# Digital Finance v4: Automation & Intelligence

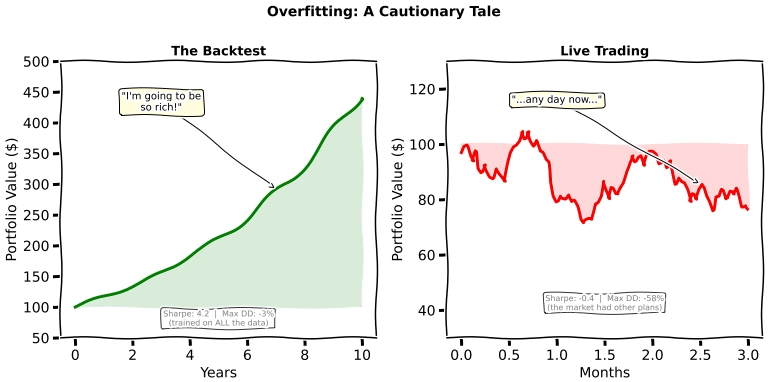
## Lesson 5.3: The Limits of Prediction – Time Series, NLP, and Market Efficiency

FHGR

February 11, 2026

**If ML is so powerful, why can't we predict stock prices? Understanding the limits matters as much as understanding the tools.**

# My Backtest Made Millions!




**The gap between backtest performance and live trading is the most expensive lesson in quantitative finance.**

# Learning Objectives

By the end of this lesson, you will be able to:

1. Explain why most ML trading strategies fail out-of-sample
2. Identify common pitfalls in financial ML (data snooping, look-ahead bias, survivorship bias)
3. Apply the Efficient Market Hypothesis to evaluate prediction claims
4. Design a sentiment analysis pipeline for financial text using NLP techniques
5. Evaluate the practical limitations of reinforcement learning for trading

---

**Critical thinking about prediction claims protects you from costly overconfidence.**

# From LLMs to Market Prediction

**Bridge from Lesson 5.2:**

- LLMs are powerful for understanding and generating text
- NLP can extract sentiment, entities, and relationships from financial documents
- **But can any model – no matter how advanced – reliably predict markets?**

**The central tension of this lesson:**

- ML excels at pattern recognition in structured, stationary data
- Financial markets are *non-stationary*, *adversarial*, and *reflexive*
- Understanding **why prediction is hard** is as valuable as any algorithm

---

**Markets are not image classifiers – the data generating process changes because participants learn.**

# What Is Time Series Stationarity?

**A time series is stationary if its statistical properties do not change over time.**

- **Constant mean:** expected value does not drift up or down
- **Constant variance:** volatility remains stable across periods
- **Constant autocorrelation:** relationship between $y_t$ and $y_{t-k}$ is time-invariant

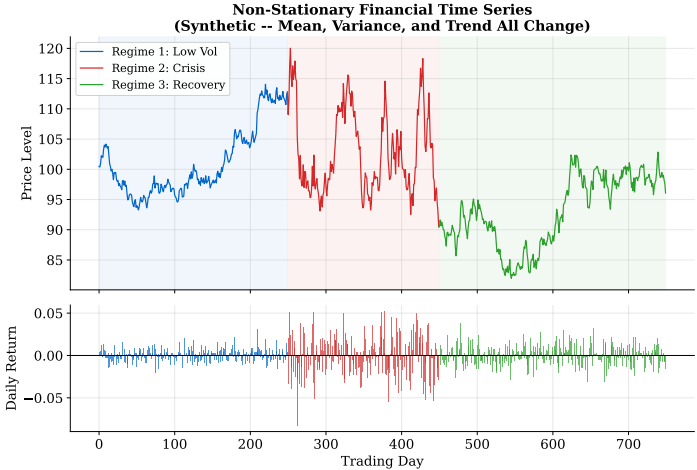
**Why this matters for ML:**

- Most ML algorithms assume training and test data share the same distribution
- Financial prices are **non-stationary** – trends, volatility clusters, regime shifts
- Returns are closer to stationary, but still exhibit structural breaks
- A model trained on calm markets may catastrophically fail in a crisis

