

Sam Voisin

SUMMARY STATEMENT

Seasoned Machine Learning Engineer and Software Developer with a strong background in creating robust data pipelines, scalable architectures, and efficient algorithms. Proven expertise in translating advanced data science and machine learning techniques into production-ready software solutions.

SKILLS

- **Machine Learning:** Adept in developing and deploying modern ML applications; Advanced PyTorch and Scikit-learn user. Experience with LlamaIndex and LangChain.
- **AI/ML at Scale:** Proven experience deploying LLMs and RAG systems to customers.
- **Statistical Analysis:** Expertise in statistical inference and predictive modeling.
- **Advanced Python Development:** Advanced Python developer with 7+ years of production experience.
- **Database Management:** Skilled with SQL and NoSQL databases and PySpark.
- **REST API Proficiency:** Deep understanding of REST API principles with hands-on experience deploying FastAPI applications.
- **Cloud Infrastructure:** Proficient with AWS services including EC2, RDS, S3, DynamoDB, etc.
- **Linux and Docker:** Advanced user of Linux and Docker containerization.
- **DevOps and Agile Practices:** Proficient with DevOps tools, Git version control, CI/CD pipelines, and Agile development methodologies.
- **Communication:** Strong communication skills, facilitating effective collaboration with both technical and non-technical stakeholders.

PROFESSIONAL EXPERIENCE

Tradewind Data Science, Chicago, IL — *Senior Data Scientist*

OCTOBER 2023 - CURRENT

- Employ state-of-the-art LLMs/LMMs for clients' generative AI needs.
- Enhance model performance through retrieval augmented generation (RAG).
- Implement modern ML algorithms (DNNs, XGBoost, Random Forest) for NLP tasks like sentiment analysis, entity extraction, and text classification.
- Construct ETL pipelines using Python, SQL, and PySpark on large-scale data clusters.
- Established company-wide DevOps program.

Infinia ML, Durham, NC — *Data Scientist*

MARCH 2022 - OCTOBER 2023

- Led initiative for integrating LLMs in the data science technology stack.
- Developed production-grade NLP pipelines for unstructured document analysis.
- Implemented asynchronous event loop for a 6x speedup in document processing time.
- Worked across functional teams to rapidly translate customer needs into solutions.

Geometric Data Analytics, Durham, NC — *Data Scientist*

JUNE 2020 - MARCH 2022

- Developed novel algorithms resulting in multiple peer reviewed publications.
- Designed tiered modeling system for anomaly detection in global shipping data.
- Planned software architecture and converted research code to deployable libraries.
- Worked with event data streamed from Kafka topics.

Ally Financial Services, Charlotte, NC — *Analyst*

JANUARY 2015 - JUNE 2018

- Automated data gathering and processing for substantial lead time and error reduction.
- Analyzed exchange data and business metric relationships to mitigate risk.

EDUCATION

Duke University, Trinity College of Arts and Sciences — *Master of Statistical Science*

Clemson University, College of Business and Behavioral Science — *Bachelor of Science in Financial Management*

PUBLICATIONS & AWARDS

Voisin, S., Hineman, J., Polly, J. B., Koplik, G., Ball, K., Bendich, P., D'Addezio, J., Jacobs, G. A., & Özgökmen, T. (2022). Topological feature tracking for Submesoscale eddies. *Geophysical Research Letters*, 49(20). <https://doi.org/10.1029/2022gl099416>

G. Koplik et al., "Topological Simplification of Signals for Inference and Approximate Reconstruction," 2023 IEEE Aerospace Conference, Big Sky, MT, USA, 2023, pp. 1-11, doi: 10.1109/AERO55745.2023.10115654.

Ball, K., Hineman, J., Voisin, S., Koplik, G., Bendich, P. (2021). Automation is All You Need: Faster Earth Systems Models with AI/ML.

First place at U of SC Big Data Health Science Conference 2020 case study competition.

OPEN-SOURCE PROJECTS

GestuReMote: Computer vision based system which provides a gesture-based interface to a personal computer.