# Machine Learning in Finance

## Challenges, Solutions, and Future Prospects

A Comprehensive PhD-level Overview

Report Date: 2025-09-05

Presented by
Chae-Shick Chung
Dept. of Economics, Sogang Univ.

# Why Machine Learning Matters in Finance

## Overcoming Human Limitations in Financial Decision-Making

### Human Decision-Making Limitations

**Emotional Biases**

Humans, influenced by emotions and biases, often struggle with objective, fact-based financial decisions, especially with conflicts of interest.

**Pattern Recognition Limitations**

Human brains are adapted to 2D/3D, limiting their ability to recognize complex patterns in high-dimensional financial data.

**Automated Decision-Making**

Next-gen automation makes judgmental decisions, surpassing human emotional constraints.

### Machine Learning Advantages

**Objective Decision-Making**

ML algorithms provide objective, optimized decisions by overcoming human emotional biases.

**High-Dimensional Pattern Recognition**

ML excels at recognizing patterns in high-dimensional spaces (e.g., 100D), beyond human capacity.

**Robust Decision Framework**

ML offers systematic, consistent decision-making for complex financial markets.

**Generative AI (LLMs)**

LLMs (e.g., GPT-5) parse unstructured financial data (statements, news) into structured datasets, surpassing human analytical limits.

**Sentiment Analysis 2.0**

Foundation models integrate multi-modal signals (text, audio,

# The Sisyphus Paradox Challenge

## Collaborative Strategy Development and Its Pitfalls

### The Paradox

#### Problem Description

Can 50 discretionary investment managers or 50 PhDs collaborate to develop an investment strategy? This often leads to overfitting backtests, factor models, or other issues.

### Manifestations

#### Overfitting Backtests

Strategies that perform perfectly on historical data but fail in actual markets due excessive curve-fitting to noise.

#### Factor Model Limitations

Reliance on limited known factors that fail to capture the complexity of market dynamics.

#### "Squeezing the Towel" Effect

The opportunity to find "alpha" in the market rapidly converges to zero, making it difficult to sustain.


#### The Sisyphus Cycle in Financial ML

- Data Mining
- Strategy Development
- Backtesting
- Failure

# Meta-Strategy Paradigm Challenge

## Lack of Systematic Approach to Investment Strategy Development

### The Challenge

**Lack of Systematic Methodology**

Investment strategy development lacks a systematic approach like a flowchart or automotive production line, relying instead on individual contributions.

**Fragmented Responsibility**

Individuals often don't take responsibility for the overall strategy development, each contributing their part without coordination.

**Missing Unity**

Lack of unified approach leads to inconsistent strategies that fail to adapt to changing market conditions.

### The Solution: Meta-Strategy

**Caissa Capitol's Approach**

Successful implementation in options trading using a meta-strategy that systematically integrates multiple models and evaluation standards.

Multiple Models

↓

**Meta-Strategy Integration**

**Practical Implementation**

Using "half a dozen different volatility models for option pricing, and evaluating performance according to 30 different standards."

# Data Quality and Technical Challenges

## Garbage In, Garbage Out: Overcoming Financial ML Obstacles

### Data Quality

ML performance relies on data quality. Financial data is complex, noisy, and hard to acquire.

**Solutions:**

- Process unique, hard-to-manipulate data
- Structure data correctly
- Generate informative labels
- Identify predictive features

### Software Customization

Financial tasks need custom tools. Standard libraries offer shared advantages, limiting competitive edge.

**Solutions:**

- Develop custom code
- Tailor functions per task
- Customize tools for problems

### Hardware Challenges

ML is computationally intensive, requiring significant resources.

**Solutions:**

- Become HPC expert
- Collaborate with labs
- Master parallel computing
- Explore quantum computing

### Synthetic Data Generation

GANs and diffusion models create realistic financial time-series data, addressing sparsity and privacy.