---

**Stationarity is the foundational assumption most financial ML models quietly violate.**

# Non-Stationary Financial Time Series




**Mean, variance, and trend all change across regimes – a model trained on Regime 1 will fail in Regime 2.**

# What Is Autocorrelation?

**Autocorrelation measures how much today's value depends on yesterday's.**

- **ACF (Autocorrelation Function):** correlation between $r_t$ and $r_{t-k}$ at lag $k$
- **Financial returns:** near-zero autocorrelation (consistent with weak-form EMH)
- **Absolute returns:** significant positive autocorrelation (**volatility clustering**)

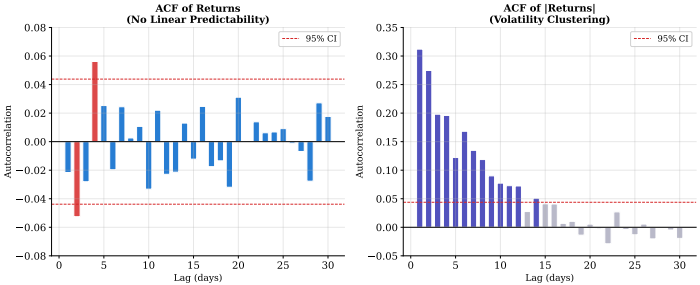
**Implication for ML:**

- Linear prediction of next-day returns from past returns is nearly impossible
- Volatility *is* predictable – GARCH models exploit this
- ML models that claim to predict return *direction* should be treated with skepticism
- Predicting *volatility* (risk) is more achievable than predicting *direction* (alpha)

---

**Returns are nearly unpredictable; volatility is clustered and forecastable – this asymmetry defines what ML can and cannot do.**

# Autocorrelation: Returns vs Absolute Returns




**Left: return ACF is within the confidence band (no linear predictability). Right: —return— ACF is highly significant (volatility clustering).**

# What Is a Regime Change?

**A regime change is an abrupt shift in the statistical behavior of markets.**

- **Volatility regime:** calm (VIX $\sim$12) $\rightarrow$ crisis (VIX $\sim$80)
- **Correlation regime:** diversified $\rightarrow$ "everything sells off together"
- **Monetary regime:** zero rates (2009–2021) $\rightarrow$ tightening (2022–2024)
- **Structural break:** new regulation, technology, or market microstructure change

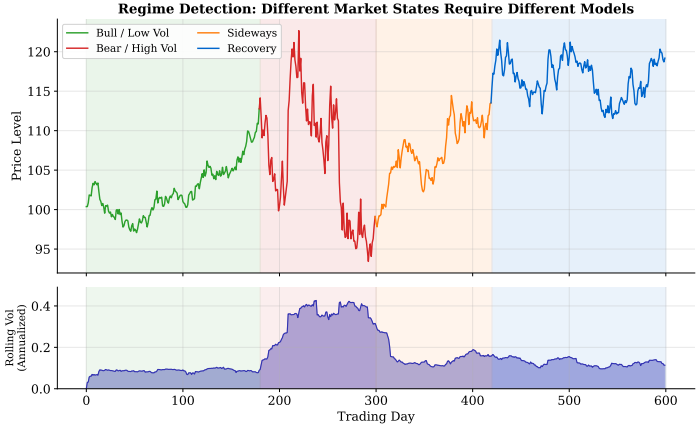
**Why regime changes break models:**

- Training data comes from one regime; deployment hits another
- Correlations, volatilities, and factor loadings all shift simultaneously
- No amount of in-sample testing protects against out-of-distribution events

---

**Models trained on 2015–2019 (low vol, low rates) were blindsided by both COVID (2020) and the rate shock (2022).**

# Regime Detection: Different Markets Need Different Models




**A single model cannot learn all regimes; regime-aware ensembles or online adaptation are needed.**

# What Is Data Snooping?

**Data snooping is testing many strategies on the same data and reporting only the winners.**

- Test 1,000 trading rules on 20 years of S&P 500 data
- By chance, $\sim$50 will be "significant" at $p < 0.05$
- Reporting the best one without correcting for multiple testing is **data snooping**

