Research Statement

Executive Summary

I am a quantitative researcher specializing in applying Bayesian statistics and causal AI to financial markets, with over 15 years of experience building and trading alpha signals across multiple asset classes. My research at the University of Maryland focuses on integrating Bayesian inference, knowledge graphs, and temporal machine learning to enhance decision-making under uncertainty. I have successfully deployed over 100 alpha signals in production across US equities, global futures, and ETFs, achieving Sharpe ratios ranging from 1.5 to 5.0 across different strategies.

My approach combines rigorous statistical methodology with practical engineering, emphasizing interpretability, economic reasoning, and systematic overfitting controls. I have built two complete trading systems from data ingestion through portfolio construction, and led research teams that have integrated alternative data sources including news sentiment, social media, macroeconomic indicators, and SEC filings into quantitative strategies.

My research philosophy centers on using modern AI and machine learning to complement—not replace—traditional quantitative methods, maintaining a high-throughput research process where models are discovered, validated, and deployed at a predictable rate while accounting for their limited lifespan in competitive markets.

My full CV is here

- Education
- Publications
- AI Research

Overview

Since 2010 I have been working on applying machine learning (ML) and artificial intelligence (AI) to financial problems, ranging from alpha-strategies to portfolio construction.

I have built, traded, reviewed more than 100 alpha signals across asset classes (US equities, futures, crypto) and frequency (from minutes to days), using market data and non-market (alternative) data.

I have interviewed numerous quantitative analysts, led research teams, and conducted thorough due diligence on data.

I have experience using and developing custom research tools, including an alpha signal backtester, portfolio optimizers, and market impact models.

Over the past four years, I have taught graduate-level courses such as "DATA605: Big Data Systems" and "MSML610: Advanced Machine Learning" at the University of Maryland, College Park, to over 800 students.

My research focuses on Bayesian statistics and Causal AI at the University of Maryland. I integrate Bayesian inference, knowledge graphs, and temporal machine learning techniques to enhance decision-making under uncertainty. I have successfully applied my research in various fields, including financial prediction, predictive maintenance, failure prediction, demand forecasting, and energy trading. These methodologies yield improved accuracy and robustness across diverse applications.

Research Philosophy

My research foundation in Causal AI and Bayesian Statistics at the University of Maryland, and applied work at Causify.AI, informs my approach to quantitative finance. I focus on:

- **Causal Inference**: Moving beyond correlation to understand causal relationships in financial markets, enabling more robust models that capture underlying economic mechanisms rather than spurious patterns
- **Bayesian Methods**: Incorporating prior knowledge and uncertainty quantification into model development. This allows principled handling of parameter uncertainty and model selection.
- **Decision-Making Under Uncertainty**: Combining probabilistic reasoning with temporal machine learning to build adaptive models. These models adjust to changing market regimes and account for epistemic uncertainty.

This foundation shapes my philosophy on alpha generation, where interpretability, economic reasoning, and rigorous statistical validation are paramount.

Core Principles for Alpha Generation

I believe that to succeed in building alpha models:

- Models need to be understandable, understood, and based on economic priors
- State-of-the-art machine learning algorithms should complement traditional statistical methods
- Limited model lifespan should be accounted for in the process itself: my focus is a high-throughput research process where models are discovered and traded in production at a predictable, steady rate
- Maximizing the impact of alternative data means investigating signals:
  - At different time scales, ranging from minutes to days
  - For multiple asset classes (futures, equities, ETFs, FX, etc.)

    - For prediction of both factors and residuals
    - For both continuous and event-driven trading
- Ability to design and implement sophisticated and proprietary data pipelines is needed in addition to expertise in quant methodologies
- Combination of strong engineering practices and rigorous research and development (R&D) accelerates research results

Applications

Portfolio Optimization

My approach to portfolio construction emphasizes pragmatism and iterative improvement:

Construction Philosophy:

- **Start Simple**: Begin with realistic portfolio optimization that respects trading constraints
- **Iterative Improvement**: Progressively enhance optimization to address signal peculiarities (e.g., sparsity, different time scales)
- **Leverage Existing Infrastructure**: Utilize available components for:
    - Mean-variance optimization with standard constraints
    - Covariance estimation
    - Transaction cost models
    - Market impact models

Portfolio Constraints and Management:

- Market neutrality (primary concern)
- Low sector exposure
- Factor exposure controls (verified ex-post)
- Integration with Barra/Axioma factors for optimization

Prediction Targets:

- Optimize for raw returns (from 5 minutes to 1 day ahead)
- Mixer blends forecasts from all models
- Future goals include single-step and multi-period optimization

Signal Integration:

- Alternative data signals are typically additive to microstructure, price-based, and fundamental signals due to low correlation
- Occasional challenges arise with strict factor exposure limits when alternative data predicts primarily a single factor
- Solution: Dedicated system for trading sectors/factors

Execution Approach:

- Leverage available infrastructure (market impact models, trading algos, fill simulators)
- Incorporate actual models into backtesting/optimization systems
- Account for costs and optimize for target trading backend
- API-based target portfolio specification for next period

Alternative Data and Alpha Generation

Alternative data sets are prioritized based on:

