Nathaniel del Rosario

natdosana@gmail.com | linkedin.com/in/natdosan | natdosan.github.io

Education

University of California, Berkeley — Berkeley, CA
M.S. Data Science — *2026*
- Artificial Intelligence, ML Systems, Data Engineering, Reinforcement Learning, Experimentation & Causal Inference

University of California, San Diego — La Jolla, CA
B.S. Data Science — *March 2025*
- Graduate (cross enrollment): Computer Vision, Recommender Systems, Deep Learning
- Principles & Techniques of Data Science, Statistics, Relational Databases, Operating Systems, Cloud Computing, Scalable ML

Skills

Python, Pandas, NumPy, PyTorch, Sci-Kit Learn, Langchain, ChromaDB, HuggingFace, OpenCV, Dash/Plotly, Dask, Streamlit
AWS, Google Cloud, Docker, Kubernetes, Airflow, SSH, Snowflake, Databricks, BigQuery, PostgreSQL, Github, Jira, ArcGIS

Experience

American Express — June 2026 - August 2026
Data Science Intern - Global Decision Science — *New York*
- Summer 2026

CVS Health — May 2025 - August 2025
Data Science Intern - Analytics & Behavioral Change — *New York*
- Democratized knowledge base of 39 million members & 77,000 unique diagnoses by implementing an Agentic-RAG Text-to-SQL tool leveraging BigQuery, Langchain, Vertex AI, ChromaDB, Docker, Kubernetes, Datadog Eppo
- Integrated MVP into existing API, optimizing performance between 82% to 93% across selected metrics through grid search
- Investigated causality of 330 million medical equipment claims, identifying potential cost-saving opportunities of up to $21 million per successful rollout

Hacioglu Data Science Institute — September 2024 - March 2025
Undergraduate Research Assistant - Zhiting Hu's Group — *La Jolla*
- Ran compute v.s. scaling experiments utilizing Monte Carlo Tree Search improving by 10% on OSWorld Benchmark
- Integrated OSWorld support into the LLM-Reasoners Python library (2k Github Stars) by implementing a Self-Evaluation Reward Function and custom Search Algorithm Configurations based on current state and action trajectory

University of California, San Diego — September 2023 - March 2025
Instructional Assistant - Halicioglu Data Science Institute — *La Jolla*
- Beta Testing assignment and exam questions, hosting Office Hours for a data science course of over 500 students
- Updated deployment of course website using github pages & Docker supervised by under Suraj and Tiefenbruck
- Grading and hosting Office Hours for upper division data science course of over 700 students under Shannon Ellis, Sam Lau

Bio-Rad — June 2024 - August 2024
Data Science Intern - Clinical Diagnostics Group — *Pleasanton*
- Leveraged AWS RDS, Docker, PostgreSQL to deploy database, reduced storage usage by 35% through schema optimization
- Built ETL pipeline w/ Pandas, RestAPI to ensure 100% data integrity & improved consistency from 91% to 99.9%
- Utilized AWS EC2 to deploy web-app & unit tests utilizing Dask to achieve 5.1x / 80.1% speedup on data validation

San Diego Supercomputer Center — June 2023 - September 2023
Machine Learning Engineer Intern — *Remote*
- Designed Content-Based Filtering Recommender System utilizing Cosine and Jaccard similarity for baseline output
- Deployed Recommender System on AWS EC2, Lambda, achieving a design that scaled to process 200,000+ points
- Trained an RL agent using Stable Baselines and Q-Learning to improve recommendation quality after 100 iterations
- Utilized AWS S3, PostgreSQL for database queries & vectorized code to achieve 1.7x runtime speedup in feature engineering

Deloitte — February 2023 - June 2023
Data Science Fellow — *Remote*
- Cleaned data w/ 3000+ features, 1 billion observations using Dask, vectorized Pandas to decrease cleaning runtime by 20%
- Leveraged XGBoost, Lasso to identify 850 significant features, predict drug use in young adults with 81% accuracy
- Tuned Hyperparameters, class weighting to improve F1 score from .35 to .70 and identify 10 highest risk demographics

Chan Zuckerburg Biohub — June 2022 - January 2023
Data Science Intern - Data Science & Computational Biology Platform — *San Francisco*
- Built 9 interactive visualizations of CRISPR screen comparisons between 20000 features using Pandas/Dash/Plotly
- Improved data processing of a Nextflow data pipeline (16,000,000 data points) to minimize runtime by 10%
- Designed algorithms to compare across 30+ virus screens to yield insights in virus-host interactions using vectorized code
- Wrote documentation for 23 functions from scratch and improved 3K+ codebase readability using Readthedocs