**Related pitfalls:**

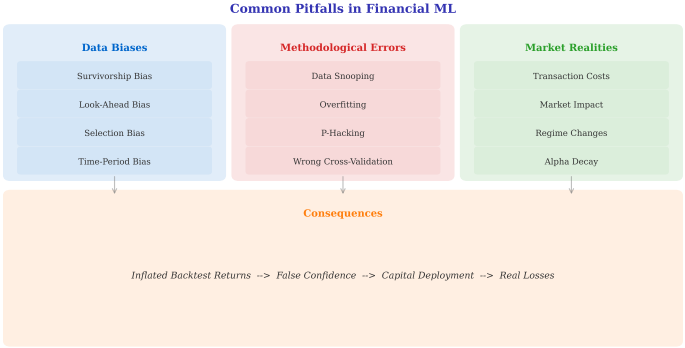
- **Look-ahead bias:** using future information unavailable at decision time (e.g., restated earnings)
- **Survivorship bias:** backtesting only on companies that still exist (excludes bankruptcies)
- **P-hacking:** tweaking model until $p < 0.05$, then calling it "discovery"

*Harvey et al. (2016) estimate that 50%+ of published finance factors are false discoveries.*

---

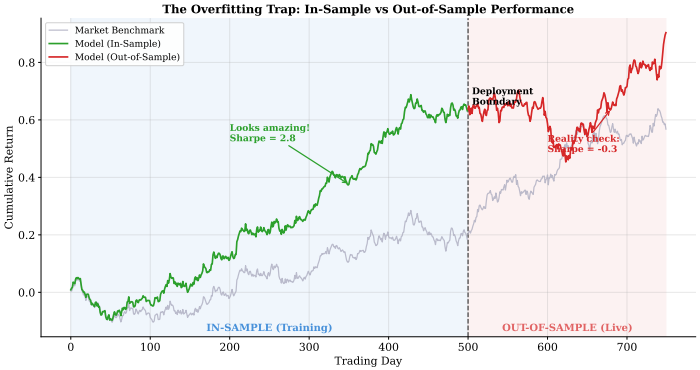
**If you torture the data long enough, it will confess to anything – but the confession is meaningless out-of-sample.**

# Taxonomy of ML Trading Pitfalls




**Pitfalls span data collection, methodology, and market realities – all lead to the same outcome: real losses.**

# The Overfitting Trap




**In-sample Sharpe of 2.8 collapsed to −0.3 out-of-sample – the model memorized noise, not signal.**

# What Is the Efficient Market Hypothesis (EMH)?

**EMH (Fama, 1970): prices fully reflect available information.**

- **Weak form:** prices reflect all past price and volume data
  - Implication: technical analysis cannot systematically profit
- **Semi-strong form:** prices reflect all publicly available information
  - Implication: fundamental analysis and news trading have limited edge
- **Strong form:** prices reflect all information, including insider knowledge
  - Implication: even insiders cannot profit (empirically rejected)

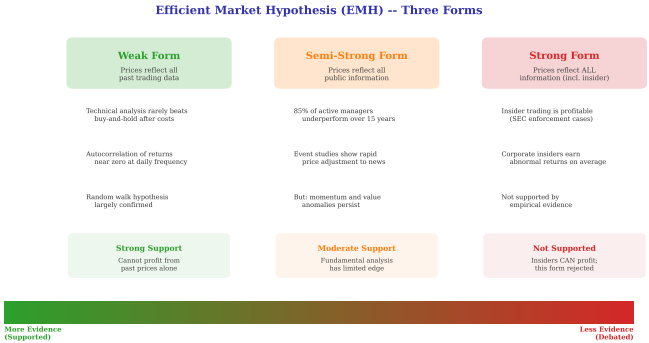
**Modern view – Adaptive Market Hypothesis (Lo, 2004):**

