**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 and Large Language Models, Data Mining with Apache Spark, PyTorch, Keras / Tensorflow, Recommender Systems, LangChain, SQL, MongoDB, Firebase, DynamoDB

**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, Algorithms and Data Structures, Time Series, Data Visualization, 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** Santa Barbara, CA

**Data Science Researcher Capstone** 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 in LaTex

**Data Mining and Recommender Systems with Spark | 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
* Applied Locality-Sensitive Functions, content-based recommendations, and collaborative filtering on Yelp Data
* Analyzed social network communities by implementing Girvan Newman algorithm in PySpark

**UpAllNight Forums | Langchain, Milvus** January 2024 – May 2024

* Scraped internet forums and public data using Selenium
* Deployed Milvus on Google Cloud Platform using Docker
* Preprocessed pharmacy text data with Langchain and ORM (PyMilvus)
* Allowed seamless integration with FastAPI and deployed dockerized application on Heroku

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

* Achieved a test accuracy of 93% and test F1 of 94.8% after finetuning EfficientNetB0, ResNet50, VGG16 on satellite images supplied by NASA for class final deep learning project
* Minimized effects of overfitting 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