- Large coverage of the trading universe
- Frequent updates
- Usable across different asset classes (e.g., equities and futures) to amortize research and acquisition costs

Data Sets with Past Research Success:

- News (e.g., news headlines or full-text using natural language processing)
- Social sentiment (e.g., from blogs, Twitter, StockTwits)
- Anonymized spending reports, credit card transactions, payroll data
- Macroeconomic data (e.g., for nowcasting inflation, industrial production, housing starts)
- Business reporting and disclosures
- Smart-phone app data for geolocation (e.g., for retail traffic)
- Supply chain information (e.g., trucking, rail freight data)
- Satellite data (e.g., for agriculture and ship tracking)
- Online retail prices (e.g., for inflation)
- Weather data

Models Developed and Deployed (with Sharpe Ratios):

- News sentiment models using Bloomberg, Reuters, Ravenpack (Sharpe ratio ~1.5-2)
- Social sentiment from Twitter and StockTwits (Sharpe ratio ~1.0-2.0)
- SEC insider trading signals (Sharpe ratio >1.5)
- SEC fundamental analysis with Compustat integration (Sharpe ratio >2.0)
- Value strategies combining Compustat and news (Sharpe ratio >2.0)
- Oil/natural gas models using news and 15-min futures (Sharpe ratio 3-5)
- Macroeconomic signals from supply chain datasets (Sharpe ratio ~1-1.5)
- Spot foreign exchange (FX) models using purchasing power parity (Sharpe ratio ~2-3)
- Short-Term Interest Rate (STIR)/bonds future arbitrage with treasury auction data (Sharpe ratio ~1.5-2)
- Microstructure models using order book data (Sharpe ratio >3.0)

Asset Classes and Trading Frequencies

Primary Targets:

- US equities (market neutral, sector neutral, directional strategies)
- Global futures (outrights, calendar spreads for financials, energy, interest rates, currencies, commodities, agriculture)
- US exchange-traded funds (ETFs) for factor and sector prediction strategies

Secondary Targets:

- Spot foreign exchange
- Global equities (Japan, Europe)

Trading Frequencies:

- Mid-frequency strategies ranging from minutes to days
- Optimal frequency depends on the alpha models from specific alternative data sets and transaction cost estimates

Frequency Examples by Data Type:

- Many models from alternative data sets trade once per day
- Some data sets (news, social sentiment, blogs) can be traded at higher frequencies, from 15-minute intervals up to every minute
- Other data sets allow trading around specific events (earnings, mergers and acquisitions, executive departures, credit rating changes) on the time scale of hours, days, or weeks
- Some models combine market microstructure data with fast alternative data (e.g., news) to trade on minutely or hourly time scales

Research Methodology

My approach to alpha-generation research is designed to combat overfitting while maintaining high research velocity. The methodology combines rigorous statistical controls with practical mechanisms to prevent common pitfalls in quantitative finance.

I follow a structured process that prioritizes data integrity and statistical rigor:

1. **Data Isolation**: Separate holdout data immediately before any work begins to prevent information leakage
2. **Exploratory Analysis**: Rigorously explore raw data without supervision from prices, understanding biases, collection processes, and data integrity

3. **Literature Review**: Review context and generation of data sets while avoiding publication bias
4. **Hypothesis Formulation**: Specify all hypotheses upfront, counting every parameter choice as a separate hypothesis
5. **Feature Engineering**:
    - Apply data transformations (e.g., differentiation, standardization, outlier detection) to extract information from data
    - Dimensionality reduction using Principal Component Analysis (PCA) and Independent Component Analysis (ICA)
    - Deseasonalization and residualization
6. **Backtesting**: Test hypotheses using cross-validation, regularization, and walk-forward testing
7. **Robustness Testing**: Bootstrap analysis, sensitivity analysis, and clipping tests
8. **Stationarity Testing**: Verify metrics remain consistent over time
9. **Multiple Hypothesis Correction**: Apply Benjamini-Hochberg adjustment and White-reality check
10. **Independent Review**: Double-check analysis and implementation with fresh eyes
11. **Out-of-Sample (OOS) Validation**: Open OOS box only when no further changes are possible
12. **Portfolio Integration**:
    - Apply transaction cost models
    - Test correlation with existing signals
    - Mean-variance optimization

Controlling Overfitting

This process combats overfitting through multiple layers of protection:

- **Statistical Controls**:

    - In-sample cross-validation and OOS testing
    - Prevent train/test information leakage to trust OOS results
    - Multiple hypothesis testing corrections (Benjamini-Hochberg adjustment, White-reality check)
    - Bootstrap and sensitivity analysis for robustness

- **Process Discipline**:

    - Specify and count all hypotheses upfront
    - Do not HARK (Hypothesize After Results are Known)
    - Maintain velocity to turn ideas into trading strategies quickly

- **Practical Safeguards**:

    - Implement mechanisms to avoid mistakes (e.g., future peaking, overly optimistic assumptions on market impact)
    - Apply common sense to model validation and interpretation
    - Independent review with fresh eyes

Machine Learning and AI Approach

My philosophy on applying AI and ML to quantitative finance emphasizes complementing traditional methods rather than replacing them:

Core Principles:

- Accelerate how research in finance is done (not change the basics of quant research)
- Extract forecasts from noisy and high-dimensional data sets
- Complement traditional statistical methods (not replace them)

Implementation Strategy:

- **Simple First**: Simple models should always be preferred over complex ones
- **Balance Complexity**: Strike a balance between model complexity and data to minimize overfitting
- **Strategic Application**: Deploy machine learning when:
    - New classes of data become available
    - Old alpha strategies are arbitraged away to extract more information
    - Simpler models lack capacity to explain market reactions
    - Widely copied strategies diminish headroom

Applications in Practice:

- Feature extraction from alternative data sets
- Natural language processing (NLP) for news and social sentiment analysis
- Deep learning for text processing
- Dimensionality reduction and residualization
- Pattern recognition in complex, high-dimensional data

Modern AI Technologies:

Recent advances in AI have opened new possibilities for quantitative finance, which I evaluate within my rigorous research framework:

- **Large Language Models (LLMs)**: Applying models like GPT and Claude for extracting nuanced information from financial text, earnings call transcripts, and regulatory filings. LLMs excel at understanding context and relationships that traditional NLP methods miss.

- **Transformer Architectures**: Using attention mechanisms for both time-series forecasting and text analysis. Transformers capture long-range dependencies in market data and can model complex temporal patterns.
- **Graph Neural Networks (GNNs)**: Leveraging GNNs to model knowledge graphs in causal inference, capturing relationships between companies, sectors, and economic factors. This aligns with my research on causal AI and enhances understanding of market interconnections.
- **Deep Learning Frameworks**: Utilizing PyTorch and TensorFlow for building custom architectures tailored to financial prediction tasks.

These modern techniques are most valuable when:

- Traditional methods cannot capture the complexity in the data
- Sufficient data is available to avoid overfitting
- The model remains interpretable or can be validated through rigorous testing
- The incremental benefit justifies the additional complexity

Key Insight: Traditional finance relies on heuristics and ordinary least squares (OLS) regression. ML offers tools to extract value that traditional methods cannot access, but must be applied rigorously within the disciplined research framework to avoid overfitting in financial markets' low signal-to-noise environment.

Technology and Infrastructure

I have built comprehensive trading infrastructure to support the research process and production trading.

Core Infrastructure Components:

- Toolkit for exploratory analysis of alternative data
- Research pipeline (feature computation, hypothesis testing, cross-validation)
- Event study framework for event-driven trading
- Backtesting system for multiple asset classes
- Mean-variance optimizer with constraints (cvxopt, numpy)
- Portfolio attribution and analysis tools
- Market impact models
- Real-time data processing pipelines
- Automated model deployment system

Technology Stack:

Core Platform:

- Linux-based infrastructure with Python as primary language

- Scientific computing: NumPy, pandas, SciPy for numerical analysis
- Traditional ML: scikit-learn for classical machine learning algorithms
- Deep learning: PyTorch and TensorFlow for neural network architectures
- Optimization: cvxopt for portfolio optimization, custom solvers for constraints

Data and ML Operations:

- Cloud infrastructure: Amazon AWS (EC2, S3, RDS) for scalable compute and storage
- Data pipelines: Real-time and batch processing for market and alternative data
- Version control: Git for code, DVC for data versioning
- MLOps: Model versioning, automated retraining, and deployment pipelines
- Monitoring: Real-time model performance tracking and alerting systems

Integration and Deployment:

- API integrations for market data (Bloomberg, Reuters), fundamental data (Compustat), and alternative data sources
- Order management system (OMS) integration for trade execution
- Database systems: PostgreSQL for structured data, time-series databases for market data
- Two complete end-to-end trading systems built from scratch (data ingestion through portfolio construction)

Example of Trading Research in 2015-2019

The following example illustrates the application of my research methodology in practice, covering the period from 2015 to 2019.

2015: Building the Foundation

Context:

- Joined as the 2nd person in the alternative data group in Berkeley
- No existing research pipeline for "low frequency" trading, no data APIs
- Everything was geared towards high-frequency trading (HFT) with a pervasive assumption that nothing was held overnight
- Initial focus: futures markets

Futures Trading Scope:

- 80 futures contracts across US, Europe, and Asia with varying liquidity levels
    - Asset classes: equity, fixed income, currency, commodities, energies
    - Instruments: outrights, calendar spreads, butterflies (curvature of term structure)
    - Trading frequencies:

        - 20 contracts: Time-Weighted Average Price (TWAP) to TWAP (trade over the day)
        - 13/19 contracts: 5 minutes
        - Intraday: around open, around close, open/mid-day (good results but not traded due to operational complexity and redundancy)

2016-2017: Scaling Up

By mid-2016:

- Deployed ~40 models across multiple strategies:
    - Macroeconomic
    - Alternative data
    - Microstructure
    - Price-volume (Commodity Trading Advisor-like)
- Several alternative data models from 4-5 datasets online:
    - Macroecon (GDP estimates tracking trains and ships)
    - Oil models using oil production and temperature data
    - News sentiment
    - Social media sentiment
    - Microstructure and price-based futures models
- Research hit rate ~50% (2-3 data sets didn't pan out)
- Historical Sharpe ratio ~4-5 for futures, before costs

