David Lu

david.linyi.lu@gmail.com • david5010.github.io • 650-305-6097

EDUCATION

University of California, Irvine | GPA: 3.86/4.00 Irvine, CA
B.S. Computer Science, Statistics; Specialization: Machine Learning and Statistical Methods *Graduating June 2024*
Relevant Courses: Data Structures, Machine Learning, Graphical Models, Generative Models, Bayesian Inference
Extracurriculars: Vice President of Technology - **Alpha Kappa Psi** | Education Director - **Commit the Change (CTC)**

SKILLS & INTERESTS

Language: Fluent in English, French, Mandarin

Languages/Framework: Python, Julia, R, SQL, C/C++, MATLAB, PyTorch, Scikit-learn, Tensorflow, Numpy, Pandas, Apache Spark, Dask, Airflow, Xarray, GeoPandas

Software: Linux, AWS, Docker, MongoDB, PostgreSQL, Git

Interests: Finance, Climate Science, Computer Vision, NLP, Art, Sailing

EXPERIENCE

Deep Data Lab Irvine, CA
Machine Learning Researcher *September 2023 - Present*

- Conducted NMEC research using Pecan Street with NOAA temperature statistics for enhanced energy usage prediction
- Developed PyTorch models to assess energy savings in programs through counterfactual scenario analysis
- Employed OpenEEmeter for benchmarking and validation of energy savings calculated by the PyTorch model, ensuring compliance with industry-standard efficiency measurement methods and enhancing the reliability of the results

Baldi Lab Irvine, CA
Machine Learning Researcher *February 2023 - Present*

- Designed various architectures such as DeepSet, and Set Transformer that can predict the quality of an antenna array
- Improved existing antenna pattern design by 89% by creating a custom gradient descent in PyTorch
- Boosted pattern cost calculation efficiency by 5000% by transitioning MATLAB code to PyTorch Tensors

EDF Innovation Lab Palo Alto, CA
Machine Learning Engineer for Energy Markets - PyTorch, Pandas, Airflow, Xarray *September 2022 - January 2024*

- Implemented various deep-learning models with time dependencies to forecast electricity load based on climate data
- Explored Dask and SageMaker integration for enhanced data processing and efficient model training with large datasets
- Developed an algorithm that integrates population density data to enhance temperature-dependent load estimation
- Expedited the team's research processes by over 30%, by engineering spatial data pipelines using Xarray and GeoPandas
- Predicted natural gas outages by analyzing the relationship with extreme temperature and spatial relations using Xarray

Lyrid Irvine, CA
Software and Data Engineer Intern - Airflow, PyMongo, Spark *December 2021 - September 2022*

- Designed an internal API to easily retrieve user data from MongoDB and facilitate data dumping for user activity analysis
- Improved the features of an authentication service by utilizing graphene and Django to create GraphQL mutations
- Saved the team 20 hours by resolving two critical bugs for Sheliak by implementing test cases using Insomnia
- Reduced development time by 8% by prototyping a scalable ETL pipeline using Airflow, Spark and Delta Lake

Yes! Star Corporation Shanghai, SH
Data Science Intern - Pandas, Numpy, Seaborn *June 2019 - July 2020*

- Reduced the data noise of the company's projected sales regression model by cleaning 36 datasets using Pandas
- Saved 50 hours for the marketing team by automating the visualization of hundreds of datasets using Seaborn
- Increased accuracy of market analysis by 10% by finding target market data by web scrapping various websites

PROJECTS

Financial Data Processor Airflow/Postgresql/Psycopg2
Extracted financial data from various sources and loaded into a Postgres database

- Customized a web scraper to bypass Yahoo Finance's automation detection to collect millions of data entries
- Automated the process of collecting data from web scraping and IEX Cloud and load onto Postgres using Airflow