Raghu Ram Sattanapalle

347-873-2177 | raghurams95@gmail.com | raghuramsattanapalle | RaghuRamSatt | raghurams.com

Education

Northeastern University — Boston, MA
Master of Science in Data Science (GPA: 3.867 / 4.00) — *Jan. 2023 – May 2025*

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

Experience

ML Engineer — June 2025 – Present
GrantX — *Remote*

- Implemented **reasoning-based document segmentation** using **PageIndex hierarchical tree structures** to extract **50+ structured fields** from funding documents without traditional chunking or vector databases.
- Architected **hybrid search engine** combining **semantic search (ELSER), vector embeddings, and BM25 retrieval** with **LLM-based query generation** using Gemini models, achieving **90.3% MRR** and **75.6% precision@10**.
- Built **data ingestion pipeline** processing **IRS Form 990-PF filings** from **756,000+ private foundations**, parsing XML tax returns to extract organizational metadata, grant histories, and financials with **cursor-based pagination** and retry logic.
- Deployed production system on **GCP/Kubernetes** serving **25K+ federal and private opportunities** with **91ms p50 latency, 99.9% uptime**; optimized enrichment pipeline by **4-6x** through batch database fetching.

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: TensorFlow, PyTorch, Deep Learning, NLP, RAG, LangChain, Text Embeddings, Ranking/Retrieval Models, XGBoost, Anomaly Detection, Fraud Detection Models, Model Deployment, Model inference optimization, A/B Testing
Data Engineering & Cloud: AWS (EMR, S3, EC2, SageMaker), GCP, Azure, Hadoop, Spark, BigQuery, Hive
Databases: MySQL, PostgreSQL, MongoDB (NoSQL), Snowflake, Elasticsearch, Supabase, Vector DBs (Pinecone, FAISS, Qdrant)
Visualization: Tableau, Power BI, Matplotlib, Seaborn, Plotly, ggplot2, D3.js
Tools: Git, Docker, Kubernetes, Linux, CI/CD, Bash, Postman

Projects

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

- Developed **diffusion models** to generate synthetic **credit card fraud data**, addressing extreme data imbalance (0.5% fraud rate) to improve **anomaly detection** and fraud identification 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.

Scalable Music Similarity Analysis with Spark | *Spark, Scala, AWS, Docker* — Oct. 2024 – Dec. 2024

- Developed a music similarity system using **Spark** on the **Million Song Dataset** (**10K songs, 110K users**), integrating **K-Means** for audio feature analysis with **collaborative filtering** based on user listening history for song recommendations.
- Engineered **two K-Means parallelization strategies** in **Spark** to improve efficiency, achieving a **4.58x speedup** on **AWS EMR**, and implemented **H-V partitioning** with a **sparse matrix** for efficient similarity computation.
- Evaluated **K-Means** using **Silhouette Score** and **Davies-Bouldin Index** and applied **user-based rating normalization** to enhance **collaborative filtering**, focusing on model accuracy and performance.