonomy for behavioral biases:

- **Exogenous biases**, like herding and contagion
- **Endogenous emotional biases**, such as overconfidence, loss, and regret
- **Perseverance,** such as confirmation, representativeness, and the illusion of control
- **Processing overload (due to too much information)**, such as anchoring, framing, and availability

Many of these biases are difficult to identify and correct. Survivorship bias is one example.[16]

Learning and Improvement

This book is about earning sustained investment performance. Unlike other investing goals that are focused on urgent business outcomes (in the Eisenhowerian[17] sense), most investment strategies constantly struggle to improve or fall behind. Lucky windfalls and complacence can precede failures. The market is, at best, indifferent and unforgiving, but often hostile and punishing: the “road to hell is paved with positive carry.”

Investing is different. An investment business, like a sell-side bank or a Registered Investment Advisor (RIA), frequently focuses on the bottom line: selling products, generating fees, raising assets, retaining clients, and growing market share. These are the lifeblood of the enterprise and are necessary and proper concerns.

As investors, we have frequently argued against convenient narratives, oversimplified implementations of conditional relationships, and shortcuts, such as the disposition of old investments to raise capital for new purchases. While other groups in an investment firm have the luxury of being casual in their scrutiny of their investment thesis, true investors are punished for it more often than not.

True attribution can be avoided. Some investment professionals whose goals are primarily to raise money and generate fees, fail to measure and attribute their performance correctly. Their narrative is accepted at face value or lightly probed, and their shortcomings are explained away. They get rewarded, so why should they expose their weaknesses to those who determine their compensation and continued employment? Why should they uncover disturbing and potentially damaging truths about their skill? Their assets grow without attribution; their losses may recover when the market “rotates” back to their investment style.[18]

Good attribution is nuanced and difficult. It is difficult to properly assign causality even with honest and transparent access to investment decisions. Without access it is guesswork at best.

Attribution is just the first step, though most firms stop there. Critical attribution and introspection are without purpose unless the culture is committed to accepting and acting on insights. Sustained investment performance requires:

- **An honest, nuanced, and accurate** evaluation and introspection of performance relative to luck and low-cost alternatives to active management

- **Forgiveness of failure** with a goal of experimentation and improvement
- **Disdain for incompetence** and a commitment to an "extremely transparent," disciplined, and rigorous evaluation of new strategies and ideas
- **Collaboration with accountability,** to allow "cross-silo" innovation without a breakdown in culture
- **Identification and acknowledgment** of success to maintain that success and proliferate best practices throughout the organization
- **Acknowledgment that** meritocracy requires robust oversight to preserve discipline and culture

■ ■ ■

In this book, we began with some fundamental but often misunderstood concepts and discussed how to design and build a solid product and business. We examined developing a reliable quantitative strategy, from gathering the data to predicting which factors would perform and under which conditions. We demonstrated how to model and backtest with professional precision. Finally, we discussed how to correctly measure risk and manage the strategy on an ongoing basis.

We wrote the book with three people in mind: managers of quantitative investment efforts, managers and analysts of investments, and advanced students (including "students of the market" in the "school of hard knocks"). The purpose was to expand the reader's perception of how others in the ecosystem view quantitative investing, allow them to better understand the broad process, and to communicate across channels.

For instance, a style-focused hedge fund manager may perform better if his style drifts, but an asset allocator who invests in his fund needs the manager to maintain his focus even if the manager's style underperforms. The manager may think his function is to deliver returns, but the allocator needs to maintain control over style weights and will hedge the style if he wants to deviate. The allocator may not want the manager to decide that.

We struck a balance. Most readers will be overly familiar with some explanations and enlightened by others. Each reader will arrive with different mental puzzle pieces and should finish the book with a more complete picture than they began with. Investment professionals are experts in their areas and businesses but may lack exposure to other specialties. We hope every reader will find a nugget of useful information, even in the most familiar chapters.

Even the most sophisticated, flexible and scalable techniques are not a guarantee of success. The critical nature of human judgment and committed application of governance, culture, standards, rules, and processes must be integrated with technological applications to produce a robust, resilient, long-term investment business.

