# Michael David Johnson

AI Engineer & Senior Data Scientist | GenAI, LLMs & Production ML

Orange County, CA | (949) 340-4330 | mchl.dvd.jhnsn@gmail.com

linkedin.com/in/data-arts-data-science | mdjhnsn.github.io

## PROFESSIONAL SUMMARY

**10-year AI/ML veteran** with production experience building **RAG systems, LLM applications, and enterprise-grade ML pipelines**. Proven track record deploying models serving **100M+ users** while maintaining reliability and governance. Recent fintech experience includes architecting data infrastructure and developing **risk scoring models for fraud detection**. Expert in **Python, PyTorch, Deep Learning, and GenAI-assisted development** (Cursor, Copilot, LLM APIs). MS Statistics and MA Economics background provides rigorous framework for building AI systems that drive measurable business outcomes in regulated financial environments.

## CORE COMPETENCIES

**GenAI & LLMs:** RAG Architecture, Prompt Engineering, LLM APIs (OpenAI, Claude), Vector Embeddings, Document Intelligence, Agentic Workflows

**ML & Deep Learning:** PyTorch, Keras/TensorFlow, Scikit-learn, XGBoost/CatBoost, NLP, Sentence-Transformers, MLFlow, Ray

**Production & MLOps:** Terraform (IaC), CI/CD Pipelines, Model Lifecycle Management, Feature Pipelines, Databricks, AWS (S3, Glue, Athena, DMS)

**AI Development Tools:** Cursor, GitHub Copilot, Claude Code, LangChain, API Development (FastAPI)

**Financial AI Applications:** Risk Scoring, Fraud Detection, Transaction Monitoring, Customer Segmentation, Churn Prediction

**Technical Stack:** Python (Expert), SQL (Expert), R, PySpark, Snowflake, Tableau, Streamlit, Plotly

## PROFESSIONAL EXPERIENCE

**Senior Data Scientist / Infrastructure Lead | Fancy | Irvine, CA | Aug 2025 – Jan 2026**

*High-growth fintech platform serving 2,500+ merchant partners.*

• Built Merchant Risk Scoring models to identify high-value vs. high-risk partners, enabling fraud detection and anomalous transaction pattern identification across 2,500+ merchants.

• Developed Customer Loyalty Scoring using ML combining behavioral signals (visit frequency, spend patterns) for personalized targeting and fraud-resilient marketing campaigns.

• Architected production-grade Data Lakehouse (AWS DMS, S3, Glue, Athena) replicating 8 schemas and 124 tables—establishing single source of truth for ML feature stores. [See: mdjhnsn.github.io/portfolio.html#lakehouse]

• Implemented Terraform for Infrastructure as Code, enabling repeatable, auditable data resource provisioning across Dev, Staging, and Production environments.

**Senior Data Scientist | Pluto TV (Paramount) | Remote | Oct 2021 – July 2025**

*Lead data scientist for Paramount's ad-supported streaming platform serving 100M+ users.*

• Production ML at Scale: Architected churn prediction models using Ray on Databricks for 100M+ users, with full model lifecycle management (training, deployment, monitoring), driving 15% increase in retention ROI.

• Spearheaded GenAI Applications for streaming user preference analytics, including prototyping LLM-powered content recommendation systems.

• Built Bayesian A/B testing framework (PyMC) with full uncertainty quantification, shortening SDLC by 20%. [See: mdjhnsn.github.io/portfolio.html#ab-testing]

• Owned revenue forecasting (Prophet, PySpark) across 30+ device types, reducing MAPE by 5%—primary input for executive financial planning. [See: mdjhnsn.github.io/portfolio.html#revenue-forecasting]

• Platform Governance: Administered Databricks workspace processing 50TB+ monthly (hundreds of millions of events), ensuring stability and compliance for 10-person data science team.

**Data Scientist | Evolus Inc | Remote | Oct 2020 – Sep 2021**

• Developed user preference segmentation using DBSCAN clustering, driving 12% improvement in trial conversion rates.

• Automated reporting pipelines via Apache Airflow, reducing time-to-insight from 48 hours to 4 hours.

• Managed analytics for $10M monthly marketing budgets, ensuring spend efficiency across digital and print channels.

**Data Scientist | Aviana Global Technologies | Brea, CA | Feb 2019 – Sep 2020**

• Deployed XGBoost churn detection models for 5 enterprise clients, achieving 18% reduction in churn through anomaly detection and risk scoring.

• Engineered real-time ML pipelines for 10M+ semi-structured records using NVIDIA GPUs, enabling low-latency predictive analytics.

• Built custom Streamlit/Plotly interfaces, enabling non-technical stakeholders to access real-time insights self-serve.

**Data Scientist | MINDBODY Software | Irvine, CA | Apr 2018 – Jan 2019**

• Built Deep Learning retention models (Keras) with 90% accuracy—the definitive signal for customer retention intervention programs.

• Validated $2M in strategic business investments through predictive modeling using R and AWS Redshift.

**Content Analyst | Hanover Research Council | Washington, DC | Jul 2013 – Mar 2018**

• Led quantitative analysis projects using R and SQL; delivered research insights shaping $5M+ in client marketing campaigns.

## GENAI & ML PROJECTS

**RAG Document Ranking Application:** Built Retrieval-Augmented Generation tool using vector embeddings to rank 100K+ documents, reducing manual review time by 30 hours/month—directly applicable to document intelligence and loan processing automation. [See: mdjhnsn.github.io/portfolio.html#llm-entity]

**LLM-based Entity Resolution (Identity Matching):** Achieved 98% accuracy in duplicate detection using Sentence-Transformers and PyTorch vector embeddings—applicable to KYC and identity verification workflows. [See: mdjhnsn.github.io/portfolio.html#llm-entity]

**Predictive Model Optimization:** Improved Census microdata model $R^2$ from 0.65 to 0.82 using CatBoost with extensive feature engineering.

## EDUCATION

**MS Statistics** — Texas A&M University

**MA Economics** — George Washington University

**BA Literature** — UC Santa Cruz