Zhengze Zhang

zhangzhengze2018@gmail.com | github.com/Garthzzz | garthzzz.github.io

Education

Columbia University New York, USA Sep 2024 – Present

M.A. in Statistics (Advanced Machine Learning Track)

- Courses: Statistical Machine Learning, Natural Language Processing, Advanced Data Science, Financial Statistics, Linear Regression, Statistical Inference, Statistical Computing and Data Science.

University of California, Santa Barbara Santa Barbara, USA Sep 2018 – Jun 2023

B.S. in Mathematics & B.A. in Physics

- Mathematics: Real Analysis, Complex Analysis, Differential Geometry, Regression Analysis, Bayesian Analysis, Estimation Theory.
- Physics: Quantum Mechanics, Statistical Mechanics, Electromagnetism, Astrophysics, Theoretical Methods, Scientific Programming.

Research Experience

Columbia University New York, USA

Constrained Symbolic Regression for Financial Return Prediction Oct 2025 – Present

Advisors: Prof. Tian Zheng, Dr. Kara Lamb, Dr. Mikhail Smirnov

- Developed Constrained Symbolic Regression (CSR) framework integrating Campbell & Thompson (2008) sign constraints with task-aware wavelet denoising and stability selection for equity premium prediction using Goyal-Welch macroeconomic predictors.
- Achieved interpretable closed-form equations with economically meaningful coefficients; validated cross-domain generalizability on Treasury yield curve modeling (R^2 >95%), Fama-French factor timing, and HAR volatility forecasting.
- Conducted rigorous validation via permutation tests (Sharpe p=0.012), placebo constraints, sub-period stability analysis, and bootstrap confidence intervals; identified signal-to-noise ratio as the critical success factor across financial domains.

Cloud Microphysics Emulation: Next-Generation Symbolic Regression Algorithms Aug 2025 – Present

Advisor: Dr. Kara Lamb, LEAP Center, Columbia University

- Designed three-stage symbolic regression framework integrating E-WSINDy (Ensemble Weak-form SINDy), LaSR (Language-Augmented SR), and AI Feynman 2.0, achieving 50-1000x noise robustness improvement by transferring derivatives from noisy data to smooth test functions via weak-form integral formulations.
- Applied Hessian-based decomposition to 9-dimensional warm rain microphysics data, proving additive structure in autoconversion processes; integrated LARS variable selection reducing search space from 9 to 4-6 variables with 4-6x computational speedup.
- Implemented Buckingham π theorem for dimensional analysis, reducing feature space to 5-6 dimensionless parameters ensuring physical consistency; developed bootstrap ensemble methods for uncertainty quantification with physics-informed conservation constraints.

Weakly Supervised Tree Species Classification via Algorithm Optimization Sep 2025 – Present

Advisor: Prof. Tian Zheng, TZstats Convergence Lab

- Migrated full APL pipeline from TensorFlow to PyTorch with GPU acceleration, processing 14,400 image patches for tropical forest tree species identification using weakly supervised learning with imprecise point labels.
- Proposed algorithmic improvements: replaced K-means with graph-based clustering (kNN + Leiden) and density-based clustering (HDBSCAN) to capture irregular embedding manifolds; designed sigmoid-weighted loss using continuous relevance scores for robustness under noisy annotations.
- Analyzed APL's limitations on general image domains and proposed adaptation strategies including multi-scale backbones, transition from density regression to instance detection, and MIL/CAM-based weak supervision.

Treatment Discontinuation Modeling in CATIE Antipsychotic Trial May 2025 – Present

Advisor: Prof. Kiyohito Iigaya, Iigaya Lab

- Established reproducible EDA and preprocessing pipelines for CATIE clinical data, integrating demographics, PANSS scores, adverse events, and neurocognitive assessments across 1,400+ schizophrenia patients.
- Applied PCA and factor analysis to identify latent symptom dimensions predicting Phase I discontinuation due to inefficacy or intolerance; modeled discontinuation risk via regression linking symptom factors and treatment variables. Manuscript in preparation.

Lineage Heterogeneity and Founder-State Modeling in Monoclonal Gastruloids May 2025 – Present

Advisor: Prof. Bianca Dumitrascu, Morpho Lab

- Studied DNA Typewriter lineage-tracing methods (Regalado et al., 2025) to inform founder-state modeling in developmental biology.
- Processed gastruloid scRNA-seq data and lineage trees using PCA/UMAP; computed pairwise lineage distance matrices and quantified inter-gastruloid heterogeneity through lineage distance–fate similarity relationships.

Dissociating Effort, Perception, and Intensity in Visual Mental Imagery Jan 2025 – Present

Advisor: Dr. Alfredo Spagna, The Living Lab

- Contributed to the development and validation of ALFIE (Assessing Layered Forms of Imagery Experiences), a novel instrument dissociating imagery vividness into three aspects: cognitive effort, perceptual features, and subjective intensity across 11 dimensions.
- Conducted statistical analyses including 2×11 mixed ANOVA (modality × dimension), PCA with variance decomposition, and reliability assessment (Cronbach's α) on 100 participants' VVIQ and ALFIE responses.