**Solutions:**

- Overcome data scarcity
- Preserve data privacy
- Enhance model robustness

# Mathematical vs. Experimental Approaches

## The Trade-off in Financial Machine Learning

### Mathematical Approach

- Proofs take years/decades/centuries to develop.
- Investors cannot wait that long for solutions.
- Rigorous proofs often come too late for practical decisions.
- Focuses on certainty over practical utility.

### Experimental Approach

- Solves problems via experimentation, not proof.
- Uses memory retention data transformation.
- Applies experimental strategy evaluation.
- Uses distributed computing and quantum methods.

### Practical Solutions for Time-Sensitive Investment Decisions

- Adopt experimental mathematics for difficult problems.
- Implement memory retention data transformation.
- AI models (e.g., AlphaGeometry) assist in proving/approximating theoretical finance results.
- Use experimental methods for poor backtesting.
- Leverage experimental strategy evaluation.
- MARL enables 'what-if' market experiments before empirical testing.

# The Overfitting Problem

## Why standard validation methods fail in finance

### Problem Definition

- ✕ Standard cross-validation methods often fail in finance
- ✕ Many "discoveries" are false due to multiple testing

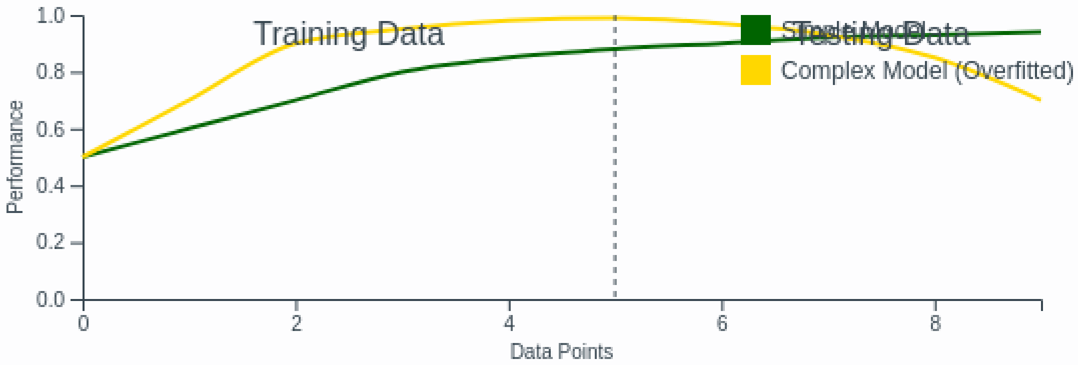
### Ethical Implications

- Overfitting is both unethical and wasteful of time
- Every backtest is somewhat overfitted

### Solutions

- Quantify the degree of overfitting
- Use robust asset allocation techniques

### Overfitting Visualization



# ML Production Chain in Finance

## Specialized Roles in the Financial ML Pipeline

**Data Curators**
Microstructure & Protocols

**Feature Analysts**
Signal Transformation

**Strategists**
Algorithm Development

**Backtesting**
Empirical Validation

**Deployment**
Product Integration

### Data Curators

- Collect, clean, index data
- Store and adjust data
- Focus on microstructure
- Handle bond, forex, stock data

### Feature Analysts

- Convert data to signals
- Information theory expertise
- Signal extraction techniques
- Feature importance analysis

### Strategists

- Develop ML algorithms
- Incorporate theories:
- - Behavioral bias
- - Asymmetric information

### Backtesting

- Test scenarios and paths
- Focus on overfitting issues
- Address second-order errors
- Validate robustness

### Deployment

- Integrate strategy code
- Algorithm implementation
- System integration
- Real-time execution

### LLM Copilots

- Assist strategists
- Suggest feature engineering
- Stress-test models
- Draft research notes

### Multi-Agent Systems

- Simulate heterogeneous investors
- Simulate central bank decisions
- Stress-test strategies under diverse beliefs

### AutoML & Orchestration

- Automated hyperparameter tuning
- Workflow orchestration (LangGraph, Ray)
- Shorten deployment cycles

# Machine Learning vs. Traditional Methods

## Bridging Algorithmic and Human Decision-Making

### Machine Learning

- Learns patterns in high-dimensional space without explicit guidance.
- Complements theory by discovering complex patterns.
- Adapts to new data and changing market conditions.

