

# THOMAS OU

(203) 243-2852 | thomasou@sas.upenn.edu | thomasou.com

## EDUCATION

### University of Pennsylvania

May 2027

BA in Mathematics, Minor in Statistics; Accelerated MSE in Computer Science

**Coursework:** Machine Learning, Big Data Analytics, Game Theory, Statistical Modeling, Cryptography

## EXPERIENCE

### Research Assistant – Defense Innovation & Policy, Penn Center for Undergraduate Research Jul 2025 – Present

- Developing end-to-end Python OCR pipeline using Tesseract and custom preprocessing to digitize 50+ years of historical U.S. military appropriations records from archival documents for Prof. Michael C. Horowitz's defense innovation research
- Designing PostgreSQL database schema with normalized tables and indexing strategies; implementing NLP system for automated document classification, named entity extraction, and metadata tagging across unstructured historical text
- Building data validation and quality assurance workflows to ensure accuracy of digitized records; creating Python scripts for batch processing and error handling across 10TB+ document corpus

### Research Assistant – Computational Physics, Princeton Plasma Physics Lab Jul 2024 – Aug 2024

- Implemented Monte Carlo simulations in Python modeling charged particle trajectories in tokamak fusion reactors under varying magnetic field configurations; developed statistical models with scikit-learn achieving  $R^2 > 0.90$  prediction accuracy
- Automated data processing pipeline with Pandas and multiprocessing handling 100GB+ simulation output; reduced analysis time by 75% through parallel processing workflows enabling faster reactor design iteration
- Created Matplotlib visualization dashboards for interdisciplinary research team; validated simulation results against experimental data with senior researchers

### Data Engineering & Analytics Intern, Flushing CPA Tax Center Feb 2024 – May 2024

- Built end-to-end ETL pipelines using Python and SQL to automate ingestion and processing of 50,000+ financial records from mult. data sources; implemented data quality checks and validation rules improving tax projection accuracy by 15%
- Developed scikit-learn regression models with feature engineering to forecast quarterly tax liabilities for small business clients achieving 92% prediction accuracy; created automated reporting system generating client-specific tax estimates
- Designed Excel VBA macros for automated compliance anomaly detection in client tax returns; implemented rule-based system flagging high-risk discrepancies and reducing manual review time by 40% across 200+ client accounts

## PROJECTS

### Predictive Analysis of MMA Fights using Ensemble Learning Aug 2025 – Present

- Engineered differential-based feature extraction using Pandas comparing 183+ fighter attributes across 10,000+ UFC fights; designed methodology capturing relative advantages in striking, grappling, and physical characteristics
- Benchmarked 7 scikit-learn supervised learning algorithms including Gradient Boosting, Random Forest, SVM, KNN, MLP, Decision Tree, and Logistic Regression; performed feature importance analysis identifying key predictive attributes
- Built interactive Python prediction system generating probabilistic confidence scores for fight outcomes; implemented cross-validation with NumPy and statistical evaluation for model robustness

### High-Performance OCR & Data Engineering Pipeline for Defense Research Jul 2025 – Present

- Developed computer vision pipeline combining Tesseract OCR with custom TensorFlow CNN-based preprocessing achieving 98.5% character accuracy on degraded historical documents spanning 50+ years using OpenCV for image processing
- Built ETL system with Apache Airflow orchestrating parallel Python processing of 10TB+ document corpus; implemented fuzzy matching algorithms in PostgreSQL for entity resolution across inconsistent data sources
- Designed Neo4j graph database modeling complex relationships between 100K+ entities; applied PageRank and community detection algorithms using Python for influence analysis with Elasticsearch indexing

### Bayesian Opponent Modeling Engine for Poker AI Feb 2024 – Jul 2025

- Engineered Monte Carlo Counterfactual Regret Minimization (MCCFR) solver in Rust implementing algorithms from Pluribus research; developed fastest open-source hand evaluator outperforming industry-standard Cactus Kev benchmarks through bitwise optimization
- Implemented optimal transport clustering using Sinkhorn iteration and Earth Mover's Distance for strategic abstraction; processed 3.1B+ poker situations through isomorphic canonicalization achieving 4-23x memory reduction
- Designed parallel training pipeline processing 268M game trees with Rayon-based batching and graceful interruption; designed generic trait-based solver extensible to imperfect-information games

## TECHNICAL SKILLS

**Languages:** Python, R, C++, Java, SQL, JavaScript, OCaml, HTML/CSS

**Frameworks & Libraries:** PyTorch, TensorFlow, Scikit-learn, OpenCV, NumPy, SciPy, Pandas, MCMC, PCA