- Markets are *adaptively* efficient – inefficiencies exist but are competed away
- Alpha is scarce, temporary, and capacity-constrained

---

**85% of active managers underperform their benchmark over 15 years (S&P SPIVA 2023). EMH is approximately right.**

# EMH: Evidence Spectrum




**Weak form is well-supported; semi-strong is debated; strong form is rejected by insider trading evidence.**

# What Is Sentiment Analysis in Finance?

**Sentiment analysis uses NLP to extract subjective opinion from text.**

- **Sources:** news articles, SEC filings (10-K, 10-Q), earnings call transcripts, social media
- **Methods:** dictionary-based (Loughran-McDonald), ML-based (FinBERT, LLM prompting)
- **Output:** positive / negative / neutral score per document or entity

**Key NLP components:**

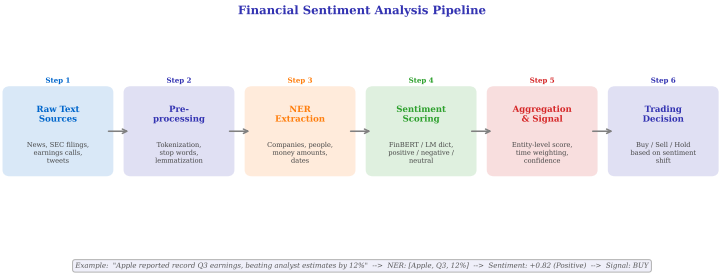
- **Named Entity Recognition (NER):** identify companies, people, amounts, dates
- **Earnings surprise:** compare actual vs. consensus estimate – sentiment shift matters
- **Temporal aggregation:** recent sentiment weighted more than old sentiment

**Limitation:** Sentiment signals are noisy, crowded, and decay rapidly once widely traded.

---

**FinBERT achieves $\sim$85% accuracy on financial sentiment classification, but accuracy $\neq$ profitable trading signal.**

# Financial Sentiment Analysis Pipeline




**Each step introduces potential errors: NER misidentifies entities, sentiment scores are context-dependent, aggregation loses nuance.**

# What Is Alpha Decay?

**Alpha decay: a profitable trading signal loses its edge over time.**

- **Discovery phase:** signal is proprietary, generates excess returns
- **Publication phase:** academic paper or word-of-mouth spreads the idea
- **Crowding phase:** many funds trade the same signal, compressing returns
- **Death phase:** signal no longer covers transaction costs after crowding

**Alpha half-life estimates:**

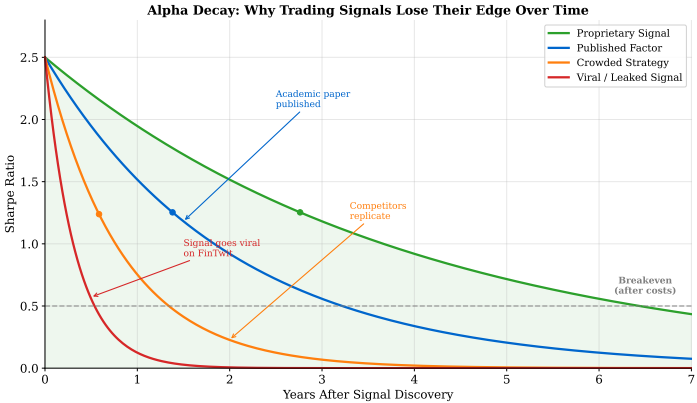
- Proprietary HFT signal: weeks to months
- Published academic factor (e.g., momentum): 3–7 years
- Viral social media signal: days to weeks

*"Alpha is not a stock; it's a decaying asset." – Marcos López de Prado*

---

**Once a signal is known, rational actors arbitrage it away – the Adaptive Market Hypothesis in action.**

# Alpha Decay: How Signals Lose Their Edge




**Proprietary signals decay slowly; published or crowded signals decay rapidly. All eventually approach the breakeven line.**

