San Francisco, CA
serigne_diaw@yahoo.com
(707) 853-8157

Serigne Diaw

Data Scientist

Portfolio
LinkedIn

Education

Master of Science in Data Science, *University of San Francisco* Jun 2025

Relevant Coursework: Advanced Machine Learning, Distributed Data Systems, Relational Databases, Probability & Statistics

Bachelor of Science in Computational Cognitive Science, *University of California, Davis* Jun 2022

Relevant Coursework: Linear Algebra, Game Theory, Computational Linguistics, Computational Theory, Data Structures

Experience

Data Scientist **Oct 2024 — Present**
Qventus Inc. *Mountain View, CA*

- As an intern, applying natural language processing and unsupervised learning techniques to group clinical procedure orders.
- Using Python for data preprocessing, feature engineering, and building models; SQL for large-scale data exploration and extraction of procedure orders.
- Analyzing the impact of procedure order groupings on clinical predictions like patient discharge timelines and post-acute care needs, improving model accuracy and operational insights.
- Developed an XGBoost model to predict classes for procedure orders, achieving a macro accuracy of 85%.

Data Analyst **Jun 2022 — Jun 2024**
UC Davis Center for Neuroscience *Davis, CA*

- Involved with all steps of the data collection and analysis pipeline for an EEG project that examined spatial memory in 8 intracranial patients.
- Applied knowledge of Python, MATLAB, Pandas, etc. to construct programs for the collection, cleaning and preprocessing of data from over 400 electrodes implanted in the brain, totaling 5000 trials.
- Implemented logistic regression to examine hippocampal-prefrontal cortex interactions during memory encoding and retrieval.

Projects

Football Player Tracking Github
OpenCV, scikit-learn, Roboflow, PyTorch

- Developed a computer vision system for real-time football analytics using YOLOv8 and Roboflow that tracks multiple objects including players, referees, goalkeepers, and the ball simultaneously.
- Optimized model performance using GPU acceleration through Google Colab integration.

Premier League Match Prediction Github
MCMC, Beautiful Soup, PostgreSQL, scikit-learn

- Developed a match prediction system using Markov Chain Monte Carlo (MCMC) methods to forecast Premier League soccer match outcomes.
- Engineered an automated ETL data pipeline that scrapes, processes, and stores match statistics from fbref.com into a PostgreSQL database.
- Achieved 55% prediction accuracy while providing granular probability estimates for match outcomes including expected goals.

arXiv Research Tool Github
Airflow, GCS, Prompt Engineering, MongoDB

- Built an automated ETL pipeline using Airflow that fetches AI papers from arXiv, validates content, stores PDFs in Google Cloud, and catalogs metadata in MongoDB.
- Integrated LLM-powered summarization to extract key research insights and relevance scoring against user topics.
- Developed a Streamlit frontend for keyword-based discovery of AI research with accessible summaries.

Technical Skills

Programming Languages	Python, SQL
Big Data & Machine Learning	Spark, MongoDB, Snowflake, PostgreSQL, PyTorch, Airflow, MLflow
Data Science & Misc. Technologies	A/B Testing, Time Series, OOP, Git, Docker

Publications

- "The hippocampus supports precise memory for public events regardless of their remoteness"