Teza Futures Returns (Jan 2016 - Jan 2017):

- Overall: Up a few percent (moderate performance)
- Alternative data signals: Performed well

2017-2018: Peak Performance and Transition

2017 Performance:

- First year when the futures system was fully completed
- $1B assets under management (AUM)
- ~25% of risk allocated to alternative data
- Mid-2016 to end of 2017: Sharpe ratio ~3, $90M profit and loss (PnL)
- 2x volatility target → returns ~56%

Strategic Shift:

- After 2017, started focusing on equities, FX, and ETFs

Company Restructuring:

- Significant organizational changes:

    - Leadership departures: Chief Investment Officer (CIO), Chief Risk Officer (CRO), and 2 Chief Technology Officers (CTOs) left
    - HFT business sold to QuantLab (70 people plus substantial core infrastructure)
    - Previously working infrastructure had to be reimplemented from scratch

2018-2019: Challenges and Resilience

Trading Challenges:

- Feb 2018: Significant drawdown in futures (10% loss in one month)
    - Mean-reversion models started losing money simultaneously
    - Risk control failure: systems that had been performing well received increased weight, amplifying losses
- New portfolio manager (PM) brought in for futures
    - Rebuilt portfolio optimization from scratch using already-researched signals
    - Started adding signals one at a time with overfitting checks
    - Sept 2018: Began adding back alternative data signals

Alternative Data Signal Deployment Issues:

- All alternative data signals (both old and new) were not traded in 2018
    - Affected both equities and futures
    - When signals were eventually put in production, backtests showed they would have performed well

Research Output:

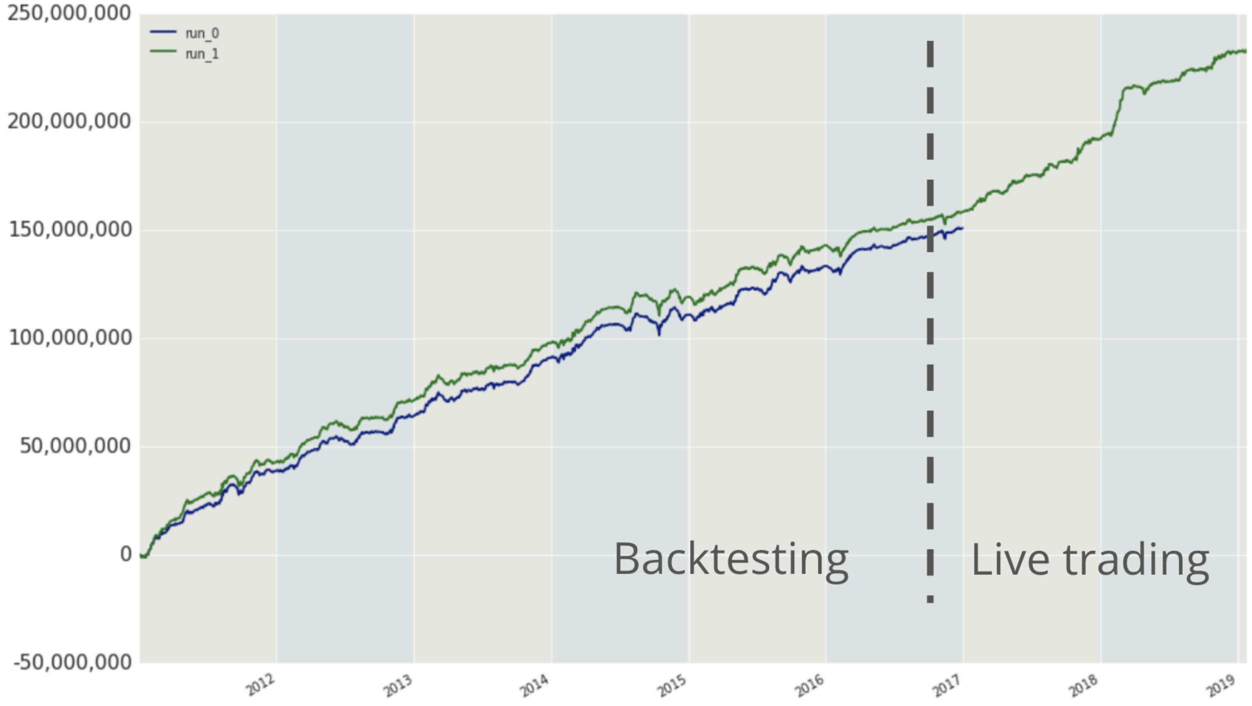
In 2018, the alternative data team released ~10 signals from 4-5 data sets:

- Compustat "latency arbitrage" (SEC data, proprietary)
- Macro-econ news for futures (RavenPack)
- Multiple news-related signals (both event-driven and continuous, RavenPack)
- FX signals estimating Purchasing Power Parity (Numbeo)
- Two social sentiment signals (PsychSignal and Social Market Analytics)
- Auditor change signals
- Company disclosure signals