- Applied ensemble clustering and co-clustering consensus methods to identify three robust imagery phenotypes; validated cluster stability via silhouette scores and bootstrap-adjusted Rand index. Manuscript in preparation.

Sentiment Dynamics in COVID-19 Tweets: Classical vs. Transformer Models Sep 2024 – Dec 2024

Advisor: Prof. Patrick Houlihan

- Analyzed 364,802 tweets comparing lexicon-based (VADER, LIWC) and transformer-based models (RoBERTa); used OLS and PCA to link sentiment scores with user features, demonstrating that contextual embeddings better capture subtle sentiment dynamics.

University of California, Santa Barbara Santa Barbara, USA

Photometric Analysis and Classification of Supernova AT2023hpb Mar 2023 – Jun 2023

Advisor: Prof. Phillip Lubin, UCSB Experimental Cosmology Group

- Conducted deep observation of AT2023hpb using Las Cumbres Observatory global telescope network; processed raw images via AstroArt 8 and Atlas; analyzed signal-to-noise ratios and plotted light curves in Python to infer Type II classification.

Panel Regression of Macroeconomic Indicators on U.S. Stock Returns Jun 2022 – Mar 2023

Advisor: Prof. Saad Mouti

- Reviewed literature on macroeconomic drivers of stock performance; applied robust and panel regression models in Python to evaluate predictive significance of inflation, interest rates, and GDP growth on cross-sectional equity returns.

Professional Experience

SDIC Securities Co., Ltd. Shanghai, China Sep 2023 – Apr 2024

Derivatives Analyst Intern, Equity Derivatives Desk

- Backtested options and futures trading strategies using Python (NumPy, pandas) and Excel VBA, analyzing historical P&L, Sharpe ratios, and maximum drawdowns to optimize entry/exit signals and improve risk-adjusted returns.
- Developed automated daily reporting system integrating real-time market data feeds with position tracking, reducing manual processing time by 60% and ensuring accurate Greeks calculation for portfolio risk management.
- Monitored trading positions across equity index options and stock index futures; proposed volatility arbitrage strategy improvements based on implied vs. realized volatility analysis.
- Collaborated with senior traders on structured product pricing and hedging strategies; assisted in client presentation materials for OTC derivatives solutions.

Taxpanda Inc. New York, USA Jun 2022 – Aug 2022

Data Analyst Intern

- Cleaned and processed large-scale SAP-extracted financial datasets ; performed data quality checks and reconciliation to ensure accuracy for tax compliance reporting.
- Designed interactive Tableau dashboards for executive reporting, visualizing key metrics including revenue trends, expense breakdowns, and tax liability projections across multiple client portfolios.
- Collaborated with domain experts and clients on data collection methodology; documented ETL pipelines and created user guides for dashboard interpretation.

Test & Certifications

- **GRE General Test**: 332 (Verbal 162 + Quant 170) Jan 2023
- **Bloomberg**: Market Concepts (BMC); Finance Fundamentals (BFF); Spreadsheet Analysis (BQL); ESG Sep – Dec 2025
- **CITI Program**: FDA-Regulated Research; Human Subjects Protection Biomed Feb 2025
- **IBM on Coursera**: Python for Data Science, AI & Development; Databases and SQL for Data Science with Python Nov 2023
- **Google on Coursera**: Data Analysis with R Programming Nov 2023

Activities

- Guest Lecturer, Columbia STAT 5205 Linear Regression Models Feb 2025
- Research Affiliate, Columbia LEAP Center Aug 2025 – Present
- Attendee, Columbia DSI ML/AI Seminar; LEAP Lecture in Climate Data Science Fall 2025
- Captain, CCST Future Soccer Team Jan – Jun 2025
- Committee Member, Columbia Chinese Soccer Team; Penn Cup Champion Jan 2025 – Present
- Vice Captain, UCSB Chinese Soccer Team; IMLeagues Championship Jun 2021 – Jun 2023

Skills

Programming: Python (NumPy, pandas, scikit-learn, PyTorch, SciPy), LaTeX, Git, SQL, Excel VBA, R.

Machine Learning & Analysis: Supervised Learning (Logistic Regression, XGBoost, Elastic Net, Cox Regression), Dimensionality Reduction & Clustering (PCA, Factor Analysis, UMAP, t-SNE, K-means, Leiden, Consensus Clustering), Time-Series & Causal Analysis (ARIMA, VAR, Granger Causality, DoWhy), Deep Learning (CNN, Grad-CAM, SHAP, Neural ODE, Symbolic Regression via PySR/AI-Feynman).

Finance: Options Pricing (Black-Scholes, Greeks), Portfolio Optimization, Risk Management, Backtesting Frameworks.

Languages: Mandarin (native), English (near-native; lived and studied in the U.S. for seven years).