### Traditional Methods

- Econometrics: Applies classic statistical methods, lacks learning.
- Requires explicit formulation of relationships.
- Struggles with complex, high-dimensional data patterns.

### Quantamental Approach

Combines quantitative analysis with fundamental analysis.

- Requires investment managers to work with ML algorithms.

### Human-in-the-Loop

*"In ML applied to high-frequency trading, there will always be a 'human in the loop'."*

- Necessary for complex, rapidly changing domains.
- When applied properly, can be powerful and scalable.

### Neuro-Symbolic AI

Combines ML pattern discovery with symbolic economics models for enhanced interpretability.

### Explainable AI (XAI)

SHAP, counterfactual reasoning, and causal ML increase trust in black-box models.

# Core Problems and Solutions Matrix

## Key ML Challenges in Finance and Their Solutions

| Category | Challenges (Pitfalls) | Solutions |
| --- | --- | --- |
| Epistemologica | Backtesting | Feature Importance Analysis |
| Data Processing | Time clock | Volume clock |
| Data Processing | Stationarity | Fractional diff. |
| Classification | Fixed time | Triple barrier |
| Classification | Learning side and size | Meta-labelling |
| Classification | Non-iid | Weighting; Bootstrapping |

# Prerequisites for Financial Machine Learning

## Multidisciplinary Knowledge Required for Successful Application

### Core ML Knowledge

**Machine Learning Algorithms**
Pattern recognition, feature learning, and prediction methods

**Statistics**
Statistical analysis and probability theory for data interpretation

**Linear Algebra**
Vector spaces and matrices for data representation and transformation

**Convex Optimization**
Optimal portfolio construction and risk management

### Financial Domain Knowledge

**Market Microstructure**
Order books, liquidity dynamics, and trading mechanisms

**Portfolio Management**
Asset allocation and risk-adjusted returns

**Mathematical Finance**
Stochastic calculus and financial derivatives pricing

**Econometrics**
Time series analysis and statistical modeling

### Technical Skills

**Signal Processing**
Time series analysis and pattern recognition

**Information Theory**
Entropy and information gain for feature selection

**OOPs**
Object-oriented programming for system development

**Parallel Processing**
High-performance computing for ML workloads

# Future Research Directions

## Emerging Areas in Financial Machine Learning

### Meta-Strategy Dev.

- Systematic, methodology-driven strategy development
- Identify cross-asset features
- Combine multiple predictions
- Robust out-of-sample methods

### Overfitting Solutions

- Develop powerful backtesting techniques
- Quantify & mitigate overfitting
- Robust validation for financial models
- Address Sisyphus Paradox

### Hardware Optimization

- High-quality data acquisition
- Custom tool development
- HPC infrastructure
- Quantum computing apps

### Experimental Math

- Solve complex, hard-to-prove problems
- Experimental over mathematical proofs
- Data-driven pattern discovery

### Quantum Computing

- Financial optimization
- Complex portfolio optimization
- Risk analysis & management
- High-dimensional parameters

### Generative Agents

- Simulate heterogeneous investors
- Stress-test monetary policy
- Analyze market microstructure

### Graph AI (TGNNs)

- Systemic risk analysis
- Supply-chain contagion
- Financial interdependence

### AI Governance & Ethics

- Ensure fairness & transparency
- Regulatory alignment
- Ethical AI deployment

### Hybrid Quantum-AI

- Quantum-enhanced optimization
- Deep reinforcement learning
- Complex portfolio strategies