# What Is Reinforcement Learning for Trading?

**RL trains an agent to maximize cumulative reward through trial and error.**

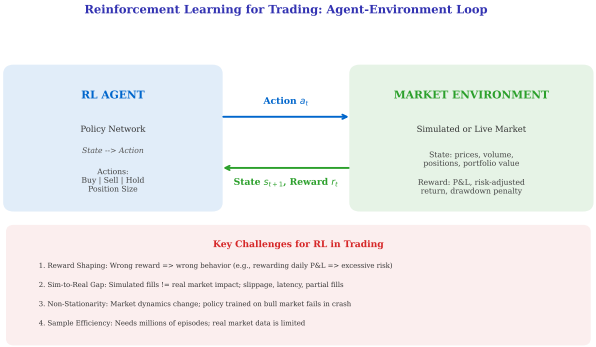
- **State:** current prices, portfolio positions, market indicators
- **Action:** buy, sell, hold; position sizing
- **Reward:** profit/loss, risk-adjusted return, drawdown penalty
- **Policy:** neural network mapping states to actions

**Key challenges for RL in finance:**

- **Reward shaping:** wrong reward $\Rightarrow$ wrong behavior (e.g., excessive risk-taking)
- **Simulation-to-reality gap:** simulated market $\neq$ real market (slippage, partial fills)
- **Non-stationarity:** policy learned in one regime fails in the next
- **Sample efficiency:** RL needs millions of episodes; real financial data is scarce

**RL has succeeded in games (AlphaGo, Atari) but markets are adversarial, non-stationary, and partially observable.**

# RL Agent–Environment Loop for Trading




**The sim-to-real gap is the Achilles heel of RL in trading: what works in simulation rarely transfers to live markets.**

# Proper Validation: Walk-Forward Testing

**Standard cross-validation (random splits) is wrong for time series.**

- Random splits create look-ahead bias (future data leaks into training)
- **Walk-forward validation:** train on past, test on next period, slide window forward

**Walk-forward protocol:**

1. Train on months 1–12, test on month 13
2. Train on months 1–13, test on month 14
3. Repeat until all data is used
4. Report average *out-of-sample* performance across all test periods

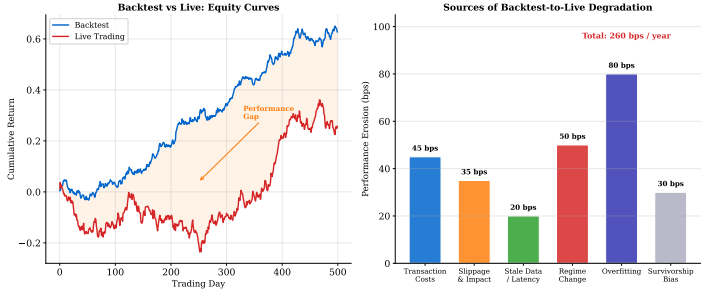
**Additional safeguards:**

- **Purging:** remove training samples near the test boundary (avoid label leakage)
- **Embargo:** leave a gap between training and test periods
- **Combinatorial purged CV:** López de Prado (2018) – multiple test paths

---

**Walk-forward validation is the gold standard for time series ML – never use random train/test splits on financial data.**

# Backtest vs Live Trading




**Transaction costs, slippage, regime change, and overfitting together erode 260 bps/year of backtest alpha.**

# Case Study: Long-Term Capital Management (LTCM)

**Nobel-Prize-winning models, $4.6B loss (1998):**

- **Team:** Myron Scholes & Robert Merton (Nobel 1997), top Wall Street traders
- **Strategy:** convergence trades – long undervalued bonds, short overvalued
- **Leverage:** 25:1 amplified small statistical edges into massive positions
- **What went wrong:** Russian default triggered global flight-to-quality
- **Model failure:** assumed normal distributions; ignored tail risk and liquidity
- **Rescue:** Fed-orchestrated $3.6B bailout by 14 banks to prevent systemic contagion

**Lesson:** Even the most sophisticated quantitative models fail when:

- Tail events exceed historical experience
- Leverage amplifies losses beyond recovery
- Liquidity vanishes when you need it most

