# Raghu Ram Sattanapalle

347-873-2177 | sattanapalle.r@northeastern.edu | raghuramsattanapalle | RaghuRamSatt | raghuramsatt.github.io

## Education

**Northeastern University** — Boston, MA
*Master of Science in Data Science (GPA: 3.833 / 4.00)* — *Jan. 2023 – Expected May 2025*

**New York University** — New York, NY
*Master of Science in Mechanical Engineering* — *Sep. 2016 – May 2018*

## Experience

**Head Teaching Assistant: Unsupervised Machine Learning & Data Mining** — Jan. 2025 – Present
*Northeastern University* — *Boston, MA*

- Provided detailed guidance to **130+ students** in **unsupervised learning algorithms** and **Python data processing** during weekly office hours.
- Coordinate a team of **6 teaching assistants**, managing **6–8 hours of weekly office hours** (online and in-person), ensuring smooth grading workflows and course operations.

**Engineering Data Scientist (Co-op)** — Jan. 2024 – June 2024
*Veeco Instruments* — *San Jose, CA*

- Unified **10+ years of manufacturing data** from multiple storage drives using **Python and SQL pipelines** to build a robust dataset for Machine Learning driven wafer analysis.
- Developed **convolutional neural networks (CNNs)** using **TensorFlow and PyTorch** to predict boron wafer resistance, achieving a **6% average error rate** by employing data augmentation, group normalization, and various CNN architectures.
- Built a **standalone wafer visualization tool** (Python) to identify trends and anomalies, improving **manufacturing efficiency by 30%** through **data-driven** process control.
- Collaborated with senior engineers on **code reviews** (GitHub) and **refactoring**, ensuring robust and **maintainable ML pipelines** aligned with software engineering best practices.

**PhD Candidate / Researcher** — Sept. 2018 – Aug. 2022
*NYU Dynamical Systems Laboratory* — *Brooklyn, NY*

- Analyzed the **MIMIC dataset** (**300M+ clinical observations**) using **SQL** for data extraction and **supervised ML models** for analysis, achieving **90% accuracy** in predicting **ICU mortality** in collaboration with **NYU Langone clinicians**.
- Led a **causal inference** study on mass shootings, media coverage, and firearm acquisition using **time series analysis** (ARIMA, Tramo/Seats) and **transfer entropy**, resulting in a ***Nature Human Behaviour*** publication.
- Developed a mathematical model using **stochastic differential equations** in **MATLAB** and **Mathematica** to examine collective behavior, contributing to a publication in *Flow* (Cambridge Core).
- Secured a **$2.1M NSF grant** as co-author to investigate the U.S. firearm ecosystem, utilizing **Tableau** for visually compelling preliminary results included in the proposal.

## Technical Skills

**Programming Languages:** Python (Pandas, NumPy), SQL, C++, Java, JavaScript, Scala, R, MATLAB
**Machine Learning:** Boosting/Bagging Models (XGBoost, Random Forest), KMeans, Clustering, Supervised Learning, Unsupervised Learning, TensorFlow, PyTorch, Deep Learning, NLP, Generative AI, RAG, Prompt Engineering, Text Embeddings
**Data Engineering & Cloud:** AWS (EMR, S3, EC2, SageMaker), Google Cloud (GCP), Hadoop, Spark, BigQuery, Hive
**Databases:** MySQL, PostgreSQL, MongoDB (NoSQL), Snowflake
**Visualization:** Tableau, Power BI, Matplotlib, Seaborn, Plotly, ggplot2, D3.js
**Tools:** Git, Docker, Kubernetes, Linux, CI/CD, Bash, Postman

## Projects

**AI Agent for Data Analysis** | *Python, GPT-4, Claude-3, Llama 3, LangChain, Streamlit* — Mar. 2025 – Present

- Building an **open-source conversational AI agent** for automating end-to-end data analysis tasks, enabling users to ask natural-language questions and receive **executable Python code** for statistical analysis and business insights.
- Leading GPT-4 integration by developing **specialized prompt-engineering strategies** and **context management systems** that transform business questions into data science code, with focus on maintaining conversation history across analysis sessions.
- Designing **comprehensive evaluation frameworks** to benchmark model performance across proprietary and open-source models using both standard benchmarks (**DS-1000**) and real-world business datasets, ensuring optimal accuracy and efficiency.

**FraudFusion: Synthetic Fraud Data Generation** | *Python, PyTorch, XGBoost, Diffusion Models* — Jan. 2025 – Present

- Developed **diffusion models** to generate synthetic **credit card fraud data**, addressing the challenge of extreme data imbalance (0.5% fraud rate) to improve fraud detection capability.
- Implemented **specialized feature engineering** techniques and **custom loss functions** to capture complex fraud patterns, iteratively improving through multiple model versions to achieve **high-quality synthetic data generation**.
- Improved fraud detection performance of **XGBoost classifiers** by increasing detection rate from 82% to ≈90%, with minimal impact on false positive rates.

**Trading at the Close: Predict US Stock Movements** | *Python, CNN, Time Series Analysis* — Oct. 2023 – Dec. 2023

- Developed **ML-driven trading models** on **NASDAQ equity data** to forecast closing auction prices, achieving a **top 20% ranking in a Kaggle competition**.
- Engineered advanced **time series features** (lag variables, rolling windows) to capture temporal dependencies, driving an **18% improvement** in prediction accuracy.
- Compared and fine-tuned **LightGBM, XGBoost, and Convolutional Neural Networks** via **cross-validation** and **hyperparameter tuning**, significantly **reducing Mean Absolute Error (MAE)** in forecasting closing auction prices.