Example of Quant Strategies

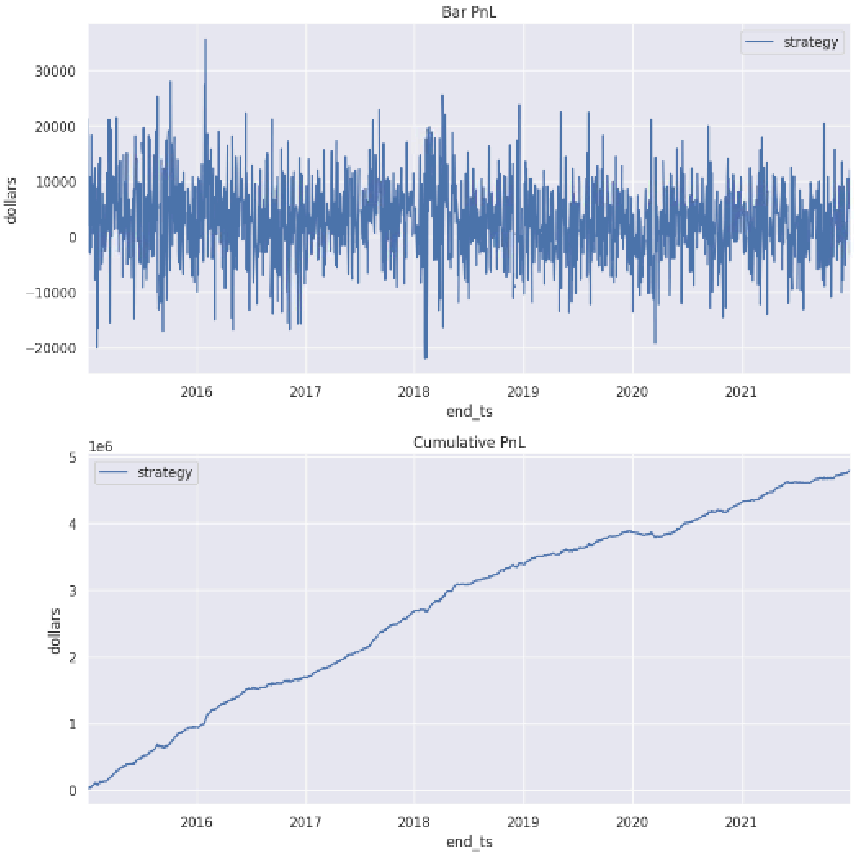
US Equity: Daily

- Frequency: multiple events per day
- 6 different models (e.g., earning announcement, stat arb, book pressure)



US Equity: Intraday

- Universe
  - 600 most liquid US equities, updated monthly
- Microstructure-based
  - Machine learning on order book data
- Trade 5 to 30 minute waves
  - With or without overnight holdings
  - Static feature weights derived independently from price data
- Execution
  - Mostly passive/market making TWAP/VWAP
- Sharpe ratio: >5
- Capacity: $10-50M



```
ratios      sharpe_ratio                             6.32
            sharpe_ratio_standard_error              0.39
            sr.tval                                 16.80
            sr.pval                                  0.00
            kratio                                   2.50
dollar      gmv_mean                         7,167,149.61
            gmv_stdev                        2,118,707.90
            annualized_mean_return             684,999.16
            annualized_volatility              108,435.71
            max_drawdown                       102,796.80
            pnl_mean                             2,810.58
            pnl_std                              6,908.47
            turnover_mean                  594,860,246.22
            turnover_stdev                 184,110,562.47
            market_bias_mean                     2,315.11
            market_bias_stdev                  134,133.96
percentage  annualized_mean_return                   9.72
            annualized_volatility                    2.28
            max_drawdown                             5.77
            pnl_mean                                 0.04
            pnl_std                                  0.15
            turnover_mean                        8,263.27
            turnover_stdev                         375.41
            market_bias_mean                         0.07
            market_bias_stdev                        1.86
```