---

**"When Genius Failed" (Lowenstein, 2000) – the definitive account of model overconfidence in practice.**

# Realistic Performance Expectations

**What is achievable in practice?**

| Metric | Good | Suspicious |
|---|---|---|
| Sharpe Ratio | 0.8–2.0 | $> 3.0$ |
| Annual Return | 8–25% | $> 50\%$ |
| Max Drawdown | $-15\%$ to $-30\%$ | $< -5\%$ |
| Win Rate | 50–55% | $> 70\%$ |

**Red flags in backtest results:**

- Sharpe $> 3.0$ – likely overfitting or look-ahead bias
- Zero losing months – model memorized the data
- No drawdowns – unrealistic or missing transaction costs
- "100% accuracy" – data leakage or survivorship bias

**The median quant hedge fund Sharpe ratio is 0.7 (2010–2023, BarclayHedge). Humility is a competitive advantage.**

# Survivorship Bias: The Invisible Error

**What is survivorship bias?**

- Analyzing only entities that "survived" (still exist) and ignoring those that failed
- Backtest on current S&P 500 members? You exclude companies that went bankrupt
- Hedge fund database? Only funds that chose to report (losers stop reporting)

**Impact:**

- S&P 500 backtest with survivorship bias overstates returns by $\sim$1–2% per year
- Hedge fund database returns are inflated by $\sim$3–5% per year (instant history + backfill)
- Mutual fund "track records" omit merged or closed funds (poor performers disappear)

**Mitigation:**

- Use point-in-time databases (constituents as of each date)
- Include delisted securities with their full return history
- Report results with and without survivorship correction

---

**Abraham Wald's WW2 bomber analysis is the classic survivorship bias example: reinforce where the missing planes were hit.**

# Named Entity Recognition and Earnings Surprise

**What is Named Entity Recognition (NER)?**

- NLP technique that identifies and classifies entities in text:
  - Organizations: "Apple Inc.", "Deutsche Bank"
  - People: "Jamie Dimon", "Jerome Powell"
  - Money: "$4.2 billion", "12% growth"
  - Dates: "Q3 2025", "fiscal year"
- Critical for mapping sentiment to the *right* company or event

**What is earnings surprise?**

- Difference between actual earnings and analyst consensus estimate
- **Positive surprise:** actual $>$ estimate $\Rightarrow$ typically bullish price reaction
- **Post-earnings announcement drift (PEAD):** prices continue drifting for weeks
- One of the most robust anomalies – but increasingly traded and decaying

---

**NER + sentiment + earnings surprise = a complete event-driven NLP pipeline for financial text.**

# Key Takeaways

**Core principles:**

1. Financial time series are **non-stationary** – stationarity assumptions break down
2. Returns have near-zero autocorrelation; **volatility** is predictable, direction is not
3. **Data snooping**, look-ahead bias, and survivorship bias inflate backtest performance
4. The EMH is approximately right: **alpha is scarce, temporary, and capacity-constrained**
5. Sentiment analysis with NLP is powerful but noisy and subject to **alpha decay**
6. RL for trading faces the **sim-to-real gap** – simulated markets $\neq$ live markets

**Practical advice:**

- Use walk-forward validation, never random splits
- Include transaction costs, slippage, and market impact in every backtest
- If results look too good to be true, they almost certainly are

---

**Understanding why prediction is hard is the first step toward building models that actually work.**

# Summary: The Limits of Prediction

**What ML Can Do:**

- Forecast volatility and risk
- Extract sentiment from text
- Detect anomalies and fraud
- Automate data processing
- Identify regime changes (after the fact)

**What ML Cannot Do:**

- Reliably predict price direction
- Avoid regime changes
- Replace risk management
- Guarantee alpha at scale
- Eliminate the need for human judgment

**Next Lesson:** 5.4 – Automation in Practice (putting all tools together in production)

---

**The most valuable skill in financial ML is knowing when not to trust the model.**