Sihan (Michelle) Zhou

michelle.zhou.701@gmail.com | (757)291-1113 | Website: https://miichellezhou.github.io/mywebsite/

Education

William & Mary (W&M) Williamsburg, VA January 2024

B.S in **Computer Science** and **Applied Mathematics, Statistics & Probability Concentration** GPA: 3.96

Honors: Phi Beta Kappa, Dean's List (all semesters)

Relevant Coursework: Adv Applied Machine Learning, Adv Statistical Data Analysis, Database, Adv Linear Algebra

Research & Publications

"Triangulation-based Spatial Clustering for Adjacent Data with Heterogeneous Density" Prof. Guannan Wang

(11/2023, submitted to *ASA Journal Statistical Analysis and Data Mining*)

- Designed a Density and Triangulation-based Clustering (DTC) method, which combines Delaunay triangulation with kernel density estimation and DBSCAN to solve adjacent problems for arbitrarily shaped data with various density
- Implemented the DTC algorithm into a full Python package and improved the algorithm speed by applying FAISS
- Presented and held poster session at ASA's *2023 Symposium of Statistics & Data Science* and W&M *SciFri* event
- Paper and code are on Github: https://github.com/MiichelleZhou/Density-and-Triangulation-based-Clustering

"The Inverse Characteristic Polynomial Problem for Graphs over Finite Fields" Prof. Charles Johnson

(5/2023, published on *IntechOpen, Recent Research in Polynomials*)

- Studied possible characteristic polynomials that can be realized by matrices over a finite field such that the graph of the matrix is a tree, designed experiments in Mathematica to test the proposed hypotheses and theories

Work Experience

Data Science - Natural Language Processing Intern June 2023 - August 2023

PangeaChat Richmond, VA

- Built model that generates user's vocab performance data with timestamps to mimic real language learner's behavior in the first 3 months, considered word frequency distribution, forgetting curve, and use of assistant tools
- Applied generated data on RNN to predict user's learning proficiency in a conversational spaced repetition system
- Analyzed app user's memory curve for language learning and the correlation between the learning proficiency and the use of memorization tools, using statistical analysis to predict the user's future performance
- Consulted experts who analyzed human behaviors in reinforcing language learning through spaced repetition system
- Collaborated with the front-end team to add a proficiency estimation feature in the learning chatbot accordingly

Machine Learning Intern August 2022 - May 2023

AidData Williamsburg, VA

- Reduced 90% of the annotation and staff training costs for the sectoral purpose coding of China's foreign financial activities by implementing the fast vote-k algorithm to select the most representative and diverse training data
- Deployed and inferenced large language model (LLM) FLAN-UL2 on AWS Sagemaker to perform autocoding task
- Improved model's autocoding accuracy from 88% to 97% by building Pinecone vector database for training data and implemented similarity-based dynamic retrieval of few-shot learning examples
- Used Python on data mining for Chinese Soft Power Indicators such as Chinese media/TV/radio availability in other countries, built SQL database for relevant articles, and prepared the documentation for future analysis
- Presented the model pipeline to Microsoft's research team and W&M IT department to facilitate cooperation

Machine Learning Intern May 2022 - May 2023

Institute for Integrative Conservation Williamsburg, VA

- Conducted topic modeling for document clustering of conservation related tools, developed the cluster results into a tool recommendation feature on the website, improved user's experience and searching efficiency
- Applied GPT3-turbo to automate a categorization pipeline for conservation tools and approaches, designed Drupal database architecture, taxonomy, content type, blocks and fields according for importing the categorized data
- Integrated OpenAI API into Drupal's search engine to augment the website's searching and ranking functionality

Data Analyst Intern June 2021 - August 2021

Bytedance Ltd. Beijing, China

- Monitored and created dashboards for TikTok and other products' global payment, analyzed the trend of payments across different channels, reported abnormal amount change and potential mistakes to the operations team
- Used SQL to retrieve bill reconciliation data between merchants and disbursement channels, queried doubtful transactions, sorted out currency and amount mismatched transactions, proposed possible solutions to the R&D team