Alternative-data for US Equities / Futures

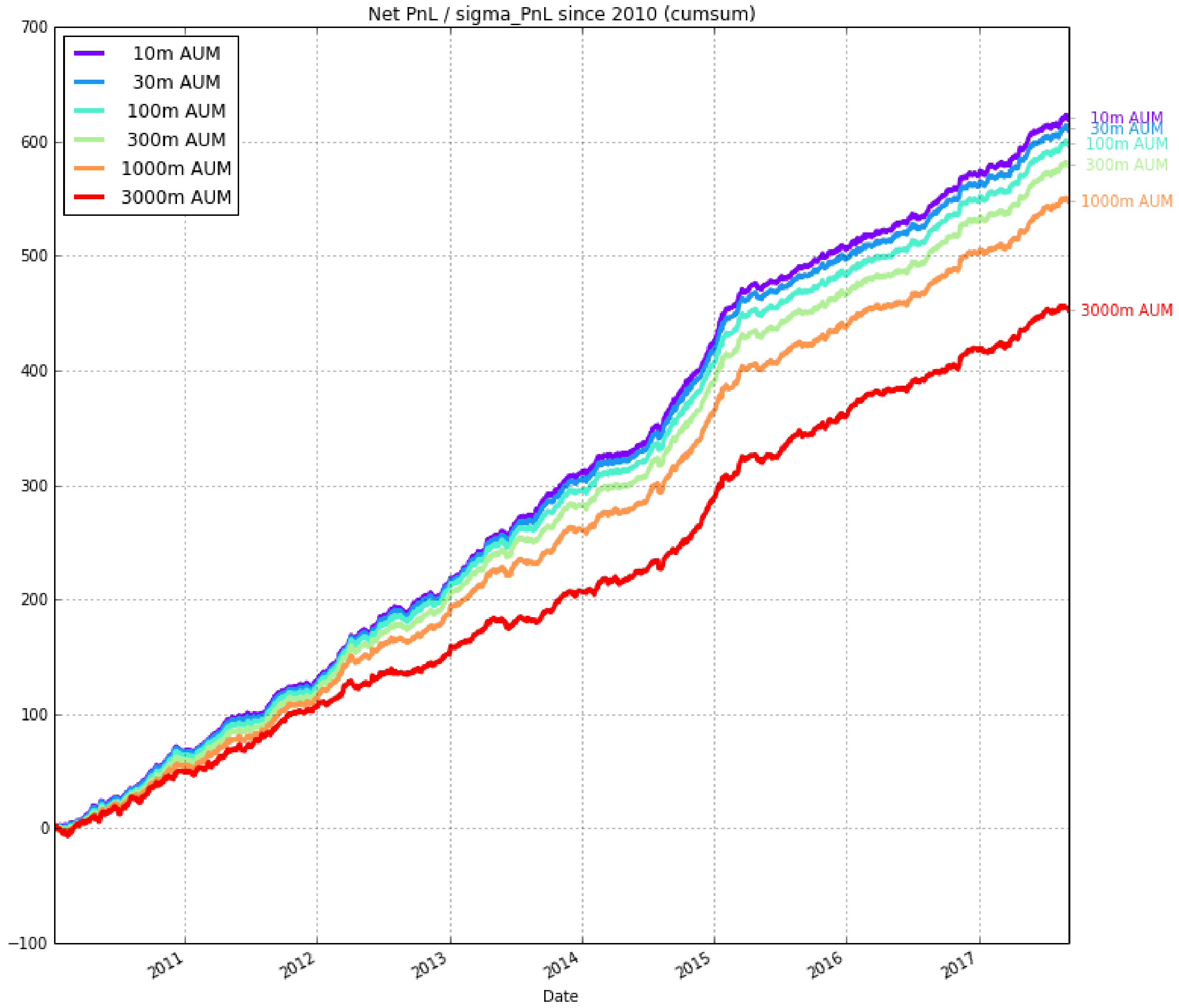
Between 2015 and 2019 I was head of data at Teza. My main duty was to produce alpha strategies for Teza funds. I've supervised a team of 5 researchers in Berkeley, CA and around 10 researchers / developers in the Moscow, Russia office.

Teza ran two funds from external investors:

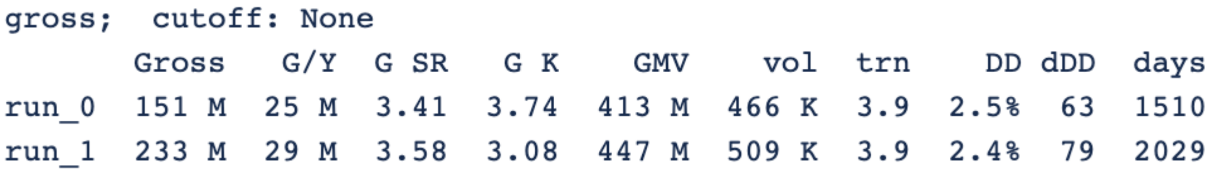
- Global futures (around $1b)
- US equities (around $600m)

My team contributed around 25% of PnL to both funds.

