# Alan Wang

858-539-5564 | Authorized to work in U.S. | zew@seas.upenn.edu | zew013.github.io

## EDUCATION

| | | |
|---|---|---|
| **University of Pennsylvania** | *M.S. in **Data Science*** | 2024 - 2026 |
| **University of California San Diego** | *B.S. in **Data Science**, B.S in **Applied Math & Economics*** | 2020 - 2024 |

***Coursework*:** Data Structure, Algorithms, Data Mining, Database, Data Visualization, Machine Learning, Deep Learning, Computer Vision, NLP, Big Data Analytics, Statistics, Probability, Optimization, Regressions, Hypothesis Testing, Forecasting

## SKILLS

***Programming language*:** Python, Java, Cpp, SQL, R, HTML, JavaScript, CSS, Shell, VBA

***Machine Learning Stack:*** Pytorch, Lightning, Tensorflow, Keras, ONNX, XGBoost, Scikit-Learn, Scipy, Numpy, Pandas

***Big Data/Database/Cloud:*** Apache (Hadoop, Spark), Dask, AWS (S3, EC2, Lambda, Redshift, EMR), Exasol

***Others*:** D3.js, Matplotlib, Tableau, Git, Heroku, Kubernetes, Excel, Microsoft Office Suite

## PROFESSIONAL EXPERIENCE

***Data Science Intern* | TE Connectivity** — **May. 2024 – Aug. 2024**

- Worked on the **digital transformation** initiative focusing on cost data. Consolidated large datasets from SQL databases, ERP systems, and Excel files into **AWS S3** and **Redshift**. Streamlined historical cost data collection and accessibility for future projects through cross-functional collaboration.
- Launched an **Auto-ML** pipeline on **AWS SageMaker** to improve cost estimations. Reduced cost estimation time from hours to 10 minutes, allowing cost analysts to focus on strategic decision-making rather than manual estimates.
- Designed cost models and generated **statistical insights** using **Excel-VBA** and retrieving cloud data for real-time updates.

***Machine Learning Intern* | Grant Street Group** — **May. 2023 – Aug. 2023**

- Proposed and developed a **machine learning** powered monitoring system. Tested models under the **Scikit-Learn**, **XGBoost**, **Prophet**, and **Pytorch** frameworks for anomaly detection.
- Manipulated hundreds of millions of data points using **SQL**, **Python**, **Pandas,** and **Spark** with **Exasol data warehouse**, and set up a CI/CD pipeline for automated model retraining and updates. Led a team of four to implement a new system that improved the F1 score from 0.15 to 0.6, replacing the previous static threshold-based system.
- Proficient at using **SQL**, **Tableau**, **Python** to deliver **data visualizations** and **statistical analysis** for daily operations.

## RESEARCH EXPERIENCE

***Data Science Capstone Owner* | Prof. Alex Cloninger** — **Oct. 2023 – Apr. 2024**

*GenAI: Diffusion Models for Image and Data Generation* | *GitHub, Webpage*

- Investigated how scene representations are generated during the diffusion process. Demonstrated that 3D properties are learned early in the denoising stage before human visual recognition by inserting probing classifiers into self-attention blocks.
- Created a synthetic dataset of generated images and their depth masks with carefully designed architecture.

***Research Assistant* | Rappel Laboratory** — **Feb. 2023 – Oct. 2023**

*Image Segmentation and Propagation Analysis Program for cAMP Waves in Cell Aggregation Stage* | *Slides Demo*

- Developed a two-stage Python program that segments more than 60 GB of images and videos, applies an unsupervised clustering algorithm for data cleaning, and constructs velocity vector fields for scientific analysis.
- Collaborated with different stakeholders to make improvement. Optimized and parallelized the code, reducing average processing time from 50 minutes to 4 minutes.

***Research Assistant* | Prof. Richard Carson & Prof. Dale Squires** — **Dec. 2021 – Dec. 2022**

*Data-Driven Analysis of Ethical Preferences in UN Membership Policies & Assumptions in Conditional Logit Model*

- Developed an ETL data mining pipeline using Python and AWS to create a large dataset from 70 years of United Nations policy documents. Improved processing efficiency and accuracy, especially for handwritten records.

- Performed statistical analysis that provided support for established and consistent ethical preferences, which could serve as a standard to guide and facilitate multilateral cooperation by reducing conflicts and information costs.

## PROJECTS

***Language Intention Classification & Model Compression Full Stack Development*** | *Webpage*

*Deep Learning and Natural Language Processing:*

- Used **BERT** as the encoder and a **Neural Network** as the decoder to classify text intentions.
- Leveraged **Knowledge Distillation** by using the trained BERT-NN as the teacher model and **BiLSTM** as the student model, compressing the model size from **439MB** to **70MB** while preserving comparable accuracy.

*Model Integration and Application Development:*

- Leveraged **ONNX Runtime** to accelerate inference speed by **6**x, reducing time per call from **0.026** to **0.0043** seconds.
- Deployed the compressed model on **Heroku** server, using **Gunicorn** and **Flask-RESTful** for the app backend, with the model stored on **Amazon S3**.

***Enhancing Scientific Literature Understanding through Large Vision Language Model***

- Scraped and collected 1,059 scientific articles to build a richly labeled, auto-generated dataset, enhancing training data quality for complex document layouts.
- Fine-tuned Qwen2-VL-2B, a large vision-language model, specifically for layout parsing and reading order comprehension, enabling improved interpretability of structured scientific content.

***Using CNN and LSTM models for Image Captioning on COCO Dataset*** | *GitHub, Report*

***Building Neural Network from Scratch & Building Transformer in PyTorch*** | *GitHub, Report*

- Implemented a neural network in Python and coded backpropagation, mini-batch gradient descent, and cross-validation using NumPy from scratch. Added early stopping, momentum, and L1 & L2 regularization to enhance the model.
- Conducted performance experiments with sigmoid, tanh, ReLU, and softmax as activation functions.

## PART-TIME EXPERIENCE

***Head Data Science and Machine Learning Teaching Assistant*** *(paid)* **| HDSI, Penn Engineering** **Mar. 2023 – Present**

- Automated the grading process by developing test cases and grading systems on Python and Jupyter Notebook.
- Leveraged extensive knowledge of statistics and machine learning with excellent communication between professors, other teaching assistants, and students, assisted over 800 students by conducting office hours, leading labs and discussions, as well as creating and grading course content.

***CSE-PACE Program Designer*** *(paid)* **| UCSD CSE Department |** *Webpage* **May. 2022 – Sep. 2022**

- Addressed issues that disproportionately affect students from historically marginalized groups by crafting programs that prioritized communication and peer relationships over sheer knowledge acquisition.
- Successfully implemented the funded program as part of the computer science curriculum and supported over a thousand students.

***Data Analyst, Tech VP*** **| Lumnus Consulting (Student Enterprise) |** *Webpage* **Nov. 2021 – Feb. 2023**

- Led team projects by building data analysis models and creating visualizations, facilitating clear communication of data insights. Launched and maintained the company website using React.js and Heroku.
- Organized data analysis projects and alumni speaker sessions, fostering collaboration and knowledge sharing.