Emily Marie Williams

Data Scientist with 15+ Years Experience | AI & Machine Learning Expert | Predictive Modeling & Data Analysis Specialist
San Diego, CA 92117 | emw314159@gmail.com | (805) 637-0856 | GitHub | LinkedIn | Substack

Skills

Programming & Tools: Python | R | SQL | NoSQL | Apache Spark | Neo4j | Java | JavaScript | Django | MySQL | PostgreSQL | Oracle | Bash | Perl | C++ | Git | Jupyter | Claude Code

Python ML/AI Stack: Keras | TensorFlow | scikit-learn | PySpark | Pipeline Creation | statsmodels | NumPy | SciPy | Pandas | NetworkX

Data Science & Machine Learning: Data Product Deployment | Machine Learning | Statistical Modeling | Natural Language Processing (NLP) | Time Series Forecasting | Clustering & Topic Modeling | Model Validation | Feature Engineering | Predictive Analytics | Translating Business Needs Into Data Science Solutions | Interactive Dashboards & Web-Based Reporting/Visualization

Platforms & Infrastructure: AWS (EC2, S3, Lambda) | Databricks (Admin & Deployment) | ETL | MLflow | Prefect | Airflow | REST API Development | Web Dashboards (Interactive UI) | Cloud-Based ML Workflows | Linux System Administration | Apache & Tomcat | Agile/Scrum Collaboration

Agentic AI: Model Context Protocol (MCP) | Pydantic | Skill Development | OpenClaw | Claude Code

Communication & Business: Cross-Functional Team Collaboration | Translating Technical Results for Non-Technical Stakeholders | Executive/Stakeholder Reporting | Requirements-to-Delivery Ownership | Technical & Business Writing

Critical Thinking: Systems Thinking | Critical Analysis of Data & Results | Thoughtful Decision-Making

Experience

Senior Data Scientist | Audentis Systems
November 2023 – April 2026

- Founded and operated a self-funded systematic foreign-exchange trading company, personally owning the full technology stack from data ingestion through machine learning research through production deployment and live trade execution
- Created deep learning algorithms (Keras, TensorFlow, scikit-learn, PySpark) to forecast currency price movements from historical time-series data, and built the trade-execution software that acted on them via the broker's API
- Engineered production data systems for time-series and trade-outcome data management (MySQL, MongoDB, Django, Spark, InfluxDB)
- Modernized the trading system's legacy notebook-based ML pipeline into a production grade, version-controlled package — MLflow experiment tracking and model registry, Prefect workflow orchestration, and DVC data/model versioning — using Claude Code as a primary AI-assisted development tool across a multi-repository refactor and bug remediation effort
- Strengthened the system's time-series validation methodology with purged chronological train/validation/test splitting to prevent lookahead bias, ACF/PACF- and ADF/KPSS-justified feature windows, baseline-model benchmarking, and multiple-comparisons corrected (Benjamini-Hochberg) significance testing across currency pairs and model configurations
- Generalized the trading system's forecasting engine into Omen, a domain-agnostic, five-layer agentic time-series forecasting toolkit exposed via the Model Context Protocol (MCP); released it as an open-source package and documented it in a published book, *Agentic Time Series Forecasting*

Senior Bioinformatics Scientist | Epic Sciences
August 2021 – October 2023

- Led the design and implementation of a scalable AI pipeline using Apache Spark to identify subtypes of metastatic breast cancer from high-dimensional genomic data
- Developed statistical models to establish optimal cutoff thresholds for a genomic assay used to determine cancer metastasis status, directly contributing to improved diagnostic accuracy
- Conducted comprehensive validation studies, including sensitivity/specificity analysis and cross-validation, to support regulatory approval of the assay by agencies such as the FDA
- Spearheaded the internal adoption of Databricks as a centralized platform for data analysis, machine learning, and ETL; providing training and mentorship to cross-functional teams
- Served as primary administrator for the company's Databricks and AWS environments, managing cloud infrastructure, optimizing compute resources, and ensuring data security and compliance