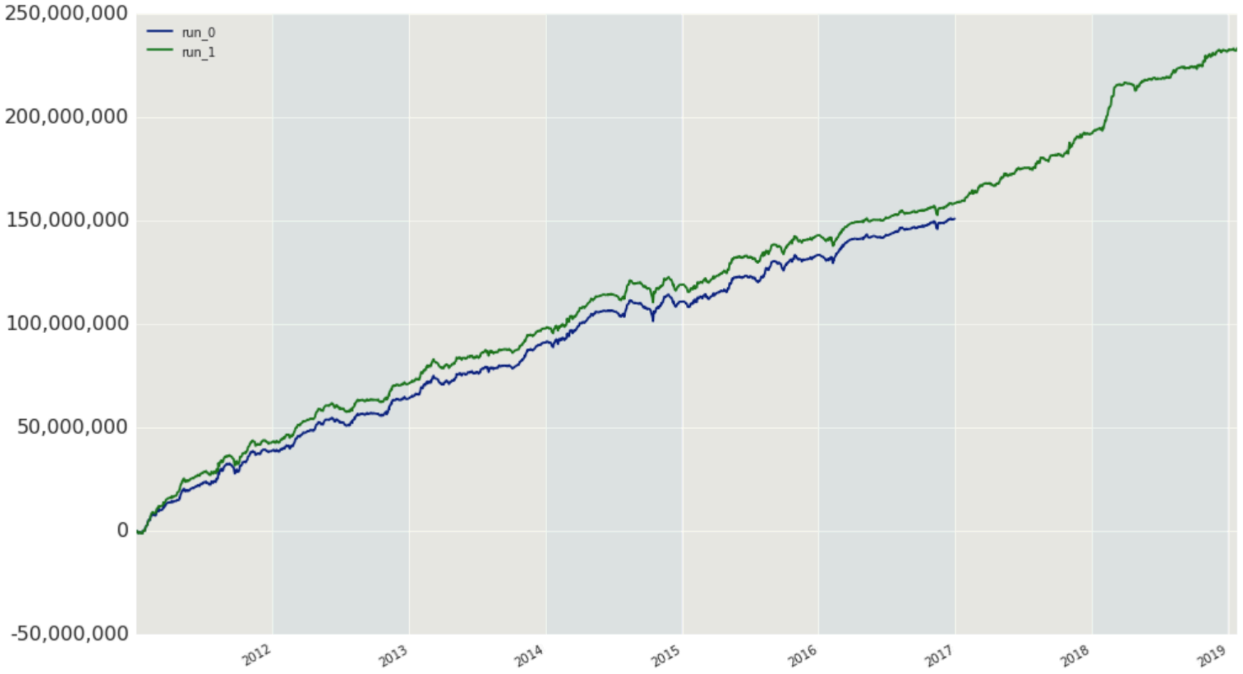
The futures fund went live in 2015. The realized performances are below.



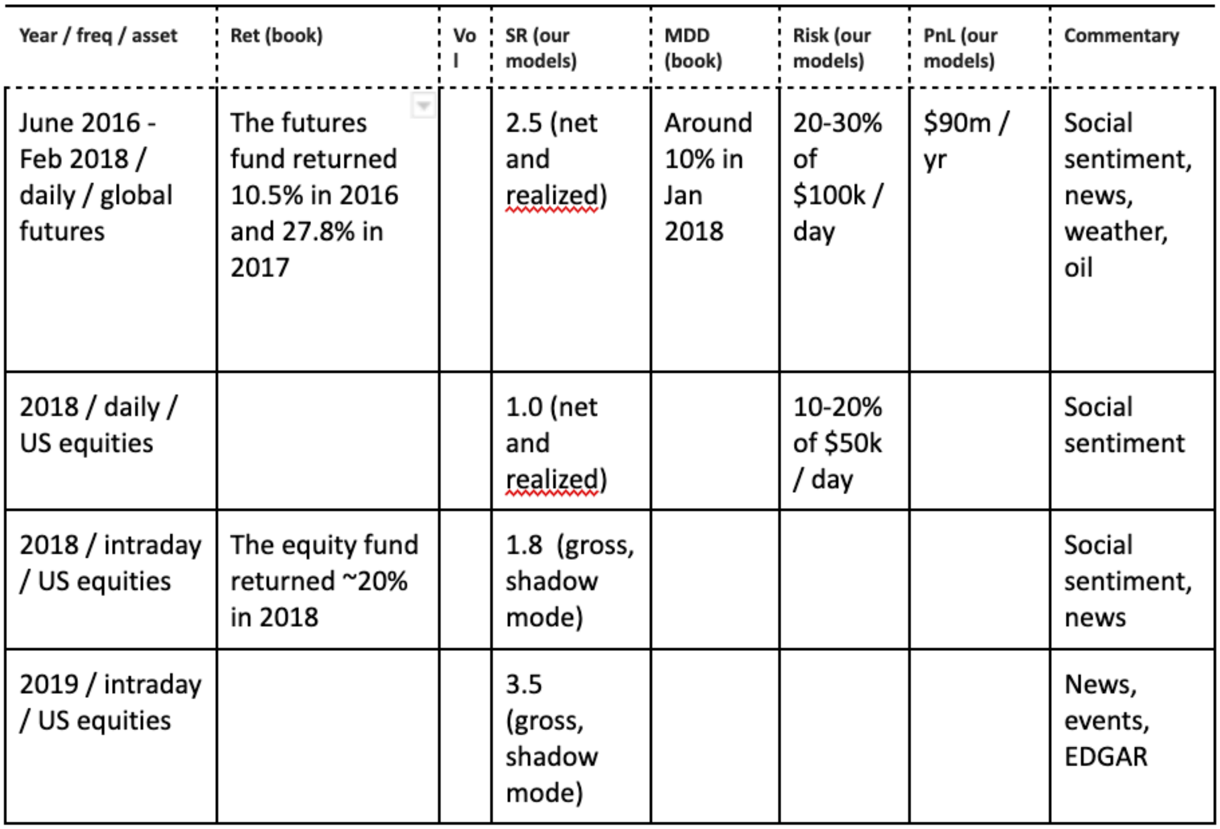
The equity fund went live around 2017. Realized performance are below.



```
gross;  cutoff: None
       Gross    G/Y  G SR   G K    GMV    vol  trn    DD dDD  days
run_0  151 M  25 M  3.41  3.74  413 M  466 K  3.9  2.5%  63  1510
run_1  233 M  29 M  3.58  3.08  447 M  509 K  3.9  2.4%  79  2029
```



Some details on the strategies we traded.



