**Nealson Setiawan**

Los Angeles, CA | 747-272-5097 | [nealsons@usc.edu](mailto:nealsons@usc.edu) | [www.linkedin.com/in/nealson‑setiawan‑4a1700252](https://www.linkedin.com/in/nealson-setiawan-4a1700252) | <https://github.com/nealsonS>

**education**

**University of Southern California** Los Angeles, CA

**M.S, Masters of Science, Applied Data Science** August 2023 - May 2025

* Relevant course work: Computer Vision, Natural Language Processing, Generative AI, Data Mining with Apache Spark, PyTorch, Keras / Tensorflow, Recommender Systems, LangChain, SQL, MongoDB, Firebase, DynamoDB, Python

**University of California Santa Barbara** Santa Barbara, CA

**B.S, Bachelors of Science, Statistics & Data Science** September 2021 - June 2023

* Relevant course work: Statistical Machine Learning, Statistical Data Science, Time Series Analysis, Time Series Forecasting, Data Visualization, Algorithms and Data Structures, Stochastic Processes, Probability Theory, Regression Analysis, Design of Statistical Experiments, Intermediate Python, R, SQL, SAS

**experience**

**Kristal.AI** Singapore, Singapore

**Machine Learning Intern** July 2024 – September 2024

* Developed an advanced RAG Q&A chatbot for confidential document processing, routing queries to Text2SQL or self-querying retrievers based on content type, leveraging quantized LLMs from Ollama and HuggingFace
* Parsed and stored documents in PostgreSQL and Milvus using Pandas, Unstructured.io, and ORMs (SQLAlchemy, PyMilvus)
* Created custom LangChain retrievers with prompt chaining for improved query handling, achieving 80% routing accuracy
* Built a user-friendly Streamlit UI and streamlined deployment with Docker Compose for reproducible environments
* Discovered 2+ clear segments of investors by analyzing investors data using cluster analysis and dimension reduction

**Evidation Health | University of California, Santa Barbara** Santa Barbara, CA

**Data Science Researcher** January 2023 - June 2023

* Analyzed patients' behavior to respiratory illness contraction by performing dimension reduction on health wearables data with 10+ features into 2 dimensions leveraging the t-SNE algorithm, and discovered 2-3 distinct patterns in participants' responses
* Scaled, centered, and binned time-series quantitative data; analyzed dimension reduced processed data by employing seaborn and plotly graphs and animations to perform explanatory data analysis
* Managed logs and records of all of 20+ weekly meetings with Evidation data science team to lead team of 5 members with action items and issues to be addressed for upcoming meetings

**projects**

**LLM Text Compressor | PyTorch, HuggingFace** September 2024 – November 2024

* Developed a practical parallelizable compression decompression algorithm to be used on any text based on Proof of Concepts from papers: Google Deepmind’s LLM Compressor, and LLMZip using the libraries Huggingface, and PyTorch
* Increased distribution and compression performance by mapping to Unicode
* Experimented and wrote a survey paper on the results of different pretrained models like Llama, Phi-3, and TinyLlama

**Developing a Recommender System | Spark, MapReduce** January 2023 – May 2024

* Implemented efficient SON algorithm in PySpark and applied to Kaggle dataset to find all combinations of frequent item sets
* Explored and implemented custom partitioning method to facilitate shuffling process in large dataset for MapReduce job
* Learning foundations of Recommender Systems like Locality-Sensitive Functions, content-based recommendations, and collaborative filtering

**CatDB - NoSQL Database System with query language | Python** August 2023 - December 2023

* Implemented a database management system manipulated chunks of 200+ MB of data with functions for scanning, external sorting, joining, filtering, grouping, aggregation, and CRUD operations
* Created a custom cat-like SQL and MongoDB inspired query language and a query execution engine to manipulate 200+ MB of data with user-defined limited memory usage to be used through a command line-interface

**Classification of Frost in NASA images | Tensorflow** November 2023 - December 2023

* Achieved a test accuracy of 93% and test F1 of 94.8% after creating data pipeline to perform transfer learning on EfficientNetB0, ResNet50, VGG16 and CNN on satellite images supplied by NASA for class final deep learning project
* Minimized effects of overfitting of CNN by applying 6 methods: L1 regularization, dropout layers, early stopping, batch normalization, data augmentation, and minimizing learning rate
* Augmented 2GB+ images on Keras to induce variability by randomly cropping, flipping, rotating, translation, brightening images