Project Experience

DHS Cross-border Migration Prediction May 2023 – Present
W&M Geolab / Department of Homeland Security (DHS) Williamsburg, VA

- Train the baseline Resnet model distributedly on HPC using Pytorch, using DHS's satellite imagery for each country's municipalities to predict the number of cross-border migrations from other countries to the United States.
- Implemented a customized Pytorch dataloader to solve inconsistent image size problem.

Explainable Synthetic Media Detection June 2023 – August 2023
W&M DisinfoLab Williamsburg, VA

- Trained the BA-TFD+ model to flag artificially and digitally generated video and audio using VideoSham dataset
- Integrated gradient attention rollout, the class specific explainability model for vision transformer, into BA-TFD+, to visualize the neural networks's focus by plotting heatmaps that highlighted the manipulated regions on media

SCOPE website (https://www.scopedata.org) September 2022 - May 2023
W&M Geolab Williamsburg, VA

- Streamlined the geo-coding process by highlighting key concepts from articles, extracted using Named Entity Recognition through fine-tuned large language models from HuggingFace and Low Rank Adaptation (LoRA)
- SCOPE is used internally for Geolab researchers and externally for World Bank.

Programs & Extracurricular Activities

IDEAS program on AI & Ethics (selected 20 out of 400) August 2023
Northeastern University & Harvard University Boston, MA

- Participated in workshops focused on building responsible and explainable AI, collaborated with practitioners and scholars in the AI Ethics domain, and engaged in challenging discussion on the transparency of current AI models
- Conducted literature review on model's tradeoff between performance and explainability under high-stake context
- Delivered a capstone project identifying the inconsistency problem in current neural networks and decision trees used by medical clinicians, matching each problem with a feasible explainability model to address the black-box dilemma

William & Mary Global Innovation Challenge (WMGIC) - External Relation Director April 2021 - May 2022
William & Mary Williamsburg, VA

- Reached out to 80+ keynote speakers, mentors and judges for WMGIC VI, recruited 150+ students from 21 schools over 5 continents to compete on finding sustainable solutions for sea level rise problem in Nigeria
- Organized and moderated WMGIC x NATO's Cybersecurity Innovation Challenge, invited 200+ participants from diverse backgrounds to devise solutions addressing disinformation and digital democracy challenges faced by NATO
- Led and coordinated a month-long fundraising event, successfully raising $4,000+ for the club's future development

Consulting Club - Member June 2021 - August 2021
William & Mary Williamsburg, VA

- Helped an auto detailing startup to find an optimal relocation site to increase exposure, traffic, and profitability while retaining proximity to old customers and ensuring the expenses fit within the budget
- Created a geospatial map in R to visualize the main roads in Williamsburg and all the important decision factors used in the decision tree model to assist the relocation decision making

Awards & Skills

Certificates:

- Azure Data Scientist Associate (DP-100) Certificate by Microsoft
- Ultimate AWS Certified Cloud Practitioner CLF-C01 & CLF-C02 Certificate by Udemy

Competitions:

- 1st place in Kaggle's CCI Machine Learning Disinformation Detection (Spring 2023)
- Successful Participant of 2023 Mathematical Contest in Modeling (MCM) (Spring 2023)
- Best Github Use and Best UI winners in Cypher VI competition (Spring 2022)

Scholarship:

- 2023 International Student Opportunity Scholarship (awarded for presenting on conferences)

Technical Skills:

- Programming Languages: Python (proficient), SQL (proficient), R (proficient), PostgreSQL, Pytorch, Java, C/C++
- Data Analysis & Visualization: Tableau, PowerBI, RShiny, PowerBI, Google Analytics
- Other tools: Azure, AWS, Github, Huggingface, LangChain