Data Scientist | Best Buy Health
February 2020 – July 2021

- Developed semi-supervised AI topic modeling solutions to analyze large-scale call center transcripts, leveraging Python, Apache Spark, and advanced NLP techniques to uncover key customer concerns and product service trends
- Built end-to-end pipelines to support this AI work, including data preprocessing, model performance refinement and data product deployment
- Designed and deployed an interactive, web-based dashboard allowing non-technical stakeholders to explore topic model results in real time, improving operational decision-making across customer service and product teams
- Collaborated cross-functionally with engineering, UX, and business teams to align data science solutions with strategic goals and ensure seamless integration of insights into existing workflows

Data Scientist & Founder | Whole-Systems Enterprises
July 2017 – February 2020

- Founded and led a data science consultancy delivering custom AI, machine learning, and bioinformatics solutions across healthcare, psychology, and finance sectors, while also building a proprietary portfolio of AI technologies
- Designed and delivered a full-stack bioinformatics search tool for genome, transcriptome, and proteome data to Pfenex Biopharmaceuticals, including a user-friendly graphical interface for internal scientists
- Collaborated with Grr-ithm, Inc. to create a dynamic mechanical simulation of thoracic cage motion under various load conditions; also developed a prototype medical expert system to provide decision support based on this model
- Conducted statistical modeling and analysis for the American Institute for Behavioral Research and Technology
- Built a currency trading algorithm leveraging deep learning and time series forecasting techniques

Bioinformatics Scientist | Thermo Fisher Scientific
November 2007 – July 2017

- Designed and implemented automated pipelines for next-generation sequencing (NGS) data analysis using Python and R
- Developed statistical and machine learning models using CRISPR/Cas9 and siRNA sequence features, thereby increasing product potency and specificity
- Created databases and web applications to support RNAi and gene editing product design, including high-performance siRNA and CRISPR/Cas9 design tools having API integration
- Led full-stack development of bioinformatics and LIMS tools using Java, Python (Django), JavaScript (AngularJS, jQuery), and SQL/NoSQL databases to support R&D and commercial workflows
- Conducted customer segmentation analysis on market research data using Python-based clustering
- Maintained legacy systems and Linux-based production environments supporting bioinformatics tools and databases

Research Analyst 3 | State of Oregon
February 2006 – September 2007

- Conducted statistical analysis of client population characteristics (SPSS, SAS, R)
- Performed time series analysis and statistical forecasting to predict future client populations
- Developed a system dynamics model of Oregon's Medicaid population (R, MATLAB)

Patents & Publications

- Co-inventor on three U.S. patent applications ("Methods and Systems for Designing Gene Panels," 15/611,233; 62/344,078; 62/510,906), built on a Neo4j-based graph database identifying gene network and sub-network relationships to specific diseases
- Caro GD, Lam ET, Bourdon D, Blankfard M, Dharajiya N, Slade M, Williams E, Zhang D, Wenstrup R, Schwartzberg L. A novel liquid biopsy assay for detection of ERBB2 (HER2) amplification in circulating tumor cells (CTCs). J Circ Biomark. 2024 Oct 4;13:27-35. doi: 10.33393/jcb.2024.3046. PMID: 39377016; PMCID: PMC11456801.
- Williams, E. (2026). Agentic Time Series Forecasting (Kindle ed.). https://www.amazon.com/dp/B0HBX8HGRC

Education

- **B.S. Mechanical Engineering, Minor in Music** | Oregon State University
- **Graduate-Level Coursework (*Professional Development*), Applied Mathematics & Controls Engineering** | Oregon State University | 27 units — Linear Systems, Linear & Non-Linear Control, Numerical Methods, Fourier Analysis
- **Graduate-Level Coursework (*Professional Development*), Statistics** | Texas A&M University | 12 units — Mathematical Statistics, Statistical Computation, Statistical Methods