Year / freq / asset	Ret (book)	Vol	SR (our models)	MDD (book)	Risk (our models)	PnL (our models)	Commentary
June 2016 - Feb 2018 / daily / global futures	The futures fund returned 10.5% in 2016 and 27.8% in 2017		2.5 (net and realized)	Around 10% in Jan 2018	20-30% of $100k / day	$90m / yr	Social sentiment, news, weather, oil
2018 / daily / US equities			1.0 (net and realized)		10-20% of $50k / day		Social sentiment
2018 / intraday / US equities	The equity fund returned ~20% in 2018		1.8 (gross, shadow mode)				Social sentiment, news
2019 / intraday / US equities			3.5 (gross, shadow mode)				News, events, EDGAR

As you know, hedge funds are quite secretive, some information leaked to the press about our machine learning research in the Berkeley office.

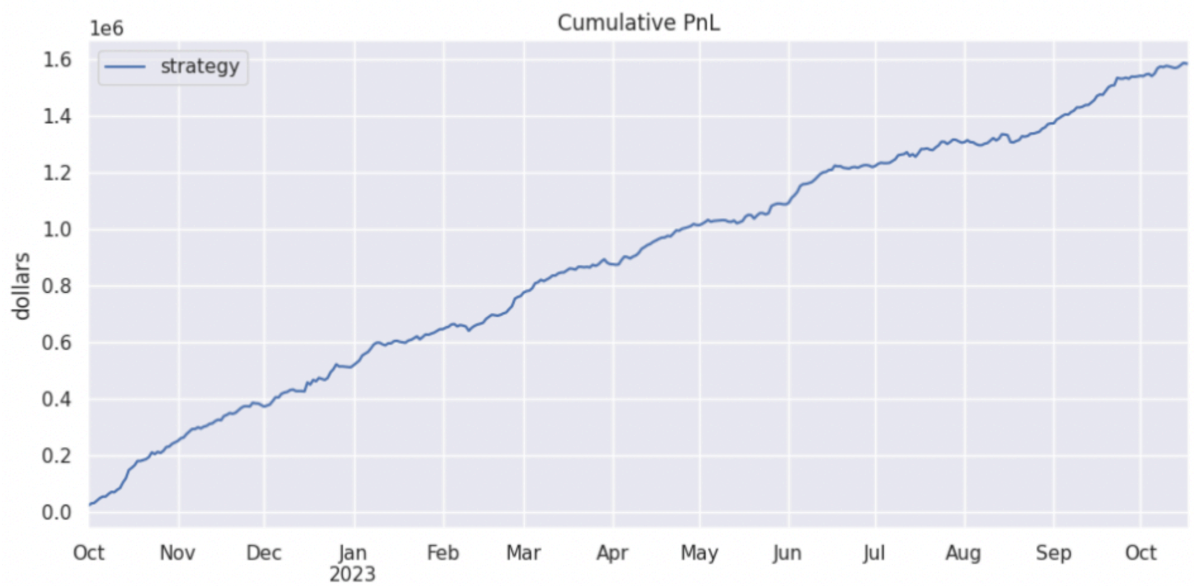
From The Future Is Bumpy: High-Tech Hedge Fund Hits Limits of Robot Stock Picking - WSJ

Teza Capital Management LLC credits machine learning in part for its more than 50% gain so far this year.

Alpha research on Crypto statarb

An example of alpha HFT strategies for crypto is below.

- *Instrument type*: Perpetual futures 25 crypto
- *Universe*: Most liquid 25 crypto tokens (excluding stable coins)
    - E.g., ETH, BTC, SAND, STORJ, GMT, AVAX, BNB, APE, MATIC, DOT, UNFI, LINK, XRP, RUNE, NEAR, FTM, WAVES, AXS, OGN, DOGE, SOL
    - Total daily trading volume: Around $1.2 billion
- *Exchange*: Binance

- 

  Performance
  - *Number of trades per month*: 216,000 (=25 cryptos * 12 trades per hour * 24 hours * 30 days)
  - *Annualized return*: 150% / yr since inception (Jan 2021)
  - *YTD returns*: 310% / yr, with a return of $1,947,168, given a target Gross Market Value (GMV) of $750,000 and target daily dollar volatility of $150,000
  - *Sharpe ratio* (measure of risk-adjusted returns): ~7
  - *Risk characteristics*: market neutral within 5%, less than 20% concentration in each token, the longest drawdown was 14 days
- *Basic economic principle*: use market microstructure and order book-based signals to predict price movement on a short-time timeframe (ranging from minutes to hours)
- *Holding period*: around 30 mins
  - We can liquidate the entire book within 15 mins
- *Brief description*: Every 5 to 15 mins we run our own proprietary statistical arbitrage model. We then find the optimal portfolio using our proprietary Bayesian mean-variance optimization, given current holdings, ideal holdings as predicted by the model, cost to execute, and risk in terms of the covariance matrix. We impose various constraints (e.g., market neutrality within 5%, less than 20% concentration in each token). Finally, we place trades through Binance.