Good Hunting,

Michael Robbins
New York, New York
February 2023

Box 19-3 The Way Forward

We would like to remind you one last time that www.QuantitativeAssetManagement.com contains many resources that could not be included in the book. The website contains copious endnotes, examples, videos, computer code, lesson plans, and other works. They are works in progress, so please visit and use these resources frequently.

---

1. This quote, and it's variations are probably attributable to Archilochus; "We don't rise to the level of our expectations, we fall to the level of our training."
2. It may seem obvious, but many investors succumb to availability bias. They remember stories of spectacular successes (and some spectacular failures) but forget that the vast majority of investors fail quietly.
3. It is common to prioritize the urgent but less important over the important but not urgent. Over time, poor choices will cause systemic problems. Oftentimes, ruin is caused by one undisciplined exception that cascades into an unrecoverable failure.
4. In Chapter 1, we discussed how to effectively design incentives that minimize the role of luck and align interests with the business.
5. VXX is the iPath S&P 500 VIX Short-Term Futures ETN.
6. "Keeping them in the dark, feeding them [fertilizer]." Tracy Kidder, *The Soul of a New Machine* (Avon Books, 1982).

7. Also known as the devil's advocate, the "tenth man" opposes the consensus to ensure that unpopular arguments are not overlooked. See Yosef Kuperwasser, "Lessons from Israel's Intelligence Reforms," the Saban Center for Middle East Policy at the Brookings Institution, 2007.

8. A competent back office can make or break a systematic investment process. Portfolio management can be less important than operational day-to-day money management.

9. The knowledge and application of behavioral heuristics can help the management and investment teams identify rational and measured responses to a crisis.

10. "No plan of operations extends with certainty beyond the first encounter with the enemy's main strength." Helmuth von Moltke, *Moltkes Militärische Werke, II*, 1892.

11. Pin risk occurs when an option's strike price is nearly the same as the price of it's underlying reference at expiration. This creates uncertainty about whether the option will expire in-the-money or not.

12. A *free rider* problem results from a misalignment between benefit and responsibility. An *agency problem* occurs when one agent is responsible for another's well being but is incentivized to violate that responsibility, creating a *conflict of interest*.

13. A *moral hazard* exists when an agent is not properly incentivized to act properly, such as when there is an economic conflict of interest.

14. Thinking systems will develop their own heuristics. If recent evidence is a guide, machine behavioral heuristics will reflect our own without the tempering of shame. See Oscar Schwartz, "In 2016, Microsoft's Racist Chatbot Revealed the Dangers of Online Conversation: The bot learned language from people on Twitter—but it also learned values," *IEEE Spectrum*, November 2019.

15. D. Hirshleifer, "Investor Psychology and Asset Pricing," *Journal of Finance* 56, no. 4, (2001): 1533–1597.

16. In Chapter 4, we discussed that if a study were to span many years, most of the stocks that composed an index (the Russell 3000) at the beginning of a study will likely be replaced by the indexer by the end of the study. If the indexer had kept the original list, the index performance would be dramatically different from the actual, realized Russell 3000 index performance.

17. The Eisenhower method, matrix, or box refers to a time management paradigm that separates tasks into four quadrants—urgent and important, not urgent and important, urgent and not important, and not urgent and not important. In Eisenhower's words, "I have two kinds of problems, the urgent and the important. The urgent are not important, and the important are never urgent." See Dwight D. Eisenhower, "Address at the Second Assembly of the World Council of Churches," August 19, 1954.

18. In March 2021, Cathy Wood's ARK fund had lost 45 percent and underperformed the S&P 500 by 65 percent in the trailing year but still attracted net inflows.

INDEX

ABOUT THE AUTHOR

Michael Robbins is the Chief Investment Officer of a large investment firm. This is his sixth CIO appointment, including one for a bank with 8½ million clients. He has managed pensions, endowments, and family offices, and was the Chief Risk Officer for the State of Utah's systems. Michael sits on private equity boards of directors and is a professor at Columbia University, where he teaches quantitative investing, including graduate classes in global macroeconomic tactical asset allocation (GTAA) and environmental, social, and